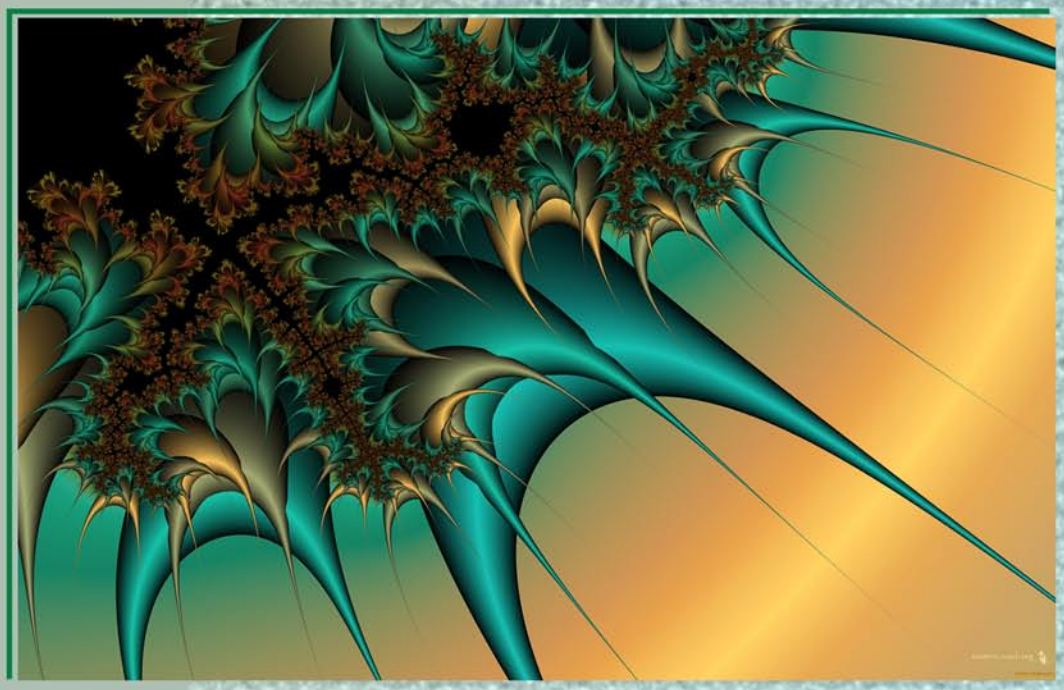
НА ДОПОМОГУ СТУДЕНТУ УДФСУ

В.В. Лаговський,
В.М. Краєвський

Бізнес-аналітика та моделювання



Бізнес-аналітика та моделювання

СЕРІЯ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ»

Серію «На допомогу студенту УДФСУ» засновано 2016 року.

Редакційна колегія:
Пашко П. В., д.е.н. (голова)
Шевчук О. А., д.е.н. (заступник голови)
Топчій В. В., д.ю.н.
Мацелюх Н. П., д.е.н.
Кужелєв М. О., д.е.н.
Швабій К. І., д.е.н.
Горбовий А. Ю., д.т.н.
Чмелюк В. В., к.ю.н.
Малинський І. Й., к.н.фіз.вих.
Шевчук В. А., к.ю.н.

У СЕРІЇ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ» ВИЙШЛИ ДРУКОМ:

2016

«Методичні основи спеціальної фізичної та технічної підготовки студентів за розділом «Легка атлетика»
«Самостійна робота студента як одна з форм впливу на функціональну, фізичну та психологічну підготовленість»
«Організація роботи командира механізованого взводу»
«5.45-мм автомати Калашникова (АК-74, АКС-74, АК-74Н, АКС-74Н) та 5.45-мм ручні кулемети Калашникова (РПК-74, РПКС-74, РПК-74Н, РПКС-74Н)»
«Гранатомет підствольний ГП-25»
«Ручні гранати»
«Кулемети Калашникова – 7.62, ПК, УЖМ, ПКТ»
«Ручний протитанковий гранатомет РПГ-7»
«9-мм пістолет Макарова (ПМ)»

2017

«Вища та прикладна математика»
«Цивільний захист»
«Програмування мовою JAVA : практикум»
«Інформаційні системи і технології в юридичній практиці»
«Дослідження операцій : практикум»
«Чисельні методи»
«English for Students of Finance»
«Основи військової розвідки»

2018

«CASE-технології. Міждисциплінарне інформаційне моделювання»
«Економічна інформатика: практикум»
«Економічна теорія (політекономія, мікроекономіка, макроекономіка). Політекономія»
«Економічна теорія (політекономія, мікроекономіка, макроекономіка). Мікроекономіка»
«Економічна теорія (політекономія, мікроекономіка, макроекономіка). Макроекономіка»
«Охорона праці»
«Економіка і організація діяльності об'єднань підприємств»
«Основи християнської культури»
«Економіка підприємства»
«Фізика»
«Трудове право України»

2019

«Основи тактичної медицини»

«Аудит»

«Збірник задач. Вища та прикладна математика»

«Міжнародні розрахунки та валютні операції»

«Підготовка озброєння механізованого взводу до бойового застосування»

«Контролінг в управлінні підприємством»

«Актуальні питання судової експертизи (у питаннях і відповідях)»

«Інноваційний менеджмент»

«Організація навчальних занять з фізичного виховання за методом колового тренування»

«Теорія судових доказів (у таблицях і схемах)»

«Статистика»

«Фінансовий менеджмент проектів і програм»

«Ручний протитанковий гранатомет РПГ-7 (РПГ-7Д)»

«Організація роботи командира механізованого відділення»

«Психологія управління»

«Deutsch und Wirtschaft»

«Основи тактичної підготовки працівників правоохоронних органів»

«Основи кінології»

«Моделювання систем»

«UML. Уніфікована мова моделювання інформаційних систем»

«Історія держави і права України»

«Економічна теорія»

2020

«Менеджмент»

«Бізнес-аналітика та моделювання»

«Безпека життєдіяльності»

«Статистичний аналіз даних»

«Хімія»

«Мікробіологія»

«Медичне право»

«Бухгалтерський облік в управлінні підприємством»

«Загальна психологія»

«Хортинг в закладах вищої освіти»

«Організація бухгалтерського обліку»

УНІВЕРСИТЕТ ДЕРЖАВНОЇ ФІСКАЛЬНОЇ СЛУЖБИ УКРАЇНИ

СЕРІЯ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ»
Заснована 2016 року

В. В. Лаговський,
В. М. Краєвський

Бізнес-аналітика та моделювання

Навчальний посібник

Ірпінь
2020

УДК 658:005.52(075.8)
ББК 65.053я73
Л 14

Рекомендовано до друку Вченою радою
Університету державної фіскальної служби України
(протокол № 5 від 30 травня 2019 р.)

Рецензенти:
Охріменко І. В., д.е.н., професор, проректор з навчальної та наукової роботи ККІБіП;
Савчук В. К., д.е.н., професор кафедри статистики та економічного аналізу НУБіП;
Лісовий А. В., д.е.н., професор, завідувач кафедри аудиту та економічного аналізу Університету ДФС України

Лаговський В. В.
Л 14 **Бізнес-аналітика та моделювання** : навчальний посібник / В. В. Лаговський, В. М. Краєвський. – Ірпінь : Університет ДФС України, 2020. – 442 с. – (Серія «На допомогу студенту УДФСУ»; т. 58).

ISBN 978-966-337-550-2

У навчальному посібнику розглянуті основи бізнес-аналітики та моделювання соціально-економічних процесів. Зокрема, досліджені основні поняття бізнес-аналітики та моделювання; основи організації баз даних; методики виявлення нових знань; основні технології бізнес-аналітики; інформаційні аспекти дослідження систем; основи інтелектуальної обробки даних (Business Intelligence, Data Mining, Knowledge Discovery, Data Warehouse, OLAP). Навчальний посібник дозволяє сформувати систему теоретичних знань і практичних навичок застосування методів бізнес-аналітики та моделювання для дослідження бізнес-процесів з метою прийняття ефективних управлінських рішень.

Призначений для студентів, аспірантів і викладачів фінансово-економічних спеціальностей та тих, хто цікавиться сучасними статистичними методами аналізу й опису інформації.

УДК 658:005.52(075.8)
ББК 65.053я73

© Лаговський В. В., Краєвський В. М., 2020
© Університет державної фіскальної служби України, 2020

ISBN 978-966-337-550-2

ЗМІСТ

Вступ.....7

РОЗДІЛ 1. ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ОСНОВИ БІЗНЕС-АНАЛІТИКИ ТА МОДЕЛЮВАННЯ.....9
§ **1. Сутність бізнес-аналітики та її роль у сучасному управлінні організаціями**.....9
1.1. Загальне уявлення про бізнес-аналітику та її місце в системі управління.....9
1.2. Сутність бізнес-аналітики та її завдання. 14
1.3. Роль бізнес-аналітики і моделювання у побудові систем управлінського обліку 26
1.4. Поняття та сутність аналітики як основи бізнес-аналітики. Структура та завдання аналітики 33
§ **2. Сутність системного підходу як методологічної основи бізнес-аналітики** 44
2.1. Сутність системного аналізу 44
2.2. Формування проблеми та виявлення цілей 49
2.3. Методи системного аналізу 54
2.4. Елементи, структура, властивості та класифікація систем 63
2.5. Управління в системі та управління системою 76
§ **3. Теоретичні основи інформаційно-аналітичної діяльності як складової бізнес-аналітики**.....79
3.1. Основні поняття 79
3.2. Параметри інформації.....82
3.3. Методи збору аналітичної інформації.....84
3.4. Базові моделі, що застосовуються під час аналізу інформаційних явищ.....86
3.5. Інформаційно-аналітична діяльність 87
3.6. Характеристика основних різновидів ІАД 97
3.7. Проблема інтерпретації фактів та типові помилки в аналітичних висновках 104
3.8. Інформаційно-аналітичні технології та методи організації ІАД.....107
3.9. Основні методи аналітики та організації ІАД.....110

§ **4. Основи моделювання економічних процесів** 118
4.1. Сутність понять «модель» і «моделювання» 118
4.2. Принципи побудови моделей 125
4.3. Моделювання бізнес-процесів 127

РОЗДІЛ 2. ПІДГОТОВКА ДАНИХ ДО АНАЛІЗУ 140
§ **5. Консолідація даних** 140
5.1. Системи підтримки прийняття рішень 140
5.2. Введення в сховища даних 145
5.3. Основні особливості концепції СД 151
5.4. Огляд архітектур СД 155
5.5. Тест FASMI 159
5.6. Багатовимірні сховища даних 161
5.7. Реляційні сховища даних 166
§ **6. Введення в ETL** 179
6.1. Поняття ETL 179
6.2. Основні цілі і завдання процесу ETL 180
6.3. Вилучення даних в ETL 182
6.4. Особливості організації процесу вилучення даних 185
6.5. Очищення даних в ETL 187
6.6. Перетворення даних в ETL 191
§ **7. Трансформація даних** 206
7.1. Введення в трансформацію даних 206
7.2. Групування даних 216
7.3. Розгруповування даних 219
7.4. Злиття даних 219
§ **8. Візуалізація даних** 224
8.1. Вступ у візуалізацію даних 224
8.2. Групи методів візуалізації 228
8.3. Візуалізатори для оцінки якості моделей 253
8.4. Візуалізатори, що застосовуються
для інтерпретації результатів аналізу 262
§ **9. Очищення і попередня обробка даних** 270
9.1. Оцінка якості даних 270
9.2. Оцінка придатності даних до аналізу 276
9.3. Технології та методи оцінки якості даних 278

9.4. Очищення і попередня обробка 282
9.5. Типовий набір інструментів попередньої обробки в аналітичному додатку 287
9.6. Фільтрація даних 290
9.7. Обробка дублікатів і протиріч 294
9.8. Відновлення пропущених значень 300
9.9. Введення в скорочення розмірності 303
9.10. Семплінг 313

РОЗДІЛ 3. МЕТОДИ БІЗНЕС-АНАЛІТИКИ ТА МОДЕЛЮВАННЯ 323
§ **10. Задачі Data Mining** 323
10.1. Класифікація задач Data Mining у бізнес-аналітиці 323
10.2. Завдання класифікації і регресії 326
10.3. Задача пошуку асоціативних правил 328
10.4. Задача кластеризації 329
10.5. Моделі Data Mining 331
10.6. Процес виявлення знань 332
§ **11. Пошук асоціативних правил** 334
11.1. Формальна постановка задачі 334
11.2. Секвенційний аналіз 338
11.3. Різновиди задач пошуку асоціативних правил 341
11.4. Надання результатів 343
11.5. Алгоритм Аргіогі 347
11.6. Різновиди алгоритму Apriori 352
§ **12. Методи класифікації і прогнозування дерева рішень** 354
12.1. Загальна характеристика 354
12.2. Загальні характеристики алгоритмів конструювання дерев рішень 359
12.3. Алгоритми побудови дерев рішень 363
§ **13. Методи класифікації і прогнозування. Метод опорних векторів. Метод «найближчого сусіда». Байєсівська класифікація** 366
13.1. Метод опорних векторів 366
13.2. Метод «найближчого сусіда» 369
13.3. Байєсівська класифікація 374

§ **14. Нейронні мережі** 376
14.1. Загальна характеристика 376
14.2. Елементи нейронних мереж 377
14.3. Архітектура нейронних мереж 379
14.4. Моделі нейронних мереж 382
§ **15. Кластеризація** 384
15.1. Загальна характеристика задачі кластеризації 384
15.2. Методи кластерного аналізу. Ієрархічні методи 392
15.3. Методи кластерного аналізу. Ітеративні методи. 400
15.4. Порівняльний аналіз ієрархічних і неієрархічних методів кластеризації 407
15.5. Нові алгоритми і деякі модифікації алгоритмів кластерного аналізу 409
§ **16. Ансамблі моделей** 412
16.1. Введення в ансамблі моделей 412
16.2. Бегінг 416
16.3. Бустінг 419
16.4. Альтернативні методи побудови ансамблів 422
16.5. Стекінг 425

Список рекомендованих джерел 435

ВСТУП

Відомо, що сучасне суспільство вступило в новий етап розвитку, так званий інформаційний або постіндустріальний. Під цим зазвичай розуміють, що знання та інформація стали основною рушійною і продуктивної суспільною силою, яка визначає як духовний, так і матеріальний стан особистості, компанії, держави в цілому. Від уміння виробляти, шукати, аналізувати, класифікувати, узагальнювати, розпізнавати, переробляти, представляти інформацію й ухвалювати рішення сьогодні безпосередньо залежать якість життя людини і суспільства, їхня інформаційна і громадська безпека.

Останнім часом багато говорять про таку спеціальність, як бізнес-аналітика. Експерти підкреслюють, що сьогодні все більше потрібні фахівці широкого профілю, які вміють аналізувати роботу бізнес-компанії в цілому, а не окремих її підрозділів, компетентні як у сфері економіки, так і в інформаційних технологіях, які вміють обґрунтовувати ухвалення управлінських рішень на основі бізнес-аналізу.

Формально під бізнес-аналізом розуміється:

- визначення необхідної для аналізу інформації;
- процес обробки зібраної інформації засобами сучасних інформаційних технологій;
- виявлення існуючих закономірностей, залежностей, трендів;
- візуалізація отриманих результатів за допомогою схем, графіків і таблиць на основі сучасних програмних продуктів;
- вироблення рекомендацій та прогнозу розвитку підприємства.

Абсолютно очевидно, що в епоху BigData (великих обсягів даних) усі перераховані операції не можуть виконуватися вручну. Тому в арсеналі бізнес-аналітика, крім аналітичних знань і навичок, має бути ще й уміння користуватися сучасними програмними продуктами у сфері Data Mining (інтелектуального аналізу даних).

Сучасному бізнесу не обійтися без інформаційно-аналітичної підтримки, рішення задач прогнозування, управління ризиками та ін. Крім програмного забезпечення, потрібні фахівці, які

впроваджуватимуть аналітичні проєкти. Цей навчальний посібник, безумовно, вносить вклад у вирішення цих завдань.

Навчальний посібник може бути використано, в першу чергу, студентами й аспірантами, які навчаються за спеціальностями, пов'язаними з управлінським обліком та інформаційними технологіями. Багато питань, що розглядаються, певним чином включені в інші дисципліни, і викладачам буде нескладно адаптувати цей матеріал до них.

Навчальний посібник складається з трьох розділів. У першому розділі «Теоретико-методологічні основи бізнес-аналітики та моделювання» надається загальне уявлення про бізнес-аналітику та моделювання і їх місце в системі управління організаціями. Дається визначення основних понять та описуються загальні методичні підходи в моделюванні й аналітиці. Зокрема, розкривається сутність системного підходу, поняття системи, моделі тощо.

Другий розділ присвячений питанням підготовки даних до аналізу.

У третьому розділі розглядаються методи бізнес-аналітики та моделювання.

РОЗДІЛ 1
ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ОСНОВИ БІЗНЕС-АНАЛІТИКИ ТА МОДЕЛЮВАННЯ

§ 1. Сутність бізнес-аналітики та її роль у сучасному управлінні організаціями

1.1. Загальне уявлення про бізнес-аналітику та її місце в системі управління

Складна соціальна й економічна ситуація в країні, зростаюча конкуренція в різних сферах життєдіяльності суспільства вимагають від керівників різних організацій підвищення якості ухвалених управлінських рішень. Конкурентні переваги компаній та вміння «виживати» в складних економічних умовах значною мірою визначаються їхньою здатністю адаптуватися під ринкові зміни (впроваджувати інноваційні технології, нові продукти, змінювати методи управління тощо). Неможливо ухвалити правильне рішення, не володіючи різнобічною інформацією про явні і приховані процеси, які відбуваються в керованій структурі і в зовнішньому середовищі. Як показує практика, питання якості управлінських рішень у сучасних умовах перетворилося в питання забезпечення високої якості роботи інформаційно-аналітичних підрозділів. Стрімке зростання обсягів доступної різнопланової інформації, що використовується у процесі ухвалення рішень, потреба в оперативному до неї доступі та її обробки, розширення кількості джерел даних, швидка зміна соціально-економічних процесів та середовища функціонування організацій стали основними причинами стрімкого підвищення інтересу до бізнес-аналітики.

У загальному уявленні бізнес-аналітика – це процес аналізу інформації для ухвалення бізнесових рішень. Вона включає у себе методи збору й обробки інформації, оцінку ризиків, моделювання і прогнозування за допомогою інформаційних технологій. Бізнес-аналітику також розглядають як сукупність методів виявлення і вирішення всіх ділових. Вона охоплює великий спектр

питань. Тобто бізнес-аналітика – своєрідний набір методик, які використовують як сполучну ланку між деякими учасниками бізнесу для того, щоб зрозуміти структуру, правила та функції організації, а за підсумками запропонувати оптимальні рішення, які дозволять організації з упевненістю досягти поставлених цілей.

Першими можливості бізнес-аналітики оцінили банки, страхові, телекомунікаційні, торгові та інші компанії. Вони навчилися «витягати» корисні знання і вигоду з великої кількості інформації, яка була зібрана і зберігалася в корпоративних базах даних. Сьогодні для все більшої кількості організацій стає очевидно, що без аналізу таких даних в умовах конкурентної боротьби не обійтися.

На сьогодні в економічній літературі єдиного визначення терміна «бізнес-аналітика» (Business Intelligence) не існує. У першу чергу, це пов'язано з тим, що цей термін включає в себе значний спектр методів і технологій. Для прикладу наведемо визначення авторитетної консалтингової компанії IDC[1] (міжнародна корпорація даних). Бізнес-аналітика – це інструменти і додатки для пошуку, аналізу, моделювання і доставки інформації, необхідної для ухвалення рішень.

Бізнес-аналітика – мультидисциплінарна сфера, що перебуває на межі таких дисциплін, як: інформаційні технології, бази даних, алгоритми інтелектуальної обробки інформації, математична статистика, методи візуалізації даних тощо.

Сфера застосування бізнес-аналітики досить широка. Її можна застосовувати у ризик-менеджменті, фінансах, охороні здоров'я, індустріальному виробництві, сфері послуг, маркетингу, під час управління містами, роботи в Інтернеті і соціальних мережах, аналізу персональних даних тощо. Основне завдання бізнес-аналітики – зробити все можливе, щоб рішення, які ухвалюються керівниками організацій, були правильними, оптимальними і своєчасними. За готовими аналітичними звітами, які можуть містити не тільки цифри в різних розрізах, а й правила, залежності, тенденції, прогнози, рекомендації (тобто нові знання), приховується праця бізнес-аналітика.

---

[1] URL : http://idc.com/

Бізнес-аналітика – процес перетворення даних в інформацію, тобто у нові знання, які можуть бути використані для збільшення ефективності та конкурентоспроможність організації (підприємства).

Бізнес-аналітика – це дисципліна, яка використовується для виявлення ділових потреб і знаходження рішень ділових проблем. Рішення часто включають компонент розробки систем, але можуть також складатися з удосконалення процесів, організаційних змін або стратегічного планування та розробки політики організації. Людину, яка виконує це завдання, називають бізнес-аналітиком.

Бізнес-аналітик – фахівець, що використовує методи бізнес-аналітики для аналізу потреб діяльності організацій з метою визначення проблем бізнесу і розробки пропозицій їх вирішення. Тих бізнес-аналітиків, які займаються виключно розробкою програмних систем, також називають IT-бізнес-аналітиками, технічними бізнес-аналітиками або системними аналітиками.

Бізнес-аналітика як дисципліна тісно пов'язана з аналізом вимог, але вона націлена на розробку системи змін для організації, які потрібні для того, щоб організація досягла стратегічних цілей. Ця система змін може включати зміни в стратегії, структурі, політиці організації, процесах та інформаційних системах. Бізнес-аналітика – це комплексне вивчення діяльності підприємства, що дозволяє дати їй об'єктивну оцінку, виявити закономірності і тенденції розвитку, розкрити резерви бізнесу та недоліки в його роботі. Аналіз організації зосереджується на розумінні потреб: бізнесу в цілому; його стратегічного керівництва; ідентифікації ініціатив, які дозволять бізнесу досягнути стратегічних цілей. У першу чергу, бізнес-аналітика потрібна для власників та топ-менеджменту підприємства (компаній), щоб зрозуміти, як воно працює, який поточний стан справ з реалізації бізнес-цілей, щоб оцінити можливості компанії стосовно реалізації її стратегічних цілей, як використати наявні можливості та потужності, що потрібно змінити в організаційній структурі, методологічних підходах, технічному плані, які продукти та послуги, що будуть користуватись попитом, варто впровадити на найближчу перспективу тощо.

Бізнес-аналітика проводиться як для оцінки існуючого стану справ у компанії, так і для визначення майбутніх потреб. У більшості випадків бізнес-аналіз проводиться з метою окреслення шляхів вирішення бізнес-потреб, бізнес-цілей чи подолання перешкод на шляху до реалізації цих цілей.

Планування і контроль бізнес-аналітики визначають, які види діяльності необхідні для її проведення. Галузь охоплює: визначення зацікавлених сторін (співробітників підприємства, акціонерів); вибір методів бізнес-аналітики; процес керування вимогами; підходи до оцінювання ходу робіт для внесення в роботу необхідних змін. Планування і контроль включають організацію діяльності з бізнес-аналітики та її координацію з потребами команди проєкту.

Аналіз організації зосереджений на дослідженні поточного стану та перспектив бізнесу з метою формування контексту вимог і функціонального проєктування інновацій та охоплює виявлення й аналіз проблем бізнесу; визначення дій для пошуку необхідного рішення; моделювання бізнес-кейсів; складання техніко-економічних обґрунтувань; визначення рішення.

Одним із завдань бізнес-аналітики є збір і дослідження неструктурованої інформації про очікування і потреби замовника з метою отримання детальних, структурованих, узгоджених вимог до бізнес-рішення, яке розробляється. Виявлення вимог спрямоване на дослідження, виявлення та документування потреб зацікавлених сторін (співробітників підприємства, акціонерів). Воно включає збір інформації від зацікавлених сторін, з'ясування їхніх потреб і контроль правильності та повноти розуміння потреб. Послідовність і функціональні особливості роботи з вимогами на різних етапах бізнес-аналітики дають змогу виділити такі її завдання:

– планування процесу розробки вимог, визначення пріоритетів вимог і управління змінами;

– збір вимог – описує методики збору вимог від зацікавлених осіб;

– аналіз вимог – описує, як розробити і визначити вимоги з деталізацією, достатньою для успішної реалізації проєктною групою;

– комунікація вимог – описує методики гарантування того, що в усіх зацікавлених осіб є загальне розуміння вимог і способів їх реалізації;

– оцінка і перевірка правильності рішення – описує, як бізнес-аналітик може перевірити правильність запропонованого рішення, як підтримати реалізацію рішення й оцінити можливі недоліки в реалізації.

Аналіз вимог описує роботу проєктної команди для вироблення рішення, яке відповідатиме визначеним потребам бізнесу та зацікавлених сторін. Ця діяльність включає аналіз і контроль заявлених вимог стейкхолдерів, оцінку поточного стану бізнесу для визначення та рекомендування рішень і в кінцевому підсумку перевірку й оцінку результатів.

Комунікації і керування вимогами забезпечують взаємодію та взаєморозуміння зацікавлених осіб і команди проєкту під час виконання робіт у рамках проєкту; описують управління конфліктами, проблемами і змінами; залежно від складності та методології проєкту, можуть включати керування формальними погодженнями, відстеження різних версій документів, відповідність виконання рішень визначеним вимогам.

Оцінка й узгодження рішень передбачають оцінювання запропонованих аналітиком рішень з метою визначення максимально відповідних стратегічним цілям і потребам бізнесу; виявлення вад у рішеннях і визначення необхідних змін; контроль відповідності реалізованих рішень початковим поставленим вимогам; оцінку продуктивності й ефективності проєктів.

Оскільки бізнес-аналітика охоплює досить широкий спектр питань, була тенденція спеціалізуватися на одній з трьох її сфер.

1. Стратег – бізнес-аналітик, який задовольняє потребу організації у вирішені стратегічних питань, є досвідченими в аналізі стратегічної мети організації та її середовища, консультуючи вище керівництво з питань належної політики організації та ефектів від політичних рішень.

2. Архітектор – часом організаціям необхідно ввести зміни, щоб вирішити ділові проблеми, знайдені стратегічним аналітиком. Бізнес-аналітики аналізують цілі, процеси і ресурси та

пропонують шляхи, якими могла бути зроблена модернізація (реорганізація бізнес-процесів) або удосконалення (модернізація бізнес-процесів). Специфічні навички цього типу аналітика – це навички: знання бізнесу, розробки вимог зацікавлених осіб і моделювання бізнес-процесу. Основними для цього аспекту бізнес-аналітики є модернізація основних бізнес-процесів; підключення технологій для підтримки основних процесів і управління організаційними змінами. Цей аспект бізнес-аналітики також називають «удосконаленням бізнес-процесу» (модернізація бізнес-процесів).

3. Системний аналітик. IT-відділи, знаючи про бізнес-проблеми, звертаються до бізнес-аналітика, щоб краще зрозуміти і визначити вимоги для їхніх IT-систем. Хоча може бути деяке перекриття з розробником і завданням тестування, фокус зусиль завжди міститься на IT-частині процесу зміни компанії, і взагалі цей тип ділового аналітика залучений, тільки коли рішення про зміни було вже ухвалено вище.

Виходячи з вищезазначеного можна виділити кілька сфер, в яких працюють бізнес-аналітики:

– управління вимогами;
– дослідження й аналіз даних;
– моделювання процесів;
– проєктування;
– впровадження;
– консалтинг.

1.2. Сутність бізнес-аналітики та її завдання

Термін «бізнес-аналітика» з’явився в економічній літературі порівняно нещодавно, й однозначності в його трактуванні, як стає зрозуміло з вищезазначеного, не існує.

Міжнародним інститутом бізнес-аналізу (IIBA) бізнес-аналітика визначається як набір завдань і методів, які застосовуються для роботи як сполучна ланка між стейкхолдерами для того, щоб зрозуміти структуру, політику і роботу організації, а також рекомендувати рішення, які дозволять організації досягти своїх цілей.

Бізнес-аналітика економістами пов'язується з керованими змінами в бізнесі та визначається як практика залучення змін в організаційному контексті шляхом визначення потреб і рекомендації рішень, що приносять користь.

Бізнес-аналітики можуть виконувати кілька ролей у рамках управління змінами в бізнесі. Тому виник ряд професійних організацій у сфері бізнес-аналітики. Найбільшими з них є:

– Міжнародний інститут бізнес-аналітики (IIBA);

– Інститут управління проєктами (PMI).

Бізнес-аналітики можуть стати членами й інших професійних співтовариств, таких як Американське товариство з якості (ASQ), Міжнародна організація зі стандартизації (ISO), Британське комп'ютерне суспільство (BCS), Інститут аналітиків і програмістів (IAP) у Великобританії, Австралійська асоціація бізнес-аналітиків (ABAA) та ін.

Міжнародний інститут бізнес-аналітики[2] є некомерційною міжнародною професійною асоціацією, яка націлена на підвищення професіоналізму своїх членів, а також самої професії бізнес-аналітика. IIBA прагне до встановлення загальних стандартів у рамках професії та співпрацює з практикуючими аналітиками по всьому світу. За їхньої безпосередньої участі відбувається оновлення стандартів діяльності, поширення інформації про ефективні інструменти. IIBA був утворений у жовтні 2003 року. Станом на 2011 р. організація включала більше 18 тис. членів і мала більше 100 представництв по всьому світу.

Австралійська асоціація бізнес-аналітики (ABAA) була утворена 2003 року для підтримки розвитку бізнес-аналітики як професії на континенті. ABAA прагне забезпечити професійну основу і підтримку тих, хто працює у сфері бізнес-аналітики. ABAA як професійне об'єднання було утворено з метою:

– визначення професії бізнес-аналітика;

– сприяння розвитку професії та підвищення обізнаності громадськості про бізнес-аналітику;

– уніфікації основних компетенцій бізнес-аналітика;

[2] URL : https://www.iiba.org/ (https://ukraine.iiba.org/)

– поліпшення практики реалізації бізнес-аналітики;

– представлення професії на національному та міжнародному рівнях;

– вільного обміну інформацією та ідеями.

У версії 1.6 Зводу знань з бізнес-аналітики (BABOK) Міжнародний інститут бізнес-аналітики (IIBA) вказував, що бізнес-аналітик працює як сполучна ланка між зацікавленими сторонами з метою отримання, аналізу, підтвердження вимог до змін у бізнес-процесах, політиці організації та інформаційних системах. Бізнес-аналітик розуміє проблеми бізнесу і його можливості в контексті вимог і рекомендує рішення, що дозволяють організації досягти своїх цілей.

2009 р. з появою наступної версії BABOK роль бізнес-аналітика була деталізована і була пов'язана з розумінням того, як організації діють для досягнення своїх цілей, з визначенням можливостей організації в реалізації основних цілей діяльності з урахуванням інтересів зацікавлених сторін. Завдання, які стоять перед бізнес-аналітиком, включають у себе визначення цілей і напрямів змін, які організація повинна здійснити для досягнення поставлених цілей, а також визначення взаємодії різних структурних підрозділів організації та стейкхолдерів всередині і поза організації.

BABOK розділяє діяльність бізнес-аналітика на шість сфер знань, які визначають, що практикуючий бізнес-аналітик повинен розуміти, які завдання має змогу виконувати.

1. Аналіз підприємства – тут як ціль ставиться визначення бізнес-завдань і коло можливих рішень.

2. Планування та моніторинг – охоплює процеси вибору методів бізнес-аналітики, ідентифікації зацікавлених сторін, способів управління вимогами та оцінки результатів.

3. Збір інформації та виявлення вимог – передбачає тісну роботу зі стейкхолдерами з метою визначення їх потреб і вивчення середовища, в якій вони містяться.

4. Аналіз вимог – за мету ставить дослідження інформації про стейкхолдерів у рамках поточного стану бізнесу, визначення можливих механізмів обліку виходячи з вимог.

5. Управління вимогами – передбачає вибір методів для можливого врегулювання конфлікту інтересів, а також узгодження ухвалених рішень з усіма сторонами.

6. Оцінка рішення і перевірка – охоплює процеси виявлення прогалин та недоліків у рішеннях, оцінки пропонованих рішень.

Нова версія BABOK з'явилася в квітні 2015 р. (BABOK v. 3[3]). Зміни в новій версії є значними (зокрема, обсяг збільшився – більше 500 сторінок, водночас як у другій версії сторінок було близько 300). Важливим є те, що нову версію BABOK можна назвати більш зрілою: тут надана нова концептуальна модель бізнес-аналітики, більш чітко і повно визначені різні сфери її застосування, доповнений список компетенцій бізнес-аналітика, описано більше технік і методів.

Концептуальна модель (BACCM, Business Analysis Core Concept Model) визначає шість базових концепцій, що утворюють добре структуровану і цілісну основу бізнес-аналізу:

– зміни;
– потреби;
– рішення;
– зацікавлені особи;
– цінність;
– контекст.

Завдяки тому, що кожна сфера знання описується як взаємодія групи таких концепцій, а BACCM дає повну картину їхніх взаємозв'язків, структура завдань і дій бізнес-аналітика стає більш зрозумілою, логічною і комплексною.

Одній з таких концепцій, а саме «Зміни», в новій версії присвячено набагато більше уваги. Тут зміни визначені як дії щодо перетворення (бізнесу) у відповідь на потреби й описані обов'язки бізнес-аналітика з визначення обґрунтованості змін і підготовки організації до таких змін. Змінам у новій версії надається настільки велике значення, що це навіть відбилося у визначенні самої бізнес-аналітики як практики, що допомагає організації змінюватися (на відміну від визначення, наданого у версії 2, де

[3] URL : https://www.iiba.org/standards-and-resources/babok/

бізнес-аналітика була покликана допомогти організації в досягненні її цілей).

Для більш успішної реалізації змін версія 3 розширює сферу відповідальності бізнес-аналітика (раніше сконцентровану навколо вимог) ще й на проєктування рішення. Це цілком відповідає реальним проєктам, в яких бізнес-аналітик може сам бути розробником рішення або його частини (наприклад, розробляти бізнес-процеси, оргструктуру, різні регламенти та ін.) чи брати участь у проєктуванні.

Визначення вимоги теж зазнало змін, стало глибше. У третій версії BABOK вимога розглядається як опис потреби замовника, що спирається на розуміння того, яку реальну цінність для замовника несе ця потреба.

Цікавим нововведенням є і опис так званих перспектив – прикладних напрямів використання бізнес-аналітики в різних предметних сферах і різних типах проєктів: гнучка розробка, інформаційні технології, бізнес-архітектура, аналіз даних, управління бізнес-процесами. Описи цих спеціалізованих підмножин бізнес-аналізу можуть допомогти в освоєнні професії початківцями, а для досвідчених можуть використовуватися як дорожня карта або чек-лист. У попередній версії BABOK передбачалося, що бізнес-аналітик певним чином пов’язаний з IT-проєктом.

Досить помітно в BABOK версії 3 перероблений список рекомендованих технік: з’явилося півтора десятка раніше не описаних технік, а також змінені назви деяких згадуваних у другій версії. Потрібно зазначити, що всі ці зміни виникли не довільно, а є відображенням тих практик і тенденцій, які прийняті в галузі бізнес-аналітики. IIBA планує оновлювати версії BABOK кожні 3–5 років, щоб цей звід знань зберігав свою актуальність і практичну спрямованість. При цьому розробляючи чергову версію зводу знань, IIBA спирається на сукупний досвід і знання великої спільноти бізнес-аналітиків, які беруть участь в обговореннях контенту BABOK і надсилають свої пропозиції про зміни і доповнення до нього.

Основна мета BABOK – визначити професію бізнес-аналітики і надати набір загальноприйнятих практик, методів, які

допоможуть фахівцям-аналітикам домовлятися між собою і розуміти, які навички необхідні їм для ефективного виконання аналітичної роботи. Бізнес-аналітика – це професія, в якій бізнес-аналітики можуть виконувати різнопланову роботу в рамках компанії. Аналітики можуть використовувати різні компетенції, знання, навички, термінологію та підходи під час виконання завдань бізнес-аналітики. BABOK – це платформа, яка об'єднує всі точки зору, що описують завдання бізнес-аналізу, які дозволяють належним чином проаналізувати зміни або оцінити їхню необхідність. Завдання можуть варіюватися за змістом, порядком або важливістю для конкретного аналітика або для різних потреб.

У BABOK зазначено, що бізнес-аналітика – це діяльність, яка дозволяє впроваджувати зміни в компанії шляхом визначення потреб і рекомендації рішень, які забезпечують цінність для зацікавлених осіб. Бізнес-аналітика дозволяє компанії визначити потреби й обґрунтувати зміни, а також розробити й описати рішення, які можуть принести користь. Бізнес-аналітика проводиться з різних ініціатив у рамках організації. Це можуть бути ініціативи в стратегічному, тактичному або оперативному плані. Бізнес-аналітика може здійснюється як у межах певного проєкту, так і у ході еволюції компанії та її безперервного розвитку. Її застосовують, щоб зрозуміти поточний стан, визначити майбутній стан, а також визначити дії, які необхідні для переходу від поточного стану до майбутнього.

Бізнес-аналітик – це будь-яка особа, яка виконує завдання бізнес-аналізу, описані в керівництві BABOK, незалежно від своєї посади або організаційної ролі. Бізнес-аналітик відповідає за виявлення, узагальнення та аналіз інформації з різних джерел у рамках компанії, в тому числі інструментів, процесів, документації, а також зацікавлених осіб. Бізнес-аналітик відповідає за виявлення реальних потреб зацікавлених осіб (що часто включає в себе аналіз і прояснення, які висловлюють побажання) для того, щоб визначити основні завдання та виявити мотиви.

Бізнес-аналітики беруть активну участь у тому, щоб спроєктоване і реалізоване рішення співвідносилося з потребами зацікавлених осіб.

Зазвичай діяльність бізнес-аналітиків включає в себе:
– осмислення проблем і завдань компанії;
– аналіз потреб і рішень;
– розробку стратегій;
– аналіз великих і часом неструктурованих даних;
– впровадження змін;
– допомогу у взаємодії зацікавлених осіб.

Бізнес-аналітиком також можуть називати:
– бізнес-архітектора;
– аналітика бізнес-систем;
– аналітика даних;
– аналітика підприємств;
– консультанта з питань управління;
– аналітика процесів;
– менеджера щодо продукції;
– власника продукту;
– фахівця з вимог;
– системного аналітика.

Британське комп'ютерне суспільство розглядає бізнес-аналітика насамперед як внутрішнього консультанта, який займається дослідженням бізнес-систем, виявленням можливостей для їхнього поліпшення з метою задоволення потреб бізнесу з використанням інформаційних технологій.

Розглянуті професійні спільноти характеризують бізнес-аналітика по-різному. Проте загальним у цих характеристиках є те, що бізнес-аналітик – це особа, що проводить зміни в бізнесі за допомогою, зокрема, інформаційних технологій з метою вирішення існуючих проблем та покращення процесів.

Можна виділити три сфери, взаємодію з якими здійснює бізнес-аналітик у рамках вирішення поставлених перед ним завдань.

1. Вищий менеджмент організації, що визначає стратегічні цілі діяльності.

2. Структурний підрозділ організації, який відповідає за окремий напрям її діяльності (постачання, збут, ведення бухгалтерського обліку тощо). При цьому кожний підрозділ має керівника, який несе відповідальність за реалізацію його завдань і звітує перед виконавчим органом.

3. Структурний підрозділ з питань розвитку. Тут можна виділити вищий керівний елемент, наприклад IT-директора, керівника проєкту і команду, що забезпечує рішення. Завданням цього підрозділу є успішна реалізація проєкту як у рамках відведеного бюджету, так і поставлених термінів і графіків. Менеджер проєкту та команда, яка забезпечує рішення, вважають: якщо вище керівництво визначило, що продукт повинен бути розроблений, то це буде вигідно для організації. Технічні фахівці розробляють рішення, і правильність цього рішення може бути перевірена тільки на завершальному етапі.

Отже, виникає потреба у функції координації дій, яка може бути покладена на виконавчий орган. Однак завданнями останнього є визначення стратегічних цілей, тобто виконавчий орган, як правило, не пов'язаний з аналізом деталей проєктів. І тут виникає питання перевірки того, що результат проєкту повністю вирішує поставлене бізнес-завдання, і ця функція покладена саме на бізнес-аналітика. Він незалежно оцінює бізнес-завдання і визначає рішення для команди, що забезпечує реалізацію, а потім гарантує, що отриманий результат повною мірою приводить до вирішення наявної проблеми.

Бізнес-аналітик, у першу чергу, орієнтований на бізнес-проблему, а не на технічні можливості її вирішення. У деяких випадках бізнес-аналітик взагалі може бути не пов'язаний з технологіями, наприклад у разі надання допомоги керівництву вищого рівня у визначенні стратегії або збору інформації та виконання контрольних процедур. Отже, в центрі уваги бізнес-аналітика завжди є вирішення бізнес-проблеми.

В узагальненому вигляді сукупність завдань, які доводиться вирішувати бізнес-аналітику у ході його роботи, є такі:

– виявлення і розуміння бізнес-проблем організації та галузі, а також визначення шляхів їхнього вирішення;

– визначення сильних і слабких сторін бізнесу;

– розробка напрямів розвитку;

– аналіз ресурсів, вимог, бізнес-процесів, що стосуються пропонованого рішення проблеми;

– документування рішення бізнес-задач у належному вигляді;

– визначення сфер удосконалення внутрішніх бізнес-процесів;

– взаємодія зі структурними підрозділами організації, які задіяні у вирішенні бізнес-проблем з метою отримання зворотного зв'язку й оцінки альтернативних рішень;

– визначення очікувань зацікавлених сторін і управління ними.

Предметом бізнес-аналітики є одна з функцій управління, яка відображає технологічний етап процесу прийняття управлінських рішень щодо змін у бізнесі під впливом внутрішніх і зовнішніх факторів і зводиться до аналітичного забезпечення цих рішень.

Об'єкт дослідження бізнес-аналітика – діяльність компаній, спрямована на здійснення змін у бізнесі виходячи з вимог стейкхолдерів.

Методи бізнес-аналітики розглядаються як системне комплексне дослідження як самого бізнесу, так і його оточення з метою забезпечення проведення змін, які задовольняють потреби всіх стейкхолдерів.

У BABOK розглядаються такі методи в бізнес-аналітиці:

– Acceptance and Evaluation Criteria (визначення критеріїв приймання та оцінки);

– Backlog Management (управління беклогом);

– Balanced Scorecard (збалансована система показників);

– Benchmarking and Market Analysis (бенчмаркінг і аналіз ринку);

– Brainstorming (метод мозкового штурму);

– Business Capability Analysis (аналіз бізнес-можливостей);

– Business Cases (бізнес-кейси);

– Business Model Canvas (бізнес-модель «Канвас»);

– Business Rules Analysis (аналіз бізнес-правил);

– Collaborative Games (спільні ігри);

– Concept Modelling (концептуальне моделювання);

– Data Dictionary (словник даних);

– Data Flow Diagrams (діаграми потоків даних);

– Data Mining (інтелектуальний аналіз даних);

– Data Modelling (моделювання даних);

– Decision Analysis (аналіз прийняття рішень);

– Decision Modelling (моделювання прийняття рішень);
– Document Analysis (аналіз документів);
– Estimation (оцінка);
– Financial Analysis (фінансовий аналіз);
– Focus Groups (фокус-групи);
– Functional Decomposition (функціональна декомпозиція);
– Glossary (словник);
– Interface Analysis (аналіз інтерфейсів);
– Interviews (інтерв'ю);
– Item Tracking (відстеження питань / елементів);
– Lessons Learned (винесені уроки);
– Metrics and Key Performance Indicators (KPIs) (метрики і ключові показники ефективності);
– Mind Mapping (інтелект-карта, або діаграма зв'язків);
– Non-Functional Requirements Analysis (аналіз функціональних вимог);
– Observation (спостереження);
– Organizational Modelling (організаційне моделювання);
– Prioritization (пріоритезація, розстановка пріоритетів);
– Process Analysis (аналіз процесів);
– Process Modelling (моделювання процесів);
– Prototyping (прототипування);
– Reviews (рецензування);
– Risk Analysis and Management (аналіз і управління ризиками);
– Roles and Permissions Matrix (матриця ролей і прав доступу);
– Root Cause Analysis (аналіз основних причин);
– Scope Modelling (моделювання кордонів);
– Sequence Diagrams (діаграми послідовності);
– Stakeholder List, Map or Personas (список зацікавлених сторін, карта зацікавлених сторін або діючі особи);
– State Modelling (моделювання станів);
– Survey or Questionnaire (опитування, або анкетування);
– SWOT Analysis (SWOT-аналіз);
– Use Cases and Scenarios (сценарій використання і сценарії);
– User Stories (призначені для користувача історії);
– Vendor Assessment (оцінка постачальника);
– Workshops (воркшопи, або семінари).

У загальному методологія передбачає системний та комплексний підходи, оскільки об'єкт дослідження представляється як система. Стосовно до бізнес-аналітики це означає, що різні напрями діяльності, такі як бізнес-аналітика, виробництво, фінанси тощо, розглядаються не як відокремлені один від одного, а в сукупності. Також бізнес-аналітика враховує, з одного боку, вплив різних чинників на процеси, досліджувані у ході його проведення, а з іншого – вплив зазначених процесів на інші напрями діяльності компанії.

Комплексний підхід передбачає, що досліджуваний об'єкт під час проведення аналізу розглядається з різних точок зору (естетичних, ергономічних, виробника і споживача, екологічних та ін.). Для цього в бізнес-аналітиці використовуються не тільки методи, запозичені з математики, а й методи соціології, психології, історичного аналізу тощо.

Системний підхід у разі проведення бізнес-аналітики передбачає, що всі поставлені завдання повинні бути вирішені повною мірою.

Компанії можуть різною мірою використовувати бізнес-аналітику: вони можуть зробити її необхідним елементом системи управління, а можуть використовувати окремі її елементи. Однак незалежно від цього об'єкт бізнес-аналітики являє собою систему, що функціонує і розвивається в просторі і часі.

Зазначимо, що існує перелік вимог, без врахування яких отриманий результат не можна розглядати як повний. Коротко зупинимось на основних.

1. Врахування факторів невизначеності і ризику. Невід'ємними елементами процесу прийняття управлінських рішень, метою яких є зміни в бізнесі з урахуванням вимог, виступають фактори невизначеності і ризику. Це належить як до факторів, на які компанія не може вплинути (зовнішні), так і до контрольованих факторів (внутрішні). Стає очевидною необхідність врахування чинників невизначеності і ризику в рамках проведених у ході бізнес-аналітики досліджень, що може бути забезпечено за рахунок застосування специфічних аналітичних процедур, зокрема таких, як методи імітаційного моделювання, теорії ймовірності, процедур розробки сценаріїв тощо.

2. Нормативність. Вимога нормативності бізнес-аналітики може бути реалізовано по-різному. По-перше, якщо розглядати нормативність як спрямованість на вирішення бізнес-проблем, то в цьому випадку вона означає, що цінність такого аналізу визначається ефективністю управлінських рішень, що стосуються змін у бізнесі, які приймаються на основі інформації, підготовленої у ході бізнес-аналітики. Самі по собі висновки в рамках дослідження цінності не мають, навіть якщо під час його проведення використовувалися актуальні та ефективні методи дослідження. По-друге, нормативними повинні бути процедури, які використовуються у процесі проведення бізнес-аналітики. Нормативність завдань передбачає використання адекватних підходів у рамках дослідження і відповідних процедур.

3. Безперервність. Ця вимога передбачає, що бізнес-аналітика повинна носити не випадковий (або епізодичний) характер, а проводитися на систематичній основі. Це дозволить контролювати всі ключові показники. Безперервність бізнес-аналітика передбачає і використання системи зворотних зв'язків.

4. Адекватність. Пошук даних повинен здійснюватися виходячи із характеру вирішуваних завдань. Умови ухвалення рішень також визначають характер використовуваної інформації. Вимога адекватності передбачає наявність у аналітика навичок вибору і розробки адекватних ситуацій аналітичних процедур, діагностики і прогнозування.

5. Адаптивність. Передбачається, що система бізнес-аналітики повинна змінюватися під впливом зміни бізнес-середовища як внутрішнього, так і зовнішнього, щоб бути йому адекватною. Метою такої адаптації є відповідність проведених досліджень тим змінам, що відбуваються.

1.3. Роль бізнес-аналітики і моделювання у побудові систем управлінського обліку

Сучасна ринкова ситуація ставить перед організаціями безліч серйозних проблем і, як наслідок, потребу у впровадженні якісно нової форми управління. Забезпеченню життєдіяльності організацій в умовах нестабільного зовнішнього і внутрішнього середовищ та гострої конкурентної боротьби сприятиме використання в процесі управління підприємством повноцінної системи управлінського обліку та звітності.

Перед практичним впровадженням управлінського обліку на промислових підприємствах необхідне правильне розуміння його сутності та значення в процесі управління підприємством. Згідно з міжнародним положенням 1А (SMA 1A) управлінський облік визначається як процес виявлення, вимірювання, накопичення, аналізу, підготовки, інтерпретації та передання фінансової інформації, що використовується управлінською ланкою для планування, оцінки і контролю всередині організації та забезпечення відповідного підзвітного контролю[4].

Багато зарубіжних науковців управлінський облік визначають як «...процес безперервного вдосконалення планування, проєктування, оцінки і функціонування системи фінансової і нефінансової інформації, який скеровує дії менеджера, мотивує поведінку, підтримує і створює культурні цінності, необхідні для досягнення стратегічних, тактичних і операційних цілей підприємства»[5].

Проведені дослідження сучасного стану і розвитку управлінського обліку вказують на те, що в процесі динамічного розвитку підприємств в умовах висококонкурентного середовища змінюються інформаційні потреби процесу управління підприємством. Це зумовлює виникнення нових напрямів управлінського обліку як системи інформаційного забезпечення апарату управління[6].

---

[4] Голов С. Ф. Управлінський облік. Х. : Фактор, 2009. 784 с.

[5] Аткинсон Э. А., Банкер Р. Д., Каплан Р. С., Янг М. С. Управленческий учет / пер. с англ. А. Д. Рахубовского, Д. А. Рахубовской ; под ред. А. Д. Рахубовского; 3-е изд. М. : Вильямс, 2003. 878 с.

[6] Плотніченко І. Б., Акіншина О. В. Теоретико-методологічні аспекти управлінського обліку на промислових підприємствах. Вісник Національного університету «Львівська політехніка». 2011. № 704. С. 57–54.

Управлінський облік забезпечує керівництво підприємства інформацією, яка необхідна для управління, що являє собою інформаційно-обчислювальну систему, що об'єднує сукупність форм і методів планування, обліку, контролю та аналізу, спрямовану на формування (контроль, зворотний зв'язок), інтерпретацію результатів, визначень та узагальнення базової інформації для ухвалення певних управлінських рішень і орієнтована на внутрішнього користувача – конкретного керівника підприємства, служби, відділу або функціонального підрозділу.

Сучасний управлінський облік вимагає організації стратегічного управлінського обліку, основною метою якого є надання впевненості апарату управління в правильності обраного стратегічного курсу компанії. Стратегічний управлінський облік можна розглядати як систему моніторингу реалізації стратегії підприємства. До стратегічного обліку належать методи прогнозування, а також економіко-математичні методи на базі даних фінансового та управлінського обліку (кожного окремо і загалом). Для стратегічного обліку можуть використовуватися носії інших видів обліку, але в них мають бути передбачені реквізити для прогностичного обліку.

Основними чинниками, що вплинули на зміни, які відбулися в управлінському обліку, є[7]:

– інформаційні технології;

– зміни в організаційній структурі;

– тактика орієнтації на покупця;

– новий стиль управління;

– електронний бізнес;

– процеси глобалізації, поглинання, об'єднання підприємств;

– наявність зарубіжних інвестицій;

– тактика орієнтації на якість;

– виробничі технології.

---

[7] Шайкан А. В. Бухгалтерський облік у прийнятті управлінських стратегічних рішень : монографія. К. : КНЕУ, 2009. 303 с.

Основними цілями сучасного управлінського обліку можна визначити[8]:

– оцінку ефективності діяльності підприємства;

– контроль витрат;

– управління бюджетом;

– стратегічне планування та прийняття управлінських рішень;

– впровадження бізнес-стратегії;

– максимізацію прибутку;

– генерування та створення вартості;

– інтерпретацію інформації про операційну діяльність.

Основним завданням управлінського обліку є забезпечення управлінського апарату оперативною, достовірною та максимально повною, інтегрованою інформацією для прийняття раціональних рішень. При цьому дані про витрати і результати господарської діяльності повинні розглядатися в перерізі необхідних для управління об'єктів з метою оптимізації фінансових результатів діяльності підприємства. Отже, від надання релевантних даних менеджерам залежить якість управлінського рішення, що є інструментом впливу на об'єкт управління та окремі його підсистеми, важливою ланкою формування та реалізації відношень управління в організації, а також становить основу реалізації кожної функції менеджменту.

Система управлінського обліку є інформаційно-аналітичною складовою механізму обґрунтування управлінських рішень.

Система управлінського обліку (далі – СУО) складається з трьох підсистем:

– обліку управлінської інформації;

– управлінської звітності;

– ключових показників.

До складу першої підсистеми входять такі елементи СУО: управлінський план рахунків (далі – УПР), класифікатори, кодифікатори, центри відповідальності, нормативи та ліміти.

---

[8] Плотніченко І. Б., Акіншина О. В. Теоретико-методологічні аспекти управлінського обліку на промислових підприємствах. Вісник Національного університету «Львівська політехніка». 2011. № 704. С. 57–54.

Розробка будь-якої СУО починається зі збору та аналізу інформації щодо функцій, процесів, документообігу та структури організації (компанії, підприємства). Звичайний підхід до аналізу діяльності організації припускає створення та аналіз різних моделей (функціональних, процесних, інформаційних тощо). Особливість розробки СУО полягає у необхідності застосування бізнес-аналітики з метою системного та комплексного аналізу. Це потребує використання множини різних аналітичних інструментів і типів моделей, які відображають різни сторони діяльності організації. При цьому необхідно мати можливість інтеграції результатів аналітичного дослідження в рамках загальної моделі.

У процесі бізнес-моделювання на підприємстві вирішується задача точної ідентифікації (формалізації) його діяльності. Це здійснюється шляхом створення повної (або комплексної) бізнес-моделі, що складається з декількох інших моделей. У процесі бізнес-моделювання під час розробки СУО будують такі моделі:

– стратегічну;
– організаційну;
– функціональну;
– модель бізнес-процесів;
– модель фінансової структури;
– інформаційну.

Послідовність виконання побудови повної бізнес-моделі складається з таких етапів, як:

– побудова стратегічної моделі;
– розбудова організаційної та функціональної моделей;
– побудова фінансової моделі;
– моделювання бізнес-процесів;
– побудова інформаційної моделі.

Необхідність процесів бізнес-моделювання для розробки СУО обґрунтовано таким чином. У процесі формування стратегічної моделі розроблюються стратегічні карти, що визначають коло стратегічних показників, які мають увійти до складу СУО. Від адекватності та коректності побудови стратегічної моделі залежить коректність підсистеми ключових індикаторів СУО, а також організаційної та функціональної моделей. Побудова фінансової моделі спирається на попередні дві моделі.

Ключовими елементами фінансової структури є центри фінансової відповідальності (далі – ЦФВ), які входять і до складу СУО (підсистеми обліку управлінської інформації). Крім ЦФВ, до складу СУО входять інші центри відповідальності, що визначаються завдяки організаційній та функціональній моделям.

Завдяки моделюванню бізнес-процесів визначається структура інформаційної моделі, яка включає кодифікатори, класифікатори, реєстри. Окрема підмножина елементів цієї моделі вживається з метою саме управлінського обліку, тому на цьому етапі визначаються складові елементи підсистеми обліку управлінської інформації.

Стратегічна модель будується на основі дерева цілей та стратегічних карт. Глобальна ціль організації (місія) є її орієнтиром, запланованим результатом діяльності. Основне завдання методу побудови дерева цілей – перетворення головної цілі організації (як деякої системи) у цілі окремих підрозділів (підсистем), що працюють разом на один результат. Таким чином, загальна ціль, відповідно до методології системного аналізу, розкладається на окремі складові. Наступним етапом є утворення дерева цілей організації з використанням конкретних критеріальних цілей. Після цього переходять до побудови стратегічних карт.

Моделювання організаційної структури є основою забезпечення ефективної роботи СУО, системи бюджетування. Для побудови організаційної та функціональної моделей використовуються, як правило, два типи моделей – деревоподібні і матричні. У більшості випадків організаційна модель будується відповідно до функціональної структури організації і при цьому застосовуються такі принципи:

– рівень деталізації функцій відповідає рівню деталізації організаційних ланок;

– взаємозв’язок організаційної та функціональної структури відбувається у процесі формування матриці-розподільника функцій за організаційними ланками.

Результати розробки організаційної моделі документуються.

Функціональна модель є ієрархічною, це – перелік функцій, що виконуються у компанії, в якому показано рівень підпорядкованості одних функцій іншим. Як правило, основна функція

компанії – виготовлення продукції та (або) надання послуг. Деталізація функціональних сфер проходить за методом вертикальної декомпозиції (за рівнями ієрархії) та методом горизонтального розподілу. Кожна функція має свого виконавця.

Під фінансовою моделлю у бізнес-моделюванні розуміється точний опис організації у розрізі джерел та структури формування її доходів і витрат з певним розподілом відповідальності. Перш ніж приступити до збору, обробки та оцінки управлінської інформації, важливо чітко визначити, які підрозділи надаватимуть необхідні дані. З цією метою моделюється фінансова структура, яка являє собою сукупність функціональних підрозділів компанії, пов'язаних єдиним бізнес-процесом з метою здійснення фінансового управління, ведення обліку та формування звітності, – центри фінансової відповідальності. Фінансова структура будується згідно з принципом ієрархічності.

Перший етап – визначення верхівки (ЦФВ першого рівня). Фінансові (ресурсні) потоки проходять крізь певні підрозділи, що являють собою елементи фінансової структури – центри обліку (далі – ЦО). Центри відповідальності фінансової структури розподіляються згідно з типом участі під час формування фінансових потоків на ЦФВ, центри фінансового обліку – ЦФО (центри прибутку (ЦП), центри доходу (ЦД), центри інвестицій (ЦІ), місця виникнення витрат (МВВ).

На другому рівні виділяються ускладнені сфери діяльності – напрями бізнес-діяльності, проєкти, дивізіони, підприємства (філії). Такий ускладнений ЦО може містити декілька юридичних осіб. На третьому та четвертому рівнях відокремлюються юридичні особи та всередині них – окремі департаменти, підрозділи, програми. Усередині підрозділів на наступних рівнях також можуть міститися МВВ та ЦФО – цехи, відділи тощо.

Завершаючий етап побудови фінансової моделі – закріплення за ЦО конкретних статей прибутків (витрат), основних бюджетних документів, затвердження способів консолідації даних.

Метою моделювання бізнес-процесів є систематизація інформації про компанію і її бізнес-процеси в графічній або табличній формі. Побудова процесної моделі дозволяє зібрати і систематизувати інформацію про структуру компанії та її бізнес-процесів,

що може бути використана для подальшої оптимізації. Оскільки наявність організаційно-функціональної моделі є необхідною умовою для початку процесного моделювання, то більшість методів опису бізнес-процесів починається з побудови саме неї. Отримані дані за основними параметрами бізнес-процесів (вхідний потік даних і ресурсів, вихідний результат, власник процесу, одержувач результату, ресурси, задіяні під час реалізації процесу, контрольні показники ефективності процесу та механізм реалізації) систематизуються і представляються в загальному звіті, який містить перелік бізнес-процесів організації на верхньому рівні та опис діяльності різних підрозділів.

Інформаційна модель визначає склад і структури різних документів, облікових реєстрів, звітів тощо, а також їхнього представлення у базах даних інформаційних систем. Інформаційна модель визначає інформаційні зв'язки між центрами відповідальності, власниками бізнес-процесів, інформаційні потоки, схему документообігу, формат даних.

Для управлінського обліку і бізнес-аналітики використовуються як бухгалтерські, так і будь-які інші дані, необхідні керівникам для ухвалення рішень. Такі дані можуть надходити в СУО як із внутрішніх, так і зовнішніх інформаційних джерел організації. Для уніфікації збору первісної інформації використовуються класифікатори управлінського обліку (далі – КУО), які визначають та описують різні об'єкти обліку. Класифікатори необхідні також для однозначності тлумачення та стандартизації. До КУО можна, зокрема, віднести:

– типи продукції (робіт, послуг);
– основні та допоміжні бізнес-процеси;
– проєкти;
– типи клієнтів;
– центри фінансової відповідальності;
– види витрат (доходів);
– статті калькуляції собівартості;
– види активів;
– види зобов'язань.

1.4. Поняття та сутність аналітики як основи бізнес-аналітики. Структура та завдання аналітики

Досліджуючи різні підходи до аналізу інформації, можна побачити значну кількість різних типів, видів і форм аналітики. Залежно від специфіки і контексту своїх досліджень різні аналітики вкладали в розуміння поняття цієї діяльності свій зміст.

П. Ю. Конотопов і Ю. В. Курносов зазначають, що їм довелося зіткнутися з великою варіацією думок про те, що ж таке аналітика, яка її сутність. Нижче наведено частину із запропонованих варіантів визначення поняття аналітика, а саме[9]:

– методологічна основа процесу обробки інформації;

– методологія пізнання, яка використовує як чітко наукові, так і інтуїтивні методи;

– форма мислення, яка спирається на науковий підхід;

– засіб перетворення інтуїтивних уявлень у логічний, раціональний план мислення;

– форма наукового знання, що застосовується в процесах управління, перш за все для вироблення управлінських рішень;

– сукупність методів, за допомогою яких можна виявляти прихований сенс у реальних соціально-політичних і економічних процесах;

– ядро науково-дослідницької роботи;

– інтелектуальна зброя;

– потужний пласт інтелектуальної культури, який використовується елітою будь-якого суспільства для управління соціумом;

– коротка назва інформаційно-аналітичної роботи;

– синонім діалектичної логіки;

– синонім системного аналізу;

– процес узагальнення й аналізу розрізнених, неповних і часто суперечливих даних про досліджувану ситуацію;

– процес виявлення причинно-наслідкових і просторово-часових зв’язків у будь-яких об’єктах;

---

[9] Конотопов П. Ю., Курносов Ю. В. Аналитика: методология, технология и организация информационно-аналитической работы. М. : РУСАКИ, 2004. 512 с.

– процес систематизації змісту за допомогою схематизації, конструювання та моделювання сутнісних елементів і зв'язків;

– процес поділу об'єкта на складові частини і подальшого синтетичного об'єднання їх у певну систему;

– процес виявлення суперечностей в об'єкті пізнання, зведення складного до простого;

– принцип конструктивного спрощення для виявлення форм взаємодії елементів цілого і розкриття внутрішньої структури будь-якого об'єкта вивчення.

Перелічені визначення показують, який різний зміст можна вкласти в один і той самий термін. Частина з них належить не до сутності, а до функцій аналітики та її форм. Отже, аналітика постає як дисципліна, яка об'єднує три важливих компонента процесу обробки інформації:

– методологію;

– організаційне забезпечення цього процесу;

– сукупність інструментальних засобів.

До кількості найбільш значущих, що визначають сутність аналітики, належать такі процеси:

– аналіз цілей управління і формулювання завдання аналітичної роботи;

– адаптивне управління збором інформації в інтересах вирішення управлінських завдань в умовах мінливої ситуації;

– аналіз і оцінювання отриманої інформації в контексті цілей управління, виявлення сутності процесів і явищ;

– побудова моделі предметної сфери досліджень, об'єкта досліджень і середовища його функціонування, перевірка адекватності моделі та її корекція;

– планування і проведення натурних або модельних експериментів;

– синтез нового знання (інтерпретація результатів, прогнозування тощо), необхідного для вирішення завдань управління;

– доведення результатів аналітичної роботи (нового знання) до суб'єкта управління (структури або особи, що приймає рішення (ОПР).

Перелік узагальнений, перелічені процеси є базисними, а з цієї причини він не є вичерпним. Процеси управління розглядаються перш за все тому, що будь-яка інформація використовується саме з метою управління діловими процесами в бізнесі.

Таким чином, аналітика – це перш за все основа інтелектуальної, логіко-розумової діяльності, спрямованої на вирішення практичних завдань. В її основі полягає не стільки принцип констатації фактів, скільки принцип «випередження подій», що дозволяє організації чи індивіду прогнозувати майбутній стан об'єкта аналізу, або принцип пошуку першопричин ситуації, що склалася (з'ясування причинно-наслідкових зв'язків). Тому можна стверджувати, що аналітика відіграє інтегруючу роль у реконструкції минулого, дослідженні сьогодення і прогнозуванні майбутнього.

Аналітика може виступати як спосіб організації пізнавальної діяльності, націленої на пошук і розкриття тих закономірностей і рушійних сил, які на момент початку досліджень невідомі. Це часом надає аналітичному дослідженню непередбачуваний, ризикований характер. Ризик обумовлений тим, що навіть помилкове рішення, якщо воно отримало правильну оцінку, теж стає предметом аналітичного осмислення, потенційно підґрунтям для отримання нового знання (інформації).

Аналітиками було зроблено чимало спроб визначити такі технології і прийоми, які були б придатні для всіх і давали б позитивні результати в будь-якій ситуації. Пошуки такого інструментарію досі не дали помітного результату, оскільки в принципі не існує «запатентованого» рецепту універсальної аналітики. У загальному можна стверджувати, що успішні і найбільш відомі в світі аналітики унікальні, кожен з них створював свою індивідуальну аналітичну систему обробки інформації.

Сутність аналітики як наукової дисципліни перш за все пов'язана з методологічною та інтелектуально-технологічною сторонами діяльності, спрямованих на вирішення завдань управління або синтезу нових знань. Тому можна дати таке визначення.

Аналітика – це цілісна сукупність принципів методологічного, організаційного та технологічного забезпечення індивідуальної і колективної розумової діяльності, що дозволяє ефективно обробляти інформацію з метою вдосконалення якості наявних і придбання

нових, пошуку латентних знань, а також підготовки інформаційної бази для прийняття оптимальних управлінських рішень.

Змістовний бік аналітики досить ємний і включає в себе значну кількість концептуальних підходів, ідей, окремих аналітичних систем. У сферу аналітики входять такі види аналізу:

– графічний;
– факторний;
– часових рядів;
– варіацій;
– дискримінантний;
– порівняльний;
– метод основних компонентів;
– причинно-наслідковий;
– теоретико-ігрове моделювання;
– структурний;
– ресурсний;
– кореляційний;
– дисперсійний;
– кластерний;
– ретроспективний;
– семантичний;
– аналіз показників ефективності;
– контент-аналіз, логіко-лінгвістичний тощо.

Перераховані види аналізу використовують для вирішення аналітичних задач специфічні методологічні підходи, які, в свою чергу, можуть розглядатися як самостійні аналітичні системи (метасистеми):

– канонічний математичний аналіз;
– теоретико-імовірнісний аналіз;
– статистичний аналіз;
– логічний аналіз;
– експертний аналіз;
– лінгвістичний аналіз;
– концептуальний аналіз.

Така значна кількість аналітичних систем невипадкова. Кожна з них має свої сильні і слабкі сторони і не здатна повною мірою добре оперувати з усім різноманіттям проявів об’єктів і

процесів, з якими доводиться мати справу аналітику. Тому будь-яка аналітична школа або система, що претендує на цілісне відображення реальної дійсності, не може обмежуватися деяким ізольованим набором методів. Більшість аналітичних шкіл мають подібний методологічний апарат і сильно перегукуються в частині використовуваних методів і прийомів проведення аналізу.

Будь-які методи аналізу мають різну «чутливість» щодо різних аспектів функціонування об'єктів дослідження. У результаті проведеної систематизації та експертної оцінки результативності застосування різних методів аналізу отримано залежність, що зображено на рисунку 1.1[10].

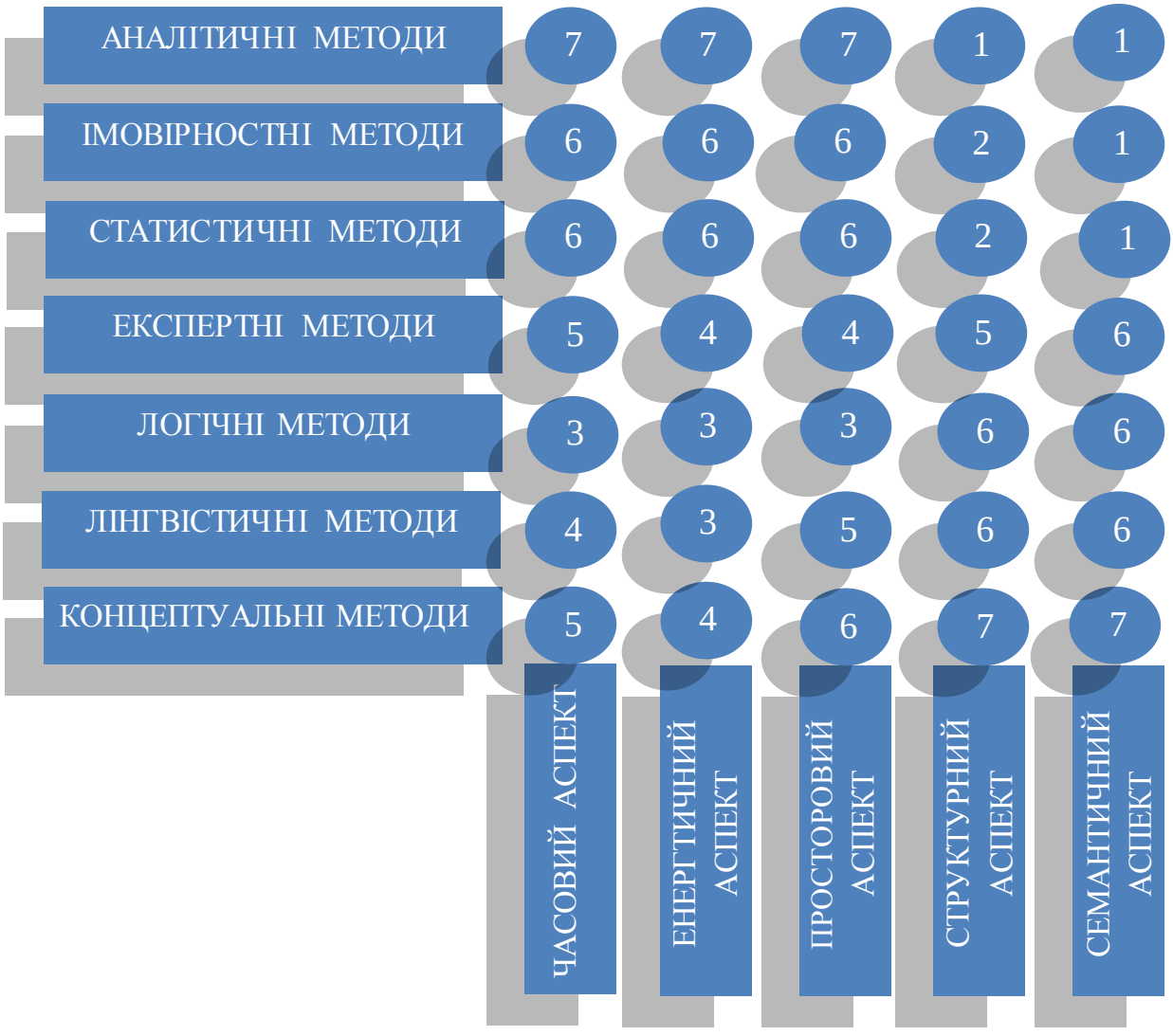


Рис. 1.1. Сприйнятливість різних методів аналізу до специфічних аспектів функціонування об'єкта

[10] Конотопов П. Ю. Аналитика: методология, технология и организация информационно-аналитической работы / П. Ю. Конотопов, Ю. В. Курносов. М. : РУСАКИ, 2004. 512 с.

Вона демонструє співвідношення різних методів аналізу і тих аспектів реальності, які розкриваються за допомогою їхнього застосування. У «таблиці» розміщені експертні оцінки в балах від 1 до 7, які вказують ступінь розкриття аспекту за допомогою певної сукупності методів.

Аналітика, як і будь-яка організована діяльність, базується на певних принципах. До основних відносять такі принципи аналітичної роботи:

1. Цілеспрямованість – орієнтація аналітичної діяльності на досягнення конкретних цілей вирішуваних завдань (результатів у практичній діяльності).

2. Системність – комплексний аналіз вирішуваних проблем з урахуванням їхнього місця, ролі і взаємозв'язків у загальній структурі забезпечення діяльності організації.

3. Актуальність – аналітична діяльність повинна витікати з потреб практики, мати високу міру важливості в цей момент, у цій ситуації, для вирішення конкретної проблеми. Крім того, дослідження можуть проводитися з питань не таких актуальних на сьогодні, але що мають перспективу розвитку, може ускладнити ситуацію в досліджуваній сфері.

4. Сучасність – отримання і видача результатів аналітичній діяльності в необхідні терміни, в зручному виді і у формі, призначеній для безпосереднього використання адресатом.

5. Активність – проведення аналітичної діяльності і видача її результатів, незалежно від конкретних запитів користувачів, з певним попередженням і елементами прогнозування. Для забезпечення ефективності дослідження потрібне визначення динаміки розвитку ситуації, що вивчається, і передбачення можливих негативних наслідків, встановлення і пояснення закономірностей зміни показників, що характеризують ситуацію, розробка сценаріїв її розвитку й експертне прогнозування.

6. Ініціативність – виявлення й опис проблем, формулювання завдань і способів їх рішення. Вироблення не лише оцінних результатів, але і конструктивних пропозицій і рекомендацій.

7. Достовірність – облік істинності початкових даних аналізу, точності використовуваних кількісних даних, міри об'єктивності й обґрунтованості висновків, оцінок, пропозицій.

8. Об’єктивність – відсутність тенденційності, неупереджене відношення аналітика до дослідження і його результатів.

9. Повнота – використання усієї наявної інформації, що належить до вирішуваних завдань. При цьому передбачається висунення і перевірка усіх можливих варіантів розвитку подій, версій про сутність і причини явища, що вивчається, визначення закономірностей його розвитку.

10. Безперервність – організація постійно діючого інформаційно-аналітичного моніторингу обстановки, своєчасно і з заданою мірою деталізації того, що відбиває основні зміни в досліджуваній ситуації (досліджуваного об’єкта і зовнішнього середовища).

11. Альтернативність – наявність у кожного співробітника аналітичного підрозділу можливості вільно висловити свою незалежну думку за результатами проведеного дослідження і довести його до вищого за рангом керівництва.

12. Гнучкість – можливість швидкої адаптації до змін соціально-економічної, суспільно-політичної обстановки без модифікації структури методів і засобів реалізації аналітичної роботи виключно за рахунок внесення змін до метазмінних.

13. Аргументованість – отримання аргументованих результатів аналітичної роботи на основі сучасних досягнень науки, ефективних інформаційних і аналітичних технологій, прагнення до об’єктивно істинного, обґрунтованого та перевіреного знання, використання усього комплексу пізнавальних принципів.

Аналітика є ядром і сутністю сучасних інтелектуальних технологій, які, в свою чергу, дозволяють з випередженням відображати дійсність, розробляти нові стратегії розвитку всіх сфер життєдіяльності суспільства, створювати ідеї проривів на вирішальних ділянках, структури й організації, що дозволяють їх реалізувати на практиці.

Структура аналітики як науки утворена системною сукупністю трьох взаємопов’язаних напрямів: методологічного, технологічного й організаційного. Тому аналітика повинна вирішувати як мінімум три класи завдань:

– методологічного плану;

– технологічного плану;

– організаційного плану.

Проте все не так просто: будь-яка наука «бореться на два фронти». Перший пов'язаний з розвитком методології, технології та організації самої науки, її власного методологічного, технологічного та організаційного забезпечення. Воно покликане забезпечити науковців і практиків у сфері бізнес-аналітики зручною формальною системою для вираження наукових суджень, сукупністю базових методів досліджень, технологією та інструментальними засобами для їхнього проведення. Для більшості наук цей напрям має безліч спільних рис, тут спостерігається феномен взаємопроникнення та взаємодоповнення наукових напрямів – методи, технології і принципи організації, розроблені в одних науках, запозичуються і знаходять широке застосування в інших. Проте кожна наука виробляє і свої суто специфічні методи та технології, які не можуть бути застосовані в рамках інших наукових напрямів.

Другий напрям пов'язаний із застосуванням досягнень науки в практичній діяльності – це її прикладний напрям. У рамках цього напряму виробляються ті методи, методики і технології, які за належної організації діяльності забезпечать споживачів, отриманих цією наукою результатів, засобами вирішення завдань, що стоять перед ними на практиці.

Існує клас наук, в яких методи, технології і принципи організації досліджень тісніше вплетені в прикладну сферу; саме до цього класу і варто відносити аналітику.

Аналітика має і свій невидимий для зовнішнього спостерігача бік, однак сфера перетину теоретичної аналітики і прикладної аналітики вміщує в себе майже повну сукупність її методів, технологій і організаційних принципів. Аналітика має і ще одну цікаву особливість: чимала частина прикладної аналітики представлена інтуїтивною компонентою аналітичної діяльності. На рис. 1.2[11] наведено ілюстрацію, що демонструє співвідношення між складовими частинами аналітики. Нижня частина сфери відповідальності прикладної аналітики, яка не увійшла в перетин, за

[11] Конотопов П. Ю., Курносов Ю. В. Аналитика: методология, технология и организация информационно-аналитической работы. М. : РУСАКИ, 2004. 512 с.

задумом авторів, і покликана відобразити ще не освоєну наукою галузь аналітики. Тут варто розглядати і так звані трансові методики, і методики, пов'язані із застосуванням гіпнотичного впливу, які досить часто використовуються у специфічних галузях досліджень.

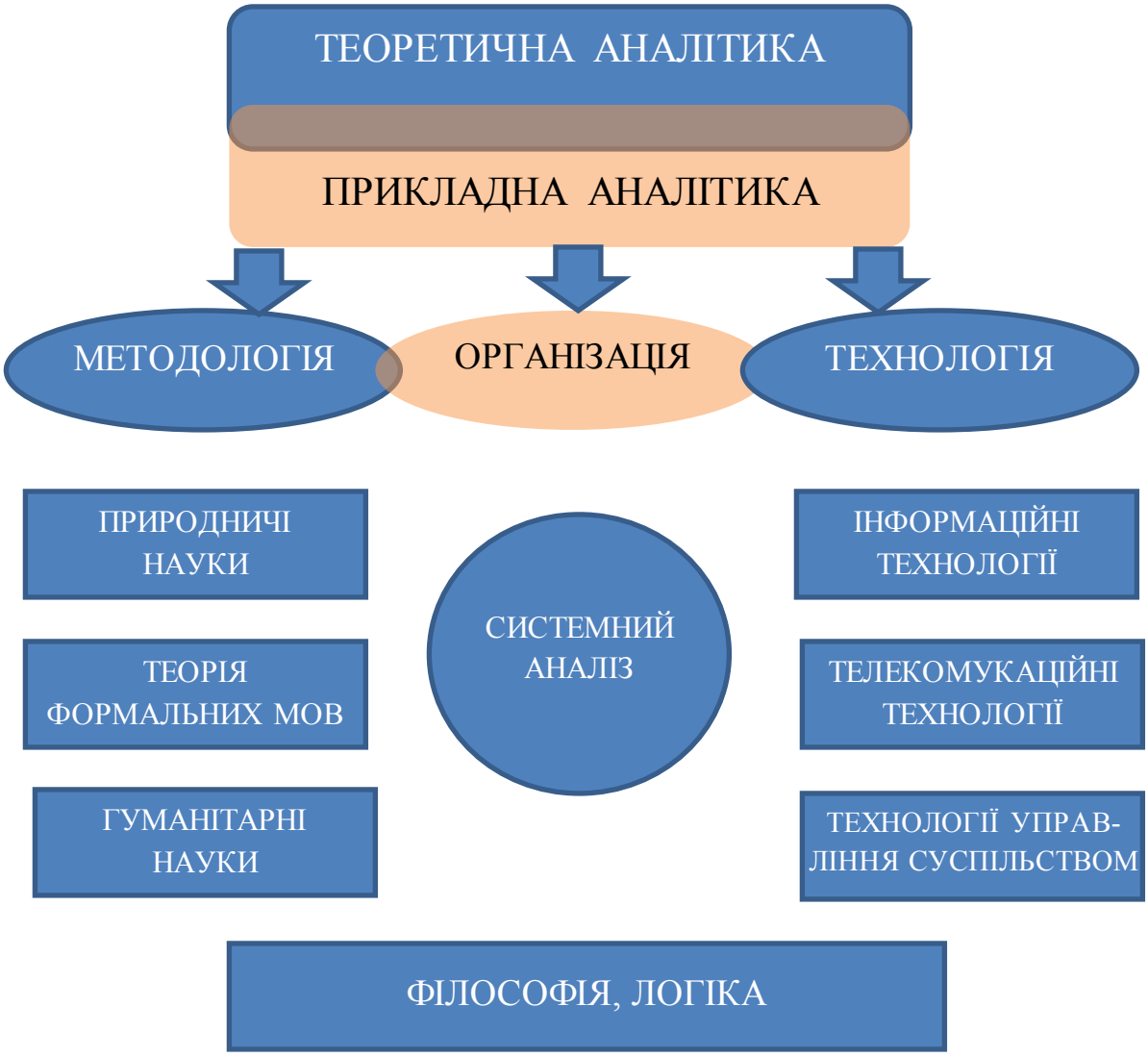


Рис. 1.2. Структура аналітики і суміжні галузі знання

Найпотужнішим інструментом під час ведення аналітичної роботи є системний аналіз. Теоретичні основи цієї методологічної системи закладені на початку 1950-х років Людвігом фон Берталанфі в його роботах із загальної теорії систем (далі – ЗТС). Перший варіант ЗТС був запропонований 1912 року А. А. Богдановим. Варіант ЗТС, запропонований Берталанфі, привернув увагу широких кіл міжнародної наукової громадськості. Проте він не тотожний ЗТС Берталанфи, наприклад, не вдалося охопити різні концептуальні системи, обмежено і його розуміння системи.

Аналітика повинна бути здатна до узагальнень, прогнозів, пояснень, постановки нових питань, зв'язків з найважливішими науковими теоріями і принципами, ці властивості їй додає саме системність.

Методологічні основи системного аналізу стають більш «прозорими», якщо розглянути рівні термінологічної ізоляції сутностей, якими оперує ця теорія. Мовою цієї теорії поставлена проблема, пов'язана з визначенням умов існування системи «аналітичний центр – об'єкти аналізу + довкілля» та виявленням певного значення реляційного й атрибутивного системних параметрів.

Реляційний системний параметр – це набір відносин таких, що будь-які системи перебувають у будь-якому відношенні з цього набора. Атрибутивний системний параметр – це набір таких властивостей, одним з яких володіє будь-яка система. Будь-яке з цих властивостей є одним зі значень атрибутивного системного параметра.

Класичний підхід у застосуванні наукових методів до вирішення проблем полягає в тому, що базовий алгоритм включає три етапи: перший – постановка задачі та збір вихідної інформації; другий – висування гіпотези; третій – перевірка цієї гіпотези.

Дослідження різного класу систем з метою виявлення спільності їхніх структур, створення загальної теорії, справедливої для всіх систем цього класу, є характерною рисою сучасного етапу розвитку науки.

Часто з поняттям системи асоціюється поняття структури, іноді поняття системи і структури навіть ототожнюють. Проте структура – це певна сторона, аспект системи. Про структуру можна свідчити в тому випадку, якщо повністю відволіктися від сутності елементів системи, зробити основним предметом аналізу відносини між її елементами. Дослідження структури й у відверненні від субстрату системи показало свою плідність у структурній лінгвістиці, топології і ряді інших наук.

У процесі розгляду систем необхідно розуміти важливість розрізнення внутрішніх і зовнішніх відносин, а також усвідомлювати відносну цінність будь-якої класифікації.

Аналіз тенденцій розвитку сучасних інформаційних технологій показує, що одним з основних напрямів є напрям, пов'язаний з наданням створюваним програмним і технічним системам інтелектуальних властивостей. Інструментарій сучасного аналітика збагачується засобами, які враховують специфіку організації розумового процесу, адаптуються до потреб аналітика. У стадії формування перебувають технологічні засоби, що здійснюють моніторинг пошукових запитів до інформаційних ресурсів з метою проведення ініціативних інформаційно-пошукових процедур в інтересах інтенсифікації аналітичної роботи. Однак очевидно, що ці засоби не мають змоги самостійно здійснювати аналітичну роботу, вони безпосередньо залежать від професійних якостей аналітиків і їхнього теоретичного та методологічного озброєння.

Запитання до параграфа 1

1. Що таке бізнес-аналітика?
2. Яка сфера застосування бізнес-аналітики?
3. Хто такий бізнес-аналітик?
4. Які основні завдання бізнес-аналітики?
5. Що таке аналіз вимог?
6. Які існують сфери бізнес-аналітики?
7. Які існують асоціації бізнес-аналітиків?
8. Що таке BABOK?
9. На які сфери знань BABOK розділяє діяльність бізнес-аналітиків?
10. Яка основна мета BABOK?
11. Що включає в себе діяльність бізнес-аналітика за версією BABOK?
12. Які можна виділити три сфери, взаємодію з якими здійснює бізнес-аналітик у рамках вирішення поставлених перед ним завдань?
13. На що орієнтований, у першу чергу, бізнес-аналітик?
14. Що є предметом бізнес-аналітики?
15. Які існують методи в бізнес-аналітиці?
16. Сутність комплексного підходу в бізнес-аналітиці.
17. Що входить у перелік вимог, без врахування яких отриманий результат не можна розглядати як повний?

18. Яка роль управлінського обліку в бізнес-аналітиці?

19. Які основні чинники вплинули на зміни, які відбулися в управлінському обліку?

20. Які основні цілі сучасного управлінського обліку?

21. Які основні завдання управлінського обліку?

22. З яких підсистем складається система управлінського обліку?

23. Які моделі будують у процесі бізнес-моделювання під час розробки систем управлінського обліку?

24. Яка послідовність виконання побудови повної бізнес-моделі?

25. Яка мета моделювання бізнес-процесів?

26. Які дані використовуються для бізнес-аналітики?

27. Що таке аналітика?

28. Які процеси визначають сутність аналітики?

29. Які види аналізу входять у сферу аналітики?

30. Які є принципи аналітичної діяльності?

§ 2. Сутність системного підходу як методологічної основи бізнес-аналітики

2.1. Сутність системного аналізу

Вивчаючи будь-яку наукову дисципліну, в першу чергу, необхідно виділити та чітко визначити її об'єкт і предмет. Об'єкт наукової дисципліни – це те, на що направлене вивчення, вибрана для вивчення частина зовнішньої реальності. Об'єктом теорії систем та системного аналізу є матеріальні й ідеальні системи навколишнього світу.

Предмет науки (предметна сфера) – це ті ознаки, зв'язки, відношення об'єкта, які вивчаються певною наукою. Предметом теорії систем та системного аналізу є вивчення великих систем навколишнього світу на основі системного підходу, вивчення внутрішніх і зовнішніх, найбільш загальних характеристик систем, розробка методів аналізу систем та методів вирішення проблем, що виникають у цих системах під час практичної діяльності людини.

Системний аналіз є методологічною науковою дисципліною, що означає: основним напрямом є розробка методів аналізу складних систем. Системний аналіз – це методологічна дисципліна, заснована на системному підході. Вона об'єднує методи вивчення систем різної складності й призначення, розробляє ці методи, узагальнює їх, дає практичні рекомендації для їх використання. Системний аналіз застосовується для підготовки й обґрунтування шляхів вирішення складних проблем політичного, соціального, військового, економічного, технічного характеру тощо.

Основна процедура системного аналізу – побудова узагальнених моделей, в яких відображені закономірності реальної ситуації. Моделі системного аналізу відображають структуру, взаємозв'язки у складних системах, реальну ситуацію та проблеми, які в них виникають. За допомогою створених моделей досліджують системи й знаходять шляхи вирішення складних проблем практичної діяльності.

Технічна основа системного аналізу – інформаційні системи, обчислювальна техніка і сучасні методи керування.

Системний аналіз вивчає такі питання:

– утворення цілого;

– побудова цілого;

– зростання і розвиток цілого;

– відношення між цілісною системою та іншими системами;

– відношення між системою та метасистемою, великою зовнішньою системою, до складу якої вона входить.

Цілями системного аналізу є вивчення загальних закономірностей складних систем різної природи й характеру. Практична направленість системного аналізу полягає у вирішенні непростих проблем.

Спеціалісти із системного аналізу вивчають процеси у складних системах, аналізують результати, до яких приводять ці процеси, оцінюють перспективи розвитку систем. Їх завданням є прогнозування можливих сценаріїв розвитку, запобігання небажаним явищам та катастрофам. Системний аналіз є основою прийняття рішення у складних ситуаціях, коли ефективність цих рішень неоднозначна й оцінити її важко.

Отже, системний аналіз розробляє наукові методи вирішення складних проблем у системах навколишнього світу. Він виступає як каркас, що об'єднує методи і знання для вирішення проблем. Основними методами системного аналізу є побудова моделей систем різного рівня і складності та аналіз систем за допомогою побудованих моделей.

Методи системного аналізу широко використовують у: бізнес-аналітиці та моделюванні; вивченні ринків; керуванні й зв'язку; організації виробництва; керуванні матеріальними і інформаційними потоками; запасами і резервами та інших напрямах економічної діяльності.

Системний аналіз засновується на системному підході, а також на ряді математичних дисциплін та сучасних методах керування.

Системний підхід – це напрям дослідження і вивчення світу, в основу якого покладено розгляд об'єктів як системи, орієнтований на розкриття цілісності об'єкта, виявлення різноманітності зв'язків у ньому і приведення їх до єдиної теоретичної картини. Основними принципами системного підходу є:

1. Принцип взаємозв'язку – система вивчається як частина певної макросистеми. Вона пов'язана безліччю зв'язків з іншими системами, взаємодіє та існує в єдності з ними.

2. Принцип багатоплановості – система як деяка самостійна одиниця вивчається з різних сторін зі своїми особливостями.

3. Принцип багатомірності, який полягає в тому, що вивчаються різні характеристики систем, які об'єднують у групи (кластери): об'єкт описується як сукупність деяких характеристик та взаємозв'язків між ними.

4. Принцип ієрархічності – система розглядається як складна структура з різними рівнями, між якими встановлюються певні зв'язки.

5. Принцип різнопорядковості – полягає у тому, що різні ієрархічні рівні системи породжують закономірності різного порядку. Одні закономірності властиві тільки всім елементам або деякій групі елементів, а інші – тільки окремим елементам.

6. Принцип динамічності – система розглядається в русі й розвитку.

Системний аналіз (далі – СА) – це сукупність методів і засобів, які використовують під час дослідження та проєктування складних технічних, економічних, організаційних структур. Теоретичну і методологічну основу СА складають системний підхід та загальна теорія систем. Методика системного аналізу розробляється і застосовується в тих випадках, коли немає достатньої інформації для формалізованого представлення досліджуваної системи.

Основні етапи СА:

1. Формування варіантів представлення системи:

– відокремлення системи від навколишнього середовища;

– вибір підходу до представлення системи;

– формування варіантів представлення системи.

2. Вибір найкращого варіанта представлення системи:

– вибір підходу до оцінки варіантів;

– вибір критеріїв оцінки та обмежень;

– проведення оцінки;

– обробка результатів оцінки;

– аналіз одержаних результатів і вибір найкращого варіанта (або коректування варіанта, якщо він був один).

Евристичні методи СА застосовуються на етапі формалізації задачі, формування варіантів, вибору підходу до оцінки, використання досвіду фахівців, який не завжди може бути виражений у кількісних оцінках.

Кількісні методи СА висувають на перший план кількісні характеристики системи й оцінки їх коректності, точності, похибок. Кількісні методи СА – це методи сучасної математики, починаючи від розрахункових формул, методів формалізованого представлення систем до алгоритмічних моделей.

З точки зору опису побудови і функціонування систем розрізняють загальні методи СА: декомпозиції та агрегації. Ці методи застосовуються на всіх етапах дослідження поведінки системи (декомпозиція – в напрямі мікроскопічної точки зору на поведінку системи, агрегація – макроскопічної).

Важливою особливістю СА є поєднання формалізованих і неформалізованих засобів і методів дослідження з використанням

різних концепцій та термінологій, що призвело до виникнення, наприклад:

– кібернетичного підходу до розробки адаптивних систем керування;

– інформаційно-гносеологічного підходу до моделювання систем, заснованого на спільності процесів відображення і пізнання в системах різної фізичної природи;

– системно-структурного підходу, який базується на сполученні методів структуризації з лінгвістичними представленнями і поняттям шкали сучасної теорії вимірювання.

Загальним для всіх методик системного аналізу є формування варіантів подання системи (процесу розв'язання задачі) та вибір кращого варіанта.

У загальному вигляді системне дослідження складається з таких етапів:

– формулювання проблеми;
– виявлення цілей;
– формулювання критеріїв;
– визначення наявних ресурсів для досягнення цілей;
– генерація альтернатив та сценаріїв.

Таблиця 1.1

Послідовність етапів системного аналізу

Назва етапу	**Зміст виконаних робіт**
1	**2**
Аналіз проблеми	Точне формулювання проблеми. Аналіз структури проблеми. Зовнішні зв'язки проблеми
Формулювання завдань	Формулювання конкретних завдань виходячи з проблеми
Визначення системи	Виділення елементів. Визначення зовнішнього середовища
Аналіз структури системи	Визначення рівнів ієрархії. Виділення підсистем. Визначення функціональних і структурних зв'язків
Формулювання загальної мети та критеріїв системи	Визначення цілей. Визначення критеріїв. Декомпозиція критеріїв за підсистемами

Продовження таблиці 1.1

1	2
Виявлення ресурсів, композиція цілей	Оцінювання існуючої технології і виробничих потужностей. Оцінювання теперішнього стану ресурсів. Оцінювання можливостей взаємодії з іншими системами
Оцінювання цілей і засобів	Обчислення оцінок за критерієм. Оцінювання взаємозалежності цілей. Оцінювання вартості ресурсів. Оцінювання впливу зовнішніх факторів. Обчислення комплексних розрахункових оцінок
Вибір варіантів	Розроблення варіантів досягнення окремих цілей. Оцінювання і порівняння варіантів. Синтез комплексу взаємозалежних варіантів
Реалізація варіантів	Моделювання економічного (технологічного) процесу. Проєктування організаційної структури. Проєктування інформаційних механізмів. Виявлення та аналіз заходів щодо удосконалення організації

2.2. Формування проблеми та виявлення цілей

В основу методології системного аналізу покладені три концепції: проблема, вирішення проблеми і система.

Проблема – це невідповідність або відмінність між існуючим і необхідним станом справ певній системі.

Необхідний стан може виступати як потрібне або бажане. Необхідний стан диктується об’єктивними умовами, а бажане визначається суб’єктивними передумовами, в основі яких полягають об’єктивні умови функціонування системи.

Проблеми, що існують в одній системі, як правило, не рівнозначні. Для порівняння проблем, визначення їх пріоритету використовуються атрибути: важливість, масштаб, спільність, актуальність тощо.

Виявлення проблеми здійснюється шляхом ідентифікації симптомів, що визначають невідповідність системи своєму призначенню або недостатню її ефективність. Симптоми, що систематично виявляються, утворюють тенденцію.

Ідентифікація симптомів проводиться шляхом вимірювання й аналізу різних показників системи, нормальне значення яких відоме. Відхилення показника від норми і є симптомом.

Вирішення проблеми полягає в ліквідації відмінностей між існуючим і необхідним станом системи. Ліквідація відмінностей може проводитися або шляхом удосконалення системи, або шляхом її заміни на нову.

Після постановки проблеми наступним кроком системного аналізу є визначення цілей вирішення проблеми. Визначення цілей означає, що проблема повинна бути приведена до такого вигляду, коли чітко поставлені цілі або завдання, які дозволяють повністю чи частково вирішити проблему.

Труднощі, що виникають під час визначення цілей, зумовлені тим, що ціль є іншим, зворотним боком проблеми. Формулюючи проблему, у нас уже наявна система, яка її містить, і варто вказати на те, що нас не влаштовує. Під час визначення цілей треба встановити, чого хочемо досягнути, до чого потрібно прагнути, який буде ідеальний результат вирішення проблеми.

Ціллю називають результат або стан, якого прагнуть досягнути.

Розрізняють об'єктивні й суб'єктивні цілі. Суб'єктивна ціль – це образ бажаного майбутнього стану або результату діяльності системи відповідно до її призначення. Суб'єктивні цілі визначаються людиною (суб'єктом), як правило, для штучних або змішаних систем. Об'єктивна ціль – це майбутній реальний стан системи, якого вона досягне, рухаючись траєкторією свого життєвого циклу.

Проблема у своїй постановці передбачає досягнення ряду цілей. Ціль, яка може бути сформульована у постановці проблеми, має, як правило, загальний, дещо розпливчастий, декларативний характер.

Сформулювати ціль – означає визначити, що треба зробити для вирішення проблеми повністю чи частково. Складність формулювання цілей має ряд причин. Серед цих причин можна виділити такі:

– суб'єктивний характер цілей;

– складність та ієрархія цілей;

– взаємозумовленість цілей;

– зв’язок цілей з можливістю їх досягнення.

Складність та ієрархія цілей полягає у тому, що будь-яка проблема має ряд цілей різного рівня. Вона зумовлена тим, що будь-яка проблема виникає в системі, що входить у більш загальну метасистему і містить у собі підсистеми. Цілі можуть бути сформульовані для кожного ієрархічного рівня.

Взаємозумовленість цілей полягає у тому, що кожна ціль не самостійна, не відірвана від інших, а взаємозв’язана з ними. Досягнення однієї цілі змінює можливість досягнення інших цілей. У разі вирішення однієї проблеми (досягнення однієї цілі) система змінюється, стає іншою.

Зв’язок цілей із засобами і можливістю їх досягнення полягає в тому, щоб були поставлені такі цілі, які можна досягнути наявними засобами за визначений час. До засобів досягнення цілей відносять матеріальні, фінансові, енергетичні, інформаційні та інтелектуальні ресурси.

Під час визначення цілей важливо враховувати ситуації, коли досягнення цілей не приведе до вирішення проблеми. Це може бути в результаті:

– підміни цілей засобами їх досягнення;

– змішування цілей;

– зміни цілей упродовж часу.

Підміна цілей засобами їх досягнення є однією з небезпек, яка може призвести не до вирішення проблеми, а до її загострення. Це буває тоді, коли засоби досягнення приймаються за самостійні цілі.

Метод побудови дерева цілей. Під деревом цілей розуміють ієрархічну деревоподібну структуру, яка отримується поділом загальної цілі на підцілі, а їх, у свою чергу, – на детальніші складові: нові підцілі, функції тощо. Якщо всі ці елементи зобразити графічно, то одержимо «дерево цілей», повернуте «кроною» донизу (рис. 1.3). При цьому основну ціль розміщують на найвищому рівні.

Перевагою цього методу є те, що він дозволяє поділ складного завдання, яке важко формалізувати, на сукупність простіших завдань, для розв’язання яких існують перевірені прийоми і методи. Послідовний поділ розв’язуваної проблеми на її частини –

підпроблеми – є важливим етапом системного аналізу проблем. Поділ продовжують доти, доки не отримають прості, звичні, очевидні завдання, які можна розв'язати відомими методами.

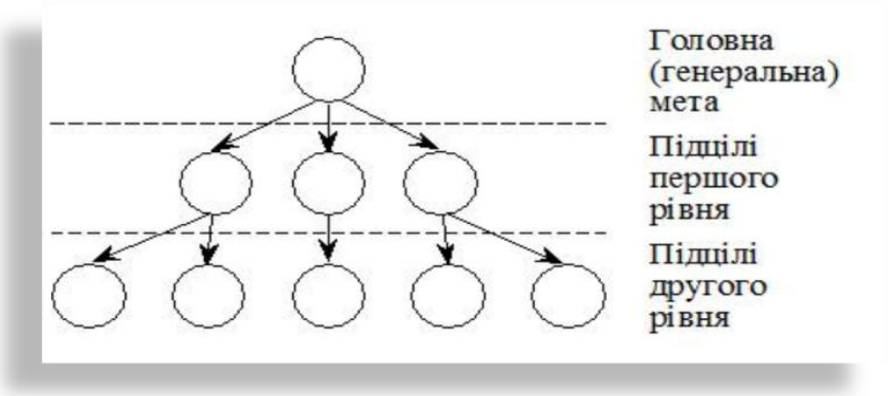


Рис. 1.3. Граф дерева цілей

Метод побудови дерева цілей являє собою один із найрозповсюдженіших та найефективніших способів аналізу слабоструктурованих завдань, що стоять перед економічними об'єктами. Він допомагає знаходити найкращі шляхи та засоби вирішення існуючих проблем. Деревоподібні ієрархічні структури використовуються і під час дослідження та удосконалення організаційних структур.

Коли будують дерево цілей, необхідно, з одного боку, здійснювати дослідження цілей зацікавленими у розв'язанні проблеми сторонами, а з іншого – передбачати можливість уточнення цілей, їх розширення або зміни. У цьому полягає одна з основних причин ітеративності системного аналізу.

Отже, у рамках цього етапу треба:

– чітко визначити цілі, досягнення яких сприяє вирішенню виявленої проблеми;

– виявити інформацію про параметри системи та зовнішнього середовища, які необхідно враховувати;

– визначити сукупність допущень та обмежень, у рамках яких буде розв'язуватись проблема.

На наступному етапі визначаються критерії та обмеження. Під критеріями розуміють кількісні показники якісних цілей, які мають точніше їх характеризувати. Критерії мають якомога точніше відповідати цілям, хоча і не можуть повністю збігатися з ними, оскільки вони фіксуються в різних шкалах вимірювання: цілі – у номінальних, а критерії – у шкалах, що передбачають упорядкування.

Під час формування критеріїв основним є не їх кількість, а те, наскільки повно вони характеризують ціль. Тому тут прагнуть досягти компромісу між повнотою опису цілей та кількістю критеріїв. Для повноти опису проблемної ситуації необхідно розглядати три взаємодіючі системи:

– систему, в якій існуюча ситуація розглядається як проблема;

– систему, в рамках якої можна вплинути на проблему для її вирішення;

– зовнішнє середовище, в якому існують та з яким взаємодіють ці дві системи.

Варто враховувати, що характер цілей цих трьох систем істотно відрізняється: для першої системи необхідно розв'язати проблему, для другої основна мета полягає у розв'язанні проблеми з найменшими витратами ресурсів, при цьому потрібно враховувати вплив зовнішнього середовища.

Початкові, переважно неформальні етапи системного дослідження проблеми подано на рис. 1.4.



Рис. 1.4. Основні етапи проведення системних досліджень

2.3. Методи системного аналізу

У філософському розумінні аналіз і синтез є методами пізнання дійсності. Сутність аналітичного методу пізнання полягає у поділі цілого на частини, в поданні складного у вигляді сукупності простіших компонентів та дослідженні властивостей цих компонентів. Потім знання про частини агрегується в знання про систему в цілому. Але під час поділу системи в процесі аналізу можуть втрачатися суттєві властивості як самої системи, так і окремих відділених від неї частин. Це обумовлено такими важливими властивостями систем, як цілісність та емерджентність.

Синтетичний метод (синтез) пізнання полягає в об'єднанні частин у ціле. Проте синтез не зводиться лише до «механічного збирання» частин, що були одержані шляхом аналізу. При цьому підході систему необхідно розглядати як складову більшої системи (надсистеми) та, дослідивши її, дезагрегувати знання про неї для пояснення частин. Це досягається вивченням значення та функцій частин у цілому. У такий спосіб аналіз і синтез доповнюють один одного.

Операції поділу цілого на частини та їх з'єднання у ціле називають, відповідно, *декомпозицією* та *агрегуванням.*

У вузькому розумінні *аналіз* системи полягає в її декомпозиції з подальшим визначенням статичних та динамічних характеристик її елементів, що розглядаються у взаємодії з іншими елементами системи та зовнішнім середовищем. *Синтез* системи полягає в її створенні (проєктуванні, організації, оптимізації) через визначення статичних та динамічних характеристик, що мають забезпечувати у сукупності максимальну відповідність системи поставленим завданням.

Розглянемо основні завдання, що вирішуються за допомогою аналізу та синтезу систем. На етапі декомпозиції системи здійснюється:

– визначення та декомпозиція загальної мети дослідження та основної функції системи як обмеження траєкторії в просторі станів системи або у сфері допустимих ситуацій. Найчастіше декомпозицію виконують побудовою дерева цілей та дерева функцій;

– виділення системи із середовища (поділ на «систему» та «несистему»);

– опис впливових факторів;

– опис тенденцій розвитку;

– опис системи як «чорного ящика»;

– функціональна (за функціями), компонентна (за типом елементів), структурна (за типом відношень між елементами) декомпозиція системи.

Глибина декомпозиції – кількість рівнів дерева цілей, що визначається метою дослідження системи.

Аналіз та синтез систем можуть здійснюватися у таких аспектах:

– структурному;

– функціональному;

– інформаційному;

– параметричному.

Структурний аналіз проводиться з метою дослідження статичних характеристик системи виділенням у ній підсистем та елементів різного рівня і зв'язків між ними. Тобто об'єктами дослідження структурного аналізу є різні можливі варіанти структури системи. Метою структурного синтезу є розробка (створення, проєктування, реорганізація, оптимізація) системи, яка повинна мати певні властивості. Структурний синтез виконується для обґрунтування множини елементів структури, відношень та зв'язків, які б забезпечували в сукупності максимальну відповідність заданим властивостям.

Сутністю *функціонального аналізу* є визначення динамічних характеристик системи через дослідження процесів зміни її станів з часом на підставі прийнятих алгоритмів (способів, методів, принципів) її функціонування. У межах функціонального аналізу досліджуються алгоритми та методи управління системою, включаючи загальний закон функціонування, що містить усі основні етапи та функції управління (формулювання цілі управління, збір та оброблення необхідної інформації, ухвалення рішень, планування, організація, контроль, виконання рішень тощо). Метою функціонального синтезу є обґрунтування оптимальних характеристик процесів функціонування системи, тобто її станів у майбутньому відповідно до поставлених перед системою цілей.

Інформаційний аналіз спрямований на дослідження якісних та кількісних характеристик інформаційних процесів у системі. При цьому вивчають:

– збір та сприйняття інформації (ці процеси характеризують взаємодію системи із зовнішнім середовищем);

– обмін інформацією між окремими підсистемами;

– аналіз, оброблення, створення нової інформації;

– використання інформації;

– обмін інформацією із зовнішнім середовищем.

Завданням інформаційного синтезу є обґрунтування необхідного обсягу та форм подання інформації, методів та засобів її передавання, оброблення, зберігання. Інформаційний синтез доповнює завдання інформаційного аналізу, що здійснюється з метою визначення необхідних кількісних та якісних характеристик інформації, яка використовується у процесі функціонування системи.

Параметричний аналіз полягає у визначенні необхідної та достатньої сукупності узагальнених і часткових показників, що утворюють ієрархічну структуру та мають характеризувати найсуттєвіші властивості системи. Сутністю параметричного синтезу є обґрунтування необхідної та достатньої сукупності показників, що уможливлюють оцінювання бажаних властивостей системи, яка створюється, та її загальну ефективність.

Системний аналіз застосовує евристичні методи. Існує кілька методів, які застосовують для цього:

– метод мозкового штурму;

– метод Дельфі;

– синектика;

– сценарний аналіз;

– ділові ігри.

Метод мозкового штурму – це метод посилення творчого підходу стимулюванням генерування ідей у процесі їх обговорення групою людей, під час якого забороняється критика. Мета цього методу полягає у стимулюванні висловлення ідей через заохочення ініціативи учасників. При цьому передбачається дотримання таких правил:

– жодна ідея не вважається безглуздою, і тому членів групи заохочують висловлювати довільні крайнощі та неймовірні ідеї;

– кожна з висловлених ідей належить колективу, а не особі, що запропонувала її. Тому кожен член групи використовує ідеї інших;

– жодна з ідей не піддається критиці, тому що основна мета – породжувати, а не оцінювати ідеї.

Метод мозкового штурму широко застосовується в рекламній діяльності та деяких інших сферах, де він, напевно, найефективніший.

Метод Дельфі передбачає одержання та зіставлення анонімних суджень про питання, яке становить для нас інтерес, через послідовне розсилання анкет, що перемежовується з обробленням отриманої інформації. Під час методу Дельфі зберігаються переваги наявності кількох суджень і водночас усувається ефект зміщених оцінок, який можливий за особистої взаємодії респондентів. Основа методу – збір поштових анкет. Наприклад, учасники опитування відповідають на першу анкету та відсилають її. Спеціалісти узагальнюють відповіді, визначаючи груповий консенсус, та відправляють цей результат респондентам разом із другою анкетою для переоцінки своїх попередніх відповідей. Основна ідея цього методу полягає в тому, що консенсус приводить до кращого розв'язку після кількох раундів опитування. Але, як свідчать дослідження, досить часто значні зміни не відбуваються вже після другого раунду.

Синектика призначена для генерування альтернатив через пошук аналогій до поставленого завдання за допомогою асоціативного мислення. На відміну від мозкового штурму основною метою тут є генерування невеликої кількості альтернатив. Для цього формується група з 5–7 осіб, які характеризуються гнучкістю мислення, широким світоглядом та практичним досвідом у різних сферах діяльності, психологічною сумісністю тощо. Після набуття певного досвіду спільної роботи група починає цілеспрямоване систематичне обговорення довільних (можливо, і фантастичних) аналогій, що виникають стосовно проблеми, яка розглядається.

Особливе значення синектика надає аналогіям, які пов'язані з відчуттям рухів, що обумовлено високою організацією наших рухових рефлексів, і їх осмислення може підказати корисну нестандартну ідею. Для успішної роботи, так само як і при мозковому штурмі необхідно дотримуватись деяких правил: забороняється обговорювати недоліки та переваги окремих членів групи, кожний має право припинити роботу без жодних пояснень, роль ведучого постійно переходить до інших членів групи. Але, на відміну від мозкового штурму, у разі застосування синектики необхідна спеціальна та тривала підготовка.

Розробка сценаріїв – це метод генерації альтернатив за допомогою аналізу ймовірних шляхів розвитку або поведінки системи у майбутньому. Отже, сценарій являє собою певний варіант можливого розвитку подій, деякий логічно обґрунтований прогноз, який з певною ймовірністю реалізується після прийняття рішення. Корисно розробляти кілька варіантів сценаріїв, як правило, песимістичних та оптимістичних, у межах яких найімовірнішим є розвиток майбутніх подій.

До сценаріїв відносять не тільки змістовні міркування, що дають змогу не втратити деякі важливі деталі, які не завжди враховуються у разі формального описування системи, а й результати кількісного техніко-економічного або статистичного аналізу із попередніми висновками, які можна одержати на їх підставі.

Отже, сценарій є попередньою інформацією, на основі якої виконується подальша робота з прогнозування чи розробки варіанта проєкту. Він дає можливість уявити проблему, а потім перейти до формального зображення системи (графіки, таблиці тощо) для проведення експертного опитування та інших методів системного аналізу.

Метод експертних оцінок полягає в опитуванні групи фахівців з метою з'ясування їхньої думки стосовно досліджуваної проблеми. У разі застосування цього методу вважається, що думка групи експертів надійніша, ніж думка окремого експерта. Він ґрунтується на тому, що невідома характеристика досліджуваного явища є випадковою величиною, а індивідуальна оцінка кожного експерта щодо істинності та значущості певної події є відображенням її закону розподілу.

Під час оброблення результатів колективної експертної оцінки застосовують методи теорії рангової кореляції. Для кількісного оцінювання рівня узгодженості оцінок експертів використовують коефіцієнт конкордації. Він дає змогу оцінити, наскільки узгоджені між собою ряди переваг, що побудовані кожним експертом. Його значення перебувають у межах $0 < K < 1$. Тіснота зв'язку вважається непоганою, якщо K полягає у межах від 0,7 до 1. Невелике значення цього коефіцієнта свідчить про слабку узгодженість думок експертів щодо досліджуваної проблеми. Це може бути зумовлено відсутністю єдності поглядів або в групі експертів є підгрупи з високою узгодженістю думок, але погляди деяких інших підгруп експертів протилежні.

Під *діловими іграми* розуміють імітаційне моделювання реальних ситуацій, за якого учасники гри поводять себе так, як би вони діяли в реальній ситуації. При цьому реальність замінюють певною моделлю. І хоча такі ігри найчастіше використовують для навчання (наприклад, робота за тренажерами під час підготовки пілотів, штабні ігри чи навчання військових або різноманітні тренінги та бізнес-табори, які стали популярними останнім часом), їх можна використовувати і для генерування альтернатив, особливо у ситуаціях, що важко піддаються формалізації. Важлива роль при цьому надається контрольно-арбітражним групам, що керують моделлю, реєструють хід гри та узагальнюють результати.

Індукція – висновок від фактів до деякої гіпотези (загального твердження), метод міркування від часткового до загального.

Дедукція – метод мислення, при якому часткове положення логічним шляхом виводиться із загального, висновок за правилами логіки; ланцюг висновків (міркування), ланки якого (вислови) пов'язані відношенням логічного проходження.

Метод міркування від загального до часткового. Початком дедукції є аксіоми, постулати або просто гіпотези, що мають характер загальних тверджень, а кінцем – наслідки з посилок, теорем. Якщо посилки дедукції істинні, то істинні і її наслідки.

Формалізація:

– процес подання інформації про об'єкт, процес, явище у формалізованому вигляді;

– метод відображення певної сфери у вигляді формальної системи, коли форма виділяється в особливий предмет дослідження незалежно від змісту.

За допомогою формалізації відбувається відображення результатів мислення в точних поняттях або твердженнях. Вона протиставляється інтуїтивному мисленню і пов'язана з процесом абстракції. Формалізація відіграє істотну роль в аналізі й уточненні наукових понять. Невідповідність між формалізацією і змістовним знанням є важливим джерелом розвитку, приводить до послідовної зміни одних елементів формалізму іншими, що точніше відображають досліджувані об'єкти і явища.

Лінеаризація – метод наближеного подання замкнутих нелінійних систем, під час якого дослідження нелінійної системи замінюється аналізом лінійної системи, в деякому розумінні еквівалентної початкової. Методи лінеаризації мають обмежений характер, тобто еквівалентність початкової нелінійної системи і її лінійного наближення зберігається лише для обмежених просторових або тимчасових масштабів системи, або для певних процесів, причому якщо система переходить з одного режиму роботи на інший, то варто змінити і її лінеаризовану модель. Застосовуючи лінеаризацію, можна з'ясувати багато якісних і особливо кількісних властивостей нелінійної системи.

Структуризація – це розподіл системи на ієрархічні підсистеми і компоненти та встановлення між ними зв'язків і відносин, що дозволяє здійснювати управління такою системою.

Основні завдання структуризації:

– розподіл системи на блоки, які підлягають управлінню;

– розподіл відповідальності за різні елементи системи, визначення зв'язків між різними структурними елементами;

– визначення витрат – часу, грошей, матеріальних ресурсів;

– створення єдиної бази для планування, складання кошторисів, контролю за витратами тощо;

– встановлення взаємозв'язків між елементами системи;

– визначення комплексу робіт та завдань.

Реструктуризація компанії – це зміна структури компанії (порядку розташування її елементів), а також елементів, що

формують її бізнес, під впливом чинників зовнішнього або внутрішнього середовища. Реструктуризація включає: вдосконалення системи управління, фінансово-економічної політики компанії, її операційної діяльності, системи маркетингу і збуту, управління персоналом.

Реінжиніринг бізнес-процесів (далі – РБП) – фундаментальне переосмислення і радикальне перепроєктування ділових процесів для досягнення різких, стрибкоподібних покращень у вирішальних основних показниках діяльності компанії (фірми, підприємства), таких як вартість, якість, швидкість виконання. На відміну від інших технологій поліпшення бізнесу (виробництва) основна риса РПБ – створення абсолютно нових і більш ефективних процесів (бізнесу, виробництва) без урахування того, що було раніше.

Алгоритмізація – метод опису систем або процесів шляхом створення алгоритмів їхнього функціонування. Алгоритмізація процесів – опис процесів мовою математичних символів для одержання їх алгоритму.

Алгоритм:

– чітко детермінована послідовність дій, що описує процес перетворення об’єкта з початкового стану в кінцевий, записана за допомогою зрозумілих команд;

– це послідовність дій, спрямованих на отримання певного результату за кінцеву кількість кроків.

Алгоритми мають ряд важливих властивостей:

1) скінченність – алгоритм має завжди завершуватися після виконання скінченої кількості кроків. Процедуру, яка має решту характеристик алгоритму, можливо, без скінченності, називають методом обчислень;

2) дискретність – процес, що визначається алгоритмом, можна розчленувати (розділити) на окремі елементарні етапи (кроки), кожен з яких називається кроком алгоритмічного процесу чи алгоритму;

3) визначеність – кожен крок алгоритму має бути точно визначений. Дії, які необхідно здійснити, повинні бути чітко та недвозначно визначені для кожного можливого випадку;

4) вхідні дані – алгоритм має деяку кількість (можливо, нульову) вхідних даних, тобто величин, заданих до початку його роботи або значення яких визначають під час роботи алгоритму;

5) вихідні дані – алгоритм має одне або декілька вихідних даних, тобто величин, що мають досить визначений зв'язок із вхідними даними;

6) ефективність – алгоритм вважають ефективним, якщо всі його оператори досить прості для того, щоб їх можна було точно виконати за скінчений проміжок часу за допомогою олівця та аркуша паперу.

Моделювання – метод дослідження явищ і процесів, що ґрунтується на заміні конкретного об'єкта досліджень (оригіналу) іншим, подібним до нього (моделлю).

Математичне моделювання – моделювання, під час якого модель являє собою систему математичних співвідношень, що описують певні технологічні, економічні чи інші процеси. Найчастіше застосовуються два способи математичного моделювання:

1) аналітичний, що передбачає можливість точного математичного опису чітко детермінованих систем;

2) ймовірнісний, що дозволяє отримати не однозначне рішення, а його імовірнісну характеристику параметрів.

Експеримент – метод дослідження деякого явища в керованих умовах. Відрізняється від спостереження активною взаємодією з об'єктом, що вивчається. Зазвичай експеримент проводиться в рамках наукового дослідження і слугує для перевірки гіпотези, встановлення причинних зв'язків між феноменами.

Ідентифікація – ототожнення, прирівнювання, уподібнення, розпізнавання. Ідентифікація – це процес розпізнавання системою (людиною, дослідником, операційною системою) іншої системи або об'єкта (людини, користувача, предмета, процесу тощо).

Класифікація – система розподілення об'єктів (процесів, явищ) за класами (групами тощо) відповідно до визначених ознак. Інколи вживають термін «категоризація» у значенні «розподілення об'єктів на категорії». Оскільки у результаті класифікації утворюється хоча б один клас (група) принаймні з одним елементом, можна визначати класифікацію як групування.

Тестування застосовується для визначення відповідності предмета випробування заданим специфікаціям. До завдання тестування не входить визначення причин невідповідності заданим вимогам (специфікаціям). Тестування – один із розділів діагностики.

Технологія тестування складається з таких частин:

– зовнішня дія;

– реакція випробовуваного;

– оцінка реакції і висновки.

Верифікація – логіко-методологічна процедура встановлення істинності наукової гіпотези (так само, як і поодинокого, конкретно-наукового твердження) на основі їх відповідності емпіричним даним (пряма або безпосередня верифікація) чи теоретичним положенням, що відповідають емпіричним даним (непряма верифікація).

2.4. Елементи, структура, властивості та класифікація систем

Фундаментальним поняттям системного аналізу є поняття «система». Термін «система» використовується у тих випадках, коли треба охарактеризувати об’єкт, який досліджується чи проєктується як дещо ціле, складне, про який неможливо одразу дістати просте уявлення.

У загальному систему розглядають як множину елементів, що перебувають у певних співвідношеннях і зв’язках один з одним, взаємодіють між собою, утворюють певну цілісність і, як ціле, взаємодіють з навколишнім середовищем. У своїй сукупності ці елементи породжують нові властивості, відсутні у кожного елемента.

До основних рис системи відносять:

– наявність об’єкта, який являє собою множину підоб’єктів (або наявність множини об’єктів, які можуть розглядатися як один складний об’єкт);

– наявність мети системи;

– наявність властивості системи, яка відрізняється від сукупності властивостей частин системи;

– наявність суб'єкта дослідження, який називається спостерігачем;

– наявність завдання, яке визначає відношення спостерігача до об'єкта і є критерієм, за яким здійснюється відбір об'єктів та їх властивостей;

– наявність зв'язку між об'єктом, спостерігачем та завданням, що виражається у наявності певної мови описування.

Система – це множина об'єктів з набором зв'язків між ними і між їх властивостями, які функціонують у часі як єдине ціле, тобто кожний об'єкт працює заради єдиної мети, що стоїть перед системою в цілому.

Підсистемою називають сукупність елементів, які об'єднані єдиним процесом функціонування та у разі взаємодії реалізують певну операцію, що необхідна для досягнення поставленої перед системою в цілому мети.

Елементом системи називають найпростішу неподільну частину системи. Властивості елемента визначаються його взаємодією з іншими елементами системи та впливають на поведінку системи.

Між елементами довільної системи та між різними системами існують *зв'язки,* за допомогою яких вони взаємодіють між собою. Система може мати зовнішні та внутрішні зв'язки. Зв'язки можуть бути також як прямими, так і зворотними.

Системи мають зовсім нові якості, які відсутні у її елементів. Ці якості виникають саме завдяки наявності зв'язків між елементами. За допомогою цих зв'язків здійснюється перенесення властивостей кожного елемента системи до інших елементів.

Зворотні зв'язки є частиною системи причинної залежності та полягають у тому, що результат попередньої дії впливає на такий перебіг процесу в системі: причина підпадає під вплив зворотного впливу наслідку. Якщо зворотний зв'язок підсилює результат впливу наслідку, то його називають позитивним, а якщо послаблює – негативним. Негативні зворотні зв'язки сприяють збереженню стійкості системи. Тільки завдяки наявності зворотних зв'язків у системах можуть відбуватися процеси цілеспрямованої діяльності та регулювання. Зв'язки перетворюють

систему з простого набору компонентів у єдине ціле і разом з компонентами визначають стан та структуру системи.

Важливими для опису систем є поняття структури та ієрархії.

Під *структурою системи* розуміють її стійку впорядкованість та зв'язки між елементами і підсистемами. Структура відбиває найсуттєвіші зв'язки між елементами та підсистемами, які мало змінюються під час змін у системі та забезпечують існування системи і найважливіших її властивостей. Для визначення структури системи потрібно провести її послідовну *декомпозицію,* тобто виділити в ній підсистеми всіх рівнів, доступних аналізу, та їх елементи, які, відповідно до завдань дослідження, не поділяються на складові. Завдяки ієрархічності структура складних систем може бути подана через структуру їх частин – від підсистем до елементів. Структуру системи можна зобразити графічно.

Ієрархія системи – впорядкованість компонентів за ступенем важливості.

Основним фактором, що утворює систему, є її **функція.** Існує кілька поглядів з приводу того, що являє собою функція системи. Так, під функцією системи можна розуміти перетворення її входів (сигнали, інформація, ресурси тощо, які входять у систему) у виходи (сигнали, інформація, ресурси тощо, які виходять з системи, продукуються нею). З іншого погляду, функція системи може полягати у збереженні її існування, підтримці її структури та впорядкованості. Іноді функцію системи ототожнюють із функціонуванням цієї ж системи, визначаючи її як спосіб, засіб або дії для досягнення мети системи.

Системи функціонують у певному зовнішньому середовищі. **Зовнішнє середовище** – це все те, що перебуває зовні системи, включаючи необхідні умови для існування та розвитку системи. Зовнішнє середовище складається з ряду природних, суспільних, інформаційних, економічних, виробничих та інших факторів, що впливають на систему та самі певною мірою перебувають під впливом цієї системи.

Взаємодія між системою та зовнішнім середовищем здійснюється за допомогою входів та виходів. **Вхід системи** – це дія на неї зовнішнього середовища. **Вихід системи** – результат функціонування системи для досягнення певної мети або її реакція на

вплив зовнішнього середовища. Загальна кількість взаємодій системи із зовнішнім середовищем надто велика, тому на практиці обмежуються аналізом найсуттєвіших зв'язків, вибір яких визначається конкретними умовами управління певним об'єктом.

Крім функції, система має ціль. **Ціль системи** – це бажаний стан її виходів. Системи, що мають ціль, називають **цілеспрямованими.**

Стан системи характеризується кількісними та якісними значеннями внутрішніх параметрів (змінних) системи в певний момент. Функціонування системи або зміну станів системи у часі називають **поведінкою** або **рухом.** Отже, поведінка системи – це розгорнута у часі послідовність реакцій системи на внутрішні зміни та зовнішній вплив.

Рівновага – це здатність системи зберігати свій стан якнайдовше (як за відсутності, так і за наявності зовнішніх збурюючих впливів).

Стійкість – це здатність системи повертатися у стан рівноваги після виведення її з цього стану впливом зовнішніх збурень. Стан рівноваги, у який система здатна повертатися, називають **стійким станом рівноваги.**

Важливе значення в системному аналізі має поняття «управління». Управління системою необхідне для забезпечення її цілеспрямованої поведінки під час зміни умов зовнішнього середовища або умов її функціонування. Управління досягається за рахунок відповідної організації системи, під якою розуміють її структуру та спосіб функціонування. Системи з управлінням називають **кібернетичними системами.**

Взаємодія системи із зовнішнім середовищем свідчить, що середовище надає системі ресурси, а одержує від неї та споживає продукти кінцевої діяльності системи (далі – ПКД). ПКД не можуть бути створені в середовищі (принаймні в достатній кількості), оскільки за таких умов немає потреби виділяти систему із середовища. Система необхідна середовищу для задоволення деяких своїх потреб в її кінцевих продуктах. Тому можна зробити висновок, що до створення нових систем спонукає наявність незадоволених потреб або, іншими словами, система створюється для вирішення деякої проблемної ситуації.

Класифікація систем. Класифікація є операцією об'єднання, а саме множина об'єктів, які підлягають класифікації, об'єднується у певні групи за характерними ознаками. Класифікація проводиться тільки тоді, коли є множина об'єктів і необхідно серед них встановити певний порядок, об'єднати їх у певні групи за певними ознаками.

Класифікацію виконують за класифікаційними ознаками. Класифікаційними ознаками є ті ознаки, які, на думку того, хто виконує класифікацію, визначальні для цього класу об'єктів. Поняття «система» охоплює всі об'єкти навколишнього світу. Тому існує велика різноманітність класифікацій систем за різними ознаками. Кожна класифікація виконується спеціалістами, які займаються певним колом проблем, і відображає підхід до проблеми класифікації з точки зору саме цих спеціалістів. Тому єдиної класифікації систем нині немає і навряд чи вона можлива. Певні узагальнення дозволяють вибрати найбільш вживану класифікацію, яку взято за основу в цьому курсі.

Принципи класифікації та класи систем подані у табл. 1.2

Таблиця 1.2

Класи та підкласи систем

№ з/п	Основа класифікації систем	Класи й підкласи систем
1	**2**	**3**
1.	За матеріалом, з якого створені	Матеріальні Ідеальні (абстрактні)
2.	За походженням	Штучні Природні Змішані
3.	За характером зв'язку з навколишнім середовищем	Відкриті Закриті
4.	За принципами поведінки	Матеріальні Гомеостатичні Вирішальні (без передбачення) Здатні передбачувати Рефлексивні

Продовження таблиці 1.2

1	2	3
5.	За складністю	Неживі Статичні структури чи їх основи (кристал) Прості динамічні із заданим законом поведінки (годинник) Кібернетичні системи з циклами керування, що мають зворотний зв'язок (термостат, робот)
		Живі Відкриті системи із самозберігаючою структурою Живі організми з низькою здатністю сприймати інформацію (рослини) Живі організми з більш розвиненою системою сприйняття інформації (тварини) Живі організми із самосвідомістю (людина) Соціальні системи (етнос, нація) Трансцендентні системи чи системи, що пребувають поза нашою свідомістю
6.	За ступенем організованості	Добре організовані Погано організовані Самоорганізуючі Саморегулівні Самонавчаючі Самонастроювані Самовідновлювальні Самовідтворювані
7.	За ступенем ресурсної забезпеченості	Малі Великі Прості Складні Звичайні Енергокритичні
8.	За характером цілей	Призначені для певної цілі Здатні обирати ціль і до неї прагнути
9.	За описом змінних	Якісний опис Кількісний опис Змішаний опис
10.	За способом керування	Керування зовні Самокерування З комбінованим керуванням
11.	За типом операторів системи	Чорний ящик Непараметризований клас Параметризований клас Білий ящик

До самоорганізуючих систем належать системи, що мають механізми регулювання. Найбільш простими з них є системи саморегулювання. Вони у своєму складі мають дві частини: частину, яка підлягає керуванню, – S, та керуючий пристрій чи регулятор – R. Зображають їх, як показано на рис. 1.5.

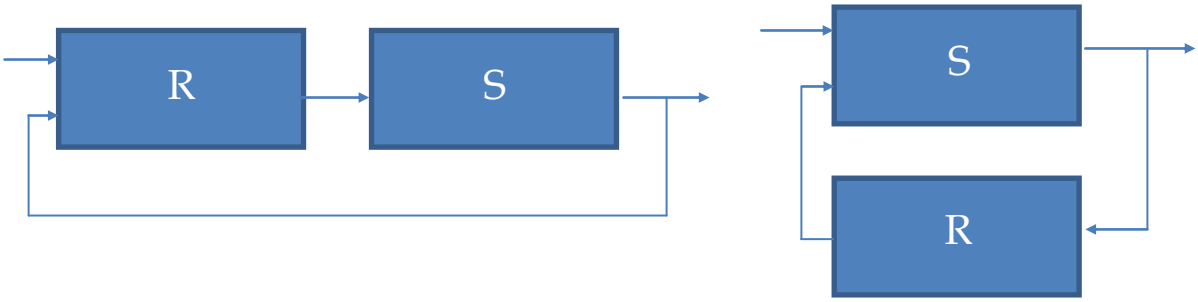


Рис. 1.5. Системи із зворотним зв'язком

У таких системах важливим є наявність зворотного зв'язку. Під останнім розуміють подання на вхід системи сигналу про вихідну величину (пропорційного, залежного від швидкості зміни чи інтегрального значення вихідної величини). Працює така система так. Вихідна величина системи чи інформація про її стан подається на регулятор. Регулятор, залежно від вимог, які стоять перед системою, та від інформації, яка надійшла через зворотний зв'язок, регулює її роботу.

Самонавчаючі системи – це системи, що мають здатність засвоювати й запам'ятовувати минулий досвід і змінювати свою поведінку відповідно до набутих знань.

Класифікація за описом змінних. Ця класифікація здійснюється залежно від ступеня складності системи, наших знань про неї і від того, як описуємо функціонування системи. Змінними системи є вхідні величини *X(i),* вихідні *Y(i)* та внутрішні параметри системи *z(t).* Відповідно до типу цих величин розрізняють системи з кількісним описом змінних, системи з якісним описом та системи зі змішаним (частково з кількісним, а частково з якісним) описом.

Системи, функціонування яких повністю відоме і можна кількісно встановити значення змінних, відносять до систем з кількісним описом змінних. На другому рівні класифікації такі системи можуть бути класифіковані за характером змінних як дискретні, чи неперервні, за кількістю змінних: одно- чи багатовимірні та за іншими параметрами.

Системи з якісним описом – це системи, в яких вхідні й вихідні величини описуються якісно. На другому рівні вони можуть бути класифіковані як системи зі змістовним, формалізованим та змішаним описами. Системи зі змістовним описом мають тільки загальний словесний опис змінних, виконаний мовою, близькою до розмовної. Системи з формалізованим описом – це вже системи, в яких більш докладно вивчені змінні, які підлягають певним формальним правилам і можуть бути описані деякою формальною мовою.

Системи зі змішаним описом змінних – це системи, в яких деякі змінні приймають числові значення, які можна виміряти й виразити числом, а інші описані тільки якісно.

Властивості систем. Поняття «система» найкраще конкретизується у процесі розгляду її властивостей. Чотири основні властивості системи:

– система є сукупністю елементів. Під час певних умов елементи можуть розглядатись як системи;

– між елементами існують суттєві зв'язки чи властивості, які за силою зв'язку перевищують зв'язки між елементами системи та елементами, які не входять у систему. Під суттєвими зв'язками розуміють такі, які закономірно з необхідністю визначають інтегровані властивості системи. Ці суттєві зв'язки визначають систему, відділяючи її від простої сукупності (конгломерату) і виділяють її з навколишнього середовища у вигляді об'єкта;

– системі властива певна організація, що виявляється у зменшенні ентропії системи порівняно з ентропією сукупності елементів, які складають систему. Організація системи приводить до зменшення безладу, зменшення кількості можливих станів системи;

– існування інтеграційних властивостей, тобто властивостей, які властиві системі в цілому і не властиві жодному елементу системи. Тобто властивості системи не зводяться тільки до властивостей її елементів.

Ознаками системи є:

– цілісність;

– якісна визначеність;

– відмежованість відносно середовища;

– гетерогенність і структурованість;

– взаємодія частин системи між собою;

– взаємодія і зв'язок з навколишнім середовищем;

– наявність інтегральних характеристик;

– емерджентність;

– наявність цілей та їх сукупності, цілеспрямованість.

Цілісність означає, що система – це об'єднання частин, яке відносно навколишнього оточення виступає як єдине ціле.

Під якісною визначеністю розуміють, що система – це така сукупність елементів, яка має свої якісні ознаки, характерні тільки для цієї системи і відсутні в інших системах. Ці ознаки проявляються тільки у цій системі. Вони визначають відношення до інших систем. Тобто система є цілісністю, якщо має властиві тільки їй якісні ознаки, за якими вона відрізняється від інших систем. Якість здебільшого може бути виражена кількісними величинами.

Відмежованість системи від середовища означає, що будь-яка система має свої межі, які відділяють систему від навколишнього середовища. Вони визначають, що входить у систему і що в неї не входить, є зовнішнім відносно системи. Переважна більшість систем має чіткі межі, проте вони не завжди можуть визначатися однозначно. Деколи визначити межі надто складно.

Гетерогенність системи і структурованість. Під гетерогенністю розуміють неоднорідність, те, що система складається з різних частин. У визначенні системи вказано, що система – це сукупність елементів. Але система – це не проста сукупність. Структурованість означає, що система є певним чином організованою сукупністю, має певну структуру. Будь-яка система також має певну структуру. Ця структура забезпечує об'єднання елементів системи таким чином, щоб це об'єднання мало свою якісну визначеність, цілісність. Гетерогенність – поняття більш вузьке, воно просто означає неоднорідність складу, наявність складових частин.

Взаємодія частин системи між собою означає, що в системі частини взаємодіють між собою і тільки у цій взаємодії вони утворюють певну систему. З наведеного прикладу зрозуміло, що

тролейбус буде системою тоді, коли складові його частини: двигун, ходова частина, електрообладнання – певним чином взаємодіють між собою, причому ця взаємодія однозначна, визначена, обумовлена в найдрібніших деталях.

Взаємодія з навколишнім середовищем означає, що система як ціле взаємодіє з іншими системами. Це зумовлене цілісністю системи, її якісною визначеністю. Під час взаємодії з навколишнім середовищем виявляються властивості системи. За характером взаємодії розрізняють відкриті й закриті системи. Відкрита система – це система, яка активно обмінюється з навколишнім середовищем речовиною, енергією та інформацією. У закритій системі такий обмін обмежений або відсутній. Поняття відкритості системи має велике значення під час вивчення питання розвитку й життєвого циклу системи. Для закритих систем характерні процеси старіння.

Під поняттям інтегрованості розуміють, що в системі властивості окремих елементів об'єднуються і виступають разом у новій якості.

Емерджентність – це поява нових якостей, не властивих елементам, що складають систему. Кожна система є сукупністю певних частин, певних елементів. Особливістю системи є те, що в результаті об'єднання декількох елементів і утворення системи з'являються нові властивості, яких не має жоден елемент до створення системи. Ця властивість системи і називається емерджентністю. Емерджентність визначає, що властивості системи не зводяться до властивостей елементів, з яких вона складається. Емерджентність системи може характеризувати ступінь організованості системи. Чим більше характеристики системи відрізняються від характеристик елементів, з яких вона утворена, тим більш організованою є система.

Наявність цілей, цілеспрямованість – одна з основних ознак системи. Кожна система має певну ціль існування або створена для певної цілі. Система може мати одну ціль або сукупність цілей. У випадку сукупності цілей всі вони утворюють певну ієрархію, в якій є головні, першочергові цілі й другорядні, підпорядковані головним. Визначення цілей системи є завданням, що має свої особливості. Цілі по-різному визначаються для систем штучного походження й природних систем.

Штучні системи, системи, створені людиною, завжди мають суб'єктивні цілі, тобто цілі, поставлені суб'єктом, який створив систему. Визначити цілі таких систем, як правило, неважко. Наприклад, телевізор створений для прийому та відображення інформації. Автомобіль створений для перевезення людей і вантажів.

Типологія структур. Структура – це все те, що вносить порядок до безлічі об'єктів, тобто сукупність зв'язків і відносин між частинами цілого, необхідні для досягнення мети. Поняття структури – одне з найбільш важливих понять як в абстрактному розумінні, так і під час його конкретизації.

Структури систем бувають різного типу, різної типології (або ж просторової структури):

1. Лінійна структура.



Рис. 1.6. Структура лінійного типу

Прикладом лінійної структури є структура станцій метро на одній незакільцьованій лінії.

2. Ієрархічна структура – систему визначають як ієрархічну цілісність.

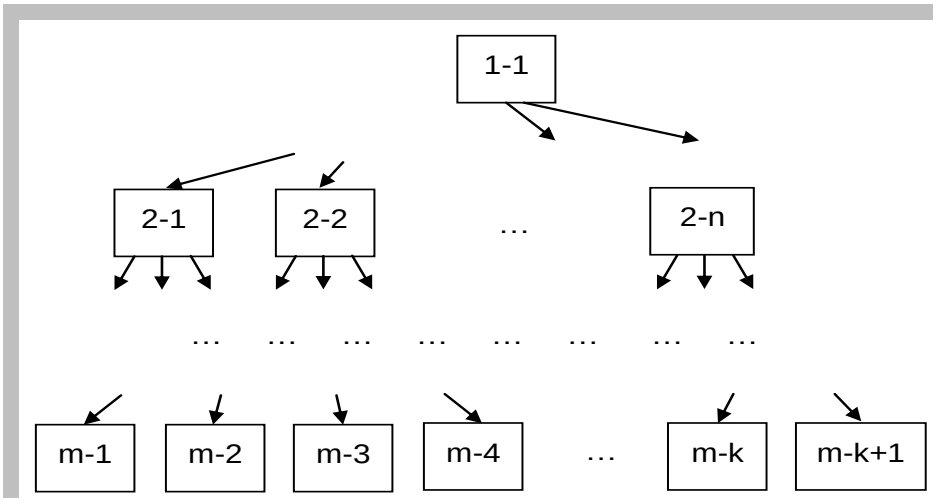


Рис. 1.7. Структура ієрархічного (деревоподібного) типу

Прикладом ієрархічної структури є структура управління ЗВО: ректор – проректори – декани – завідувачі кафедр і начальники підрозділів – викладачі кафедр і співробітники інших підрозділів.

3. Структура мережевого типу.

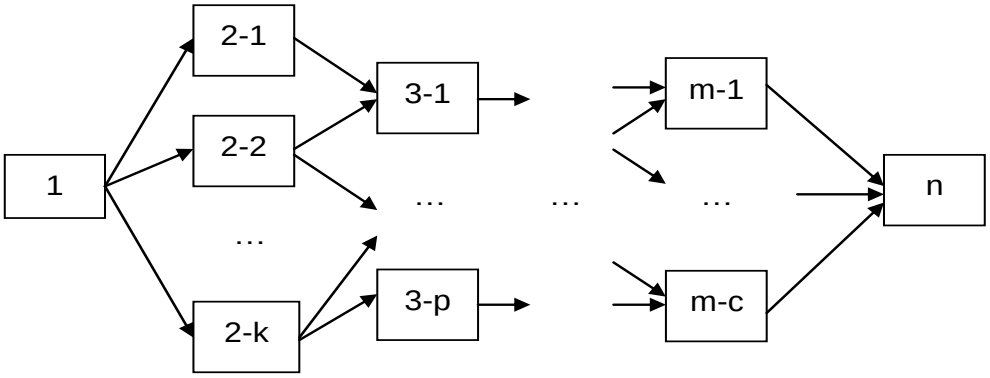


Рис. 1.8. Структура мережевого типу

Приклад мережевої структури – структура організації будівельно-монтажних робіт під час спорудження будинку: деякі роботи, наприклад монтаж стін, впорядкування території та ін., можна виконувати паралельно.

4. Структура матричного типу.

Прикладом матричної структури є структура працівників відділу НДІ, що виконують роботи з однієї теми.

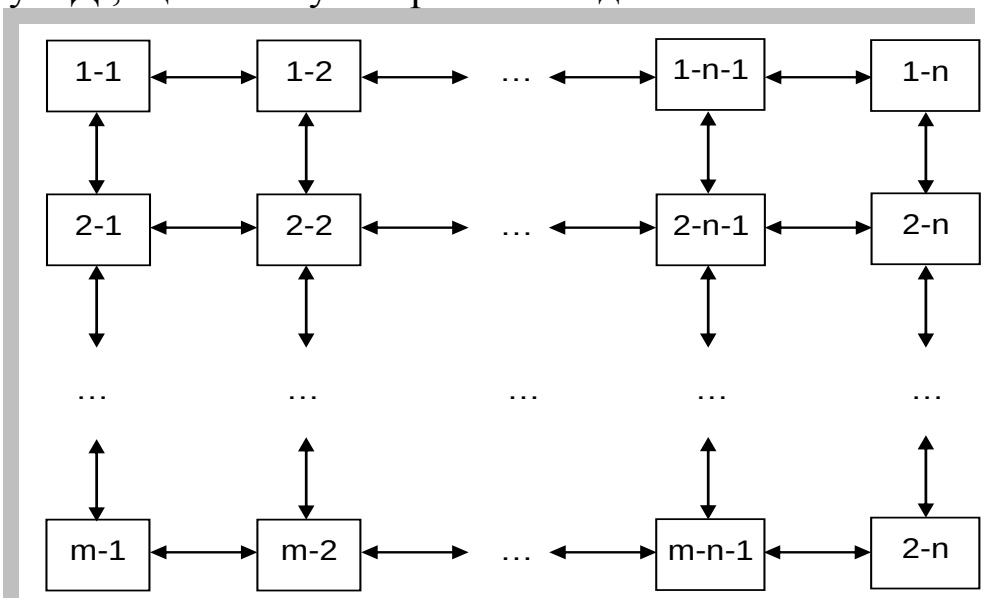


Рис. 1.9. Структура матричного типу

Крім зазначених основних типів структур, використовуються й інші, що утворюються за допомогою їх коректних комбінацій – з'єднань і укладань.

«Вкладання один в одного» площинних матричних структур може привести до складнішої структури – структури просторової

матричної (наприклад, речовини кристалічної структури типу, зображеного на рис. 1.8). Структура сплаву і навколишнього середовища (макроструктура) можуть визначати властивості й структуру сплаву (мікроструктуру).

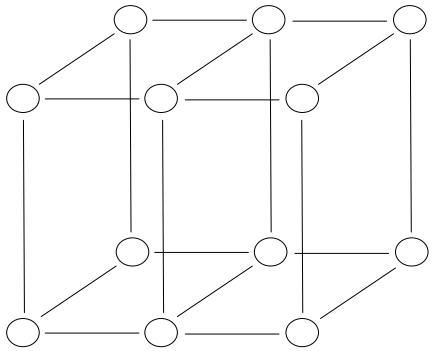


Рис. 1.10. Структура кристалічного (просторово-матричного) типу

Такого виду структури часто використовуються в системах з тісно пов'язаними і рівноправними («по вертикалі» і «по горизонталі») структурними зв'язками. Зокрема, таку структуру можуть мати системи відкритого акціонерного типу, корпорації на ринку з дистриб'ютерною мережею та інші.

З однакових елементів можна отримувати структури різного типу.

Структура є сполучною, якщо можливий обмін ресурсами між будь-якими двома підсистемами системи.

Опис (специфікація) системи – це опис усіх її елементів (підсистем), їх взаємозв'язків, мети, функцій за деякими ресурсами, тобто всіх допустимих станів.

Морфологічний опис системи – опис будови або структури системи: опис сукупності *A* елементів цієї системи і необхідного для досягнення мети набору відносин *R* між ними. Морфологічний опис задається кортежем S = {A, B, R, V, Q}, де A – множина елементів і їх властивостей; B – множина відносин з навколишнім середовищем; R – множина зв'язків в *A*; V – структура системи, тип цієї структури; Q – опис, подання системи певною мовою.

З морфологічного опису системи отримують функціональний опис системи (тобто опис законів функціонування, еволюції системи), а з неї – інформаційний опис системи (опис інформаційних зв'язків як системи з навколишнім середовищем, так і підсистем

системи) або ж так звану інформаційну систему, а також інформаційно-логічний опис системи.

Морфологічний опис системи залежить від зв'язків, що враховуються, їх глибини (зв'язки між основними підсистемами, другорядними підсистемами, елементами), структури (лінійна, ієрархічна, мережева, матрична, змішана), типу (прямий зв'язок, зворотний зв'язок), характеру (позитивна, негативна).

Інформаційний опис системи часто дозволяє отримувати додаткову інформацію про систему, витягувати нові знання про систему, вирішувати інформаційно-логічні завдання, досліджувати моделі систем.

Дві системи називаємо еквівалентними, якщо вони мають однакові цілі, складові елементи, структуру. Між такими системами можна встановити зв'язок (зв'язки) деяким конструктивним чином.

2.5. Управління в системі та управління системою

Управління в системі – внутрішня функція системи, здійснювана в системі незалежно від того, яким чином, якими елементами системи вона повинна виконуватися.

Управління системою – виконання зовнішніх функцій управління, що забезпечують необхідні умови функціонування системи.

Важливий аспект управління системою безпосередньо виражається через принцип Ешбі (принцип необхідної різноманітності): керуюча система повинна мати вищий рівень організації (більша різноманітність, більший вибір), ніж керована система, тобто різноманіття може бути кероване (зруйноване) лише різноманіттям.

Функції і завдання управління системою:

1) організація системи – повне, якісне виділення підсистем, опис їх взаємодій і структури системи (як лінійної, так і ієрархічної, мережевої або матричної);

2) прогнозування поведінки системи, тобто дослідження майбутнього системи;

3) планування (координація в часі, просторі, за інформацією) ресурсів і елементів, підсистем і структури системи, необхідних (або достатніх у разі оптимального планування) для досягнення мети системи;

4) облік і контроль ресурсів, що приводять до певних бажаних станів системи;

5) регулювання – адаптація і пристосування системи до змін зовнішнього середовища;

6) реалізація певних спланованих станів, рішень.

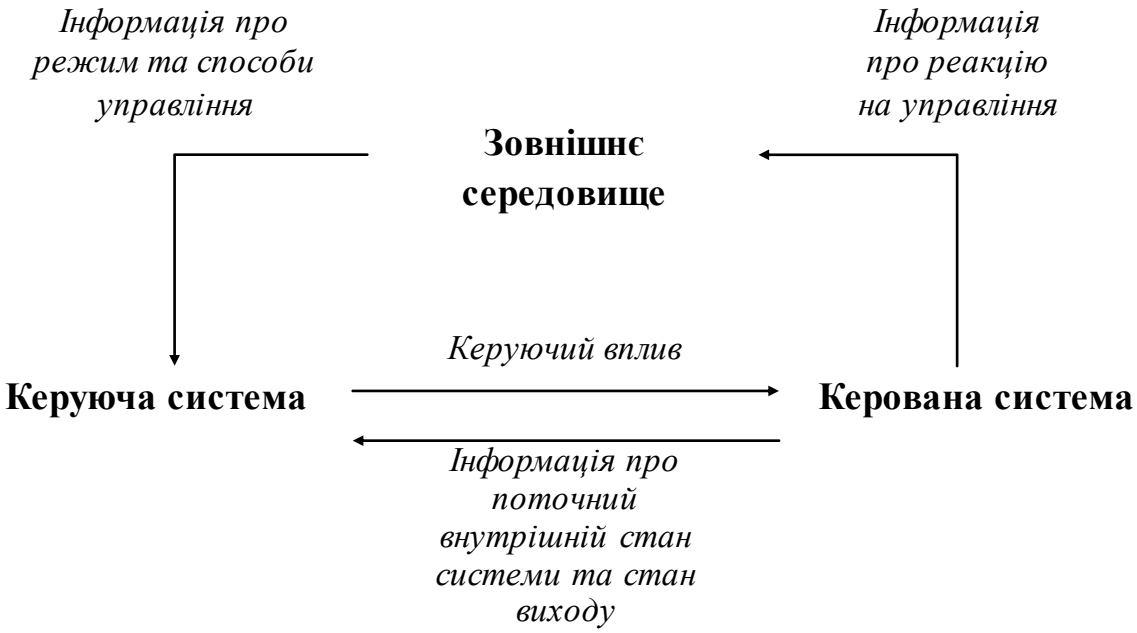


Рис. 1.11. Загальна схема управління системою

Функції і завдання управління системою взаємозв'язані, а також взаємозалежні. Не можна здійснювати повне планування в економічній системі без прогнозування, обліку і контролю ресурсів, без аналізу попиту і пропозиції – основних регуляторів ринку. Економіка будь-якої держави – завжди керована система, хоча підсистеми управління можуть бути організовані по-різному, мати різні елементи, цілі, структуру, відносини.

Виявлення керуючих параметрів і їх використання для управління системою може зменшити складність системи. У свою чергу, зменшення складності системи може зробити систему повністю керованою.

Чим різноманітніші вхідні сигнали (параметри) системи, кількість різних станів системи, тим різноманітніші зазвичай вихідні сигнали, складніша система, актуальніша проблема пошуку інваріантів управління.

Запитання до параграфа 2

1. Що таке системний аналіз?
2. Які питання вивчає системний аналіз?
3. Які цілі системного аналізу?
4. Які основні принципи системного підходу?
5. Які основні етапи системного аналізу?
6. З яких етапів складається системне дослідження?
7. З яких етапів складається системний аналіз?
8. Які три концепції покладено в основу методології системного аналізу?
9. Що таке ціль?
10. З чим пов'язана складність формулювання цілі?
11. Які ситуації можуть завадити досягненню цілі?
12. Яка сутність методу побудови дерева цілей?
13. Які є основні етапи проведення системних досліджень?
14. Які існують методи системного аналізу?
15. Які основні завдання вирішують за допомогою аналізу та синтезу систем?
16. Яка сутність функціонального аналізу?
17. Яка сутність інформаційного аналізу?
18. Що таке алгоритмізація?
19. Яка сутність математичного моделювання?
20. Які основні риси має система?
21. Що таке система?
22. Що таке структура системи?
23. Що таке ієрархія системи?
24. Що таке функції і мета системи?
25. Які є класи систем?
26. Що таке системи із зворотним зв'язком?
27. Які властивості систем?
28. Що є ознаками системи?
29. Що таке емерджетність?
30. Яка типологія структур систем?
31. Що таке морфологічний опис системи?
32. Які функції і завдання управління системою?

§ 3. Теоретичні основи інформаційно-аналітичної діяльності як складової бізнес-аналітики

3.1. Основні поняття

У загальному **бізнес-аналітика і моделювання** – це перетворення або представлення певної інформації. Тому інформаційно-аналітична діяльність виступає їх певною основою, базисом, початковим етапом. Це зумовлює потребу у більш детальному розгляді понять «інформація» та «інформаційно-аналітична діяльність».

Інформація – абстрактне поняття, що має різні значення залежно від контексту. На інтуїтивному рівні інформація означає зміст того, про що отримувач довідався. Має декілька значень:

– роз'яснення;

– виклад фактів, подій;

– тлумачення.

Погляд на інформацію з точки зору її споживачів окреслює так це поняття: **інформація** – це нові відомості, які прийняті, зрозумілі й оцінені її користувачем як корисні, **інформація** – це нові знання, які отримує споживач (суб'єкт) у результаті сприйняття і переробки певних відомостей.

Складність однозначного визначення поняття «інформація» пов'язана з тим, що цей термін використовується у багатьох сферах людської діяльності, тобто є багатогранним. Залежно від галузі використання термін «інформація» одержав велику кількість різних визначень, зокрема:

– відомості або повідомлення про щось (побутове);

– роз'яснення, виклад;

– оригінальність, новизна;

– результат інтелектуальної (аналітико-синтетичної чи евристичної) діяльності певної людини щодо подання відомостей, повідомлень, сигналів, кодів, образів;

– документовані або публічно оголошені відомості про події та явища, що відбуваються у суспільстві, державі та навколишньому природному середовищі (Закон України «Про інформацію»)[12];

[12] URL : https://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=2657-12

– відображення заміни дискретного значення енергетичного процесу в контурі керування пристрою керування іншим дискетним значенням, зокрема таким:

– розпізнаний кібернетичною системою сигнал або комплекс сигналів (образ), який зменшує кількість варіантів вибору нею чергової дії команди;

– комунікація та зв'язок, у процесі якого усувається невизначеність (інформаційна ентропія);

– позначення змісту, отриманого із зовнішнього світу в процесі нашого пристосування до нього і пристосування до нього наших почуттів;

– ймовірність вибору;

– міра неоднорідності розподілу матерії та енергії у просторі та у часі, міра змін, якими супроводжуються всі процеси, що відбуваються у світі;

– заперечення ентропії, міра хаосу в системі;

– передача різноманітності;

– міра складності структур;

– відображена різноманітність;

– результат інтелектуальної (аналітико-синтетичної чи евристичної) діяльності певної людини щодо подання відомостей, повідомлень, сигналів, кодів, образів тощо;

– універсальна субстанція, що пронизує усі сфери людської діяльності, слугує провідником знань та думок, інструментом спілкування, взаєморозуміння та співробітництва, утвердження стереотипів мислення та поведінки.

Сьогодні **інформація стає елементом ринкового механізму.** Крім того, вона є економічним ресурсом, фактором конкурентної боротьби, суспільним благом і **товаром.**

Як предмет і результат управлінської діяльності, інформація характеризується такими **властивостями:**

– виступаючи на певному етапі управління як предмет праці, в подальшому сама стає продуктом праці;

– має здатність тривалого і багаторазового застосування і не втрачає своїх властивостей;

– має властивість накопичення;

– здатна старіти і втрачати свої властивості з появою нових відомостей і знань;

– корисність та цінність отриманої інформації залежить від уміння користувача скористатися нею, його кваліфікації;

– підлягає тиражуванню й одночасному використанню в багатьох місцях незалежно від мети використання.

Споживча вартість інформації виявляється в її здатності задовольнити людську потребу в ній. Така потреба виникає у разі, коли людина відчуває недостатність у певних благах, бажає їх отримати, володіти ними, з метою переборення цього відчуття звертається за допомогою до інформації.

Особливості інформації як товару:

– володіння інформацією не є гарантією абсолютного права на її використання, тобто її можна продати і залишитись її власником;

– продаж інформації може здійснюватися значною кількістю покупців, і жоден з них не стає її одноосібним власником;

– виникає у результаті не тільки і не просто розумової праці, а її творчої частини;

– цінність інформації не має прямого зв'язку з вартістю її створення, зберігання та обробки, а залежить від потреби в ній, її актуальності, достовірності та інших якісних характеристик, очікуваних споживачем.

Інформаційний простір (національний) – інформаційне середовище, в якому здійснюються інформаційні процеси та інформаційні відносини щодо створення, збирання, відображення, реєстрації, накопичення, збереження, захисту і поширення інформації, інформаційних продуктів та інформаційних ресурсів, на яке розповсюджується юрисдикція держави.

Інформаційний продукт (продукція) – створена виробником сукупність документованої інформації, відомостей, даних і знань, яка призначена для забезпечення інформаційних потреб користувача.

Інформаційні ресурси – це увесь обсяг знань, відчужених від їх творців, зафіксованих на матеріальних носіях і призначених для суспільного використання.

Інформаційна послуга – це отримання і надання в розпорядження користувача інформаційних продуктів.

Ринок інформаційних послуг – сукупність економічних, правових і організаційних відносин з продажу і купівлі інформаційних продуктів та послуг **(ІПП),** які складаються між їхніми постачальниками і споживачами.

Інформаційна безпека – складова національної безпеки, процес управління загрозами та небезпеками державними і недержавними інституціями, окремими громадянами, за якими забезпечується інформаційний суверенітет України.

Інформаційні відносини – відносини, які виникають у всіх сферах життя і діяльності людини, суспільства і держави у разі одержання, використання, поширення та зберігання інформації.

Інформаційна система – організаційно впорядкована сукупність інформаційних ресурсів та інформаційних технологій і засобів забезпечення інформаційних процесів.

Інформаційні технології – цілеспрямована організована сукупність інформаційних процесів з використанням засобів електронної техніки, що забезпечують високу швидкість обробки даних, швидкий пошук інформації, розосередження даних, доступ до джерел інформації незалежно від місця їхнього розташування.

Інформаційне суспільство – суспільство, в якому більшість робітників займаються створенням, збиранням, відображенням, реєстрацією, накопиченням, збереженням і поширенням інформації, особливо її вищої форми – знань.

Інформаційна війна – процес боротьби між суб'єктами із застосуванням інформаційної зброї.

3.2. Параметри інформації

Параметри інформації – характеристики, за допомогою яких оцінюються інформаційні ресурси.

Добираючи необхідну інформацію, обов'язково потрібно враховувати такі **характеристики (параметри)** інформації:

1. Якісні показники:

– *актуальність* (інтегрований показник, який визначає своєчасність і цінність інформації у сукупності);

– *адекватність* (однозначна відповідність інформації відображеному об'єктові);

– *цілковита певність* (відсутність в інформації прихованих помилок);

– *об'єктивність* (очищення інформації від неминучих перекручень у процесі її передачі, а також від наслідків її суб'єктивного розуміння);

– *однозначність* (відсутність інших тлумачень наявної інформації);

– *новизна* (уперше отримана інформація);

– *корисність* (відповідність інформації певній меті);

– *доступність* (можливість отримання інформації);

– *своєчасність* (отримання інформації у межах того часу, коли вона придатна для прийняття рішення);

– *релевантність* (відповідність між змістом інформації і запитом користувача);

– *точність* (відношення релевантної інформації до загальної сукупності релевантної і нерелевантної інформації);

– *відповідність змісту* інформації потребам користувача;

– *ергономічність* (зручність форми та обсягу інформації);

– *привабливість* інформації;

– *живучість* (здатність інформації зберігати свої характеристики у часі);

– *захищеність* (неможливість несанкціонованого використання (або зміни) інформації);

– *можливість перевірки* (досяжність джерела для підтвердження істинності інформації).

2. Кількісні показники:

– *повнота* (співвідношення між наявною інформацією та інформацією, яка досяжна);

– *достатність* (можливість досягти мети у разі наявної інформації);

– *обсяг* (загальна кількість інформації).

3. Ціннісні показники:

– *цінність* (значимість, важливість, потрібність інформації для прийняття рішення);

– *вартість* (ціна, уречевлена у продукті суспільна праця).

Основним результатом інформаційно-аналітичної діяльності найчастіше є аналітичний документ у формі *доповіді, огляду, довідки, прогнозу*, де подаються **висновки та рекомендації.** Вони абсолютно необхідні для того, щоб робота мала завершений вигляд і могла бути максимально корисною. Висновки є наслідком досягнення поставленої цілі.

У багатьох документах читаються і запам'ятовуються лише *висновки,* тобто результати аналізу. У висновках мають викладатись коротко і чітко найважливіші моменти, але водночас стисліть не повинна породжувати неправильне сприйняття. Написання висновків потребує **найвищої майстерності аналітика.**

3.3. Методи збору аналітичної інформації

Аналітична інформація поділяється на зовнішню і внутрішню. Зовнішня інформація – це інформація, що надходить до аналітика із зовнішнього середовища. Вона поділяється на:

– *ринкову інформацію;*

– *інформацію про конкурентів;*

– *макроекономічну і геополітичну інформацію;*

– *інформацію про постачальників;*

– *зовнішню фінансову інформацію;*

– *інформацію про регулювання й оподаткування тощо.*

Внутрішня інформація – це інформація про сам об'єкт, процес чи явище, що досліджується.

До **джерел** аналітичної інформації належать:

– урядові заклади;

– бібліотеки;

– торгові асоціації;

– компанії, що займаються приватними дослідженнями й інформацією;

– газети і журнали;

– служби бізнес-інформації;

– бази даних у режимі он-лайн.

Організації використовують аналітичну інформацію у **чотирьох основних цілях:**

1) для створення можливості зростання прибутку і розширення ринків;

2) для зниження ризику і зменшення невизначеності;

3) для одержання влади і засобів впливу на інших;

4) для контролю й оцінки продуктивності й ефективності своєї фірми.

Активний збір інформації найчастіше буває моментальним і фокусується на конкретній події; він не часто буває постійним і усвідомлюється саме як процес збору інформації.

Регулярний збір інформації виникає як складова бізнесу, але більш цілеспрямовану роботу можна включити до складу заходів щодо щорічного планування і складання бюджету, з координації стратегічних рішень з прогнозами зовнішніх і внутрішніх умов, з розробки нових товарів, послуг, ринків або раціоналізації забезпечення.

Пасивний збір інформації виникає в більшості випадків у процесі різних чи взаємодій операцій у бізнесі, а також під час неофіційного обміну. Збір інформації як процес складається з двох частин: безупинний збір загальних ділових зведень від неформального спілкування з колегами й іншими контактами як частина щоденної діяльності, а також постійне підживлення інформацією з більш офіційних мереж, звідки надходять важливі новини.

Під час проведення аналітичних досліджень, крім аналізу статистичних показників, найчастіше застосовуються такі **методи збору інформації:**

– анкетне опитування;

– інтерв'ю;

– спостереження;

– аналіз документів;

– соціометричне опитування;

– тестування;

– експеримент тощо.

3.4. Базові моделі, що застосовуються під час аналізу інформаційних явищ

Модель – це сукупність логічних, математичних чи інших об'єктів, зв'язків і співвідношень, що відображають з необхідною або максимально досяжною мірою подібності деякий фрагмент реальності, що підлягає вивченню, а також опис усіх властивостей об'єкта, що моделюється.

Інформаційна модель – сукупність інформації, що характеризує істотні властивості і стани об'єкта, процесу, явища, а також взаємозв'язок із зовнішнім середовищем.

Інформаційні моделі називають також уявними (тому що вони не мають матеріального втілення), іноді – логічними. Усі знання людства про реальний світ – це безліч інформаційних моделей.

Образні моделі. Образні моделі (малюнки, фотографії та ін.) являють собою зорові образи об'єктів, зафіксовані на будь-якому носії інформації (папері, фото- і кіноплівці та ін). Образні інформаційні моделі широко використовуються в освіті, де потрібна класифікація об'єктів за їх зовнішніми ознаками (навчальні плакати з ботаніки, біології та фізики).

Знакові інформаційні моделі. Знакові інформаційні моделі будуються з використанням різних мов (знакових систем). Знакова інформаційна модель може бути представлена у формі тексту (наприклад, програми на мові програмування) або формули (наприклад, другого закону Ньютона).

Табличні моделі. Широко поширені інформаційні моделі у формі таблиць. У таблиці хімічних елементів Д. І. Менделєєва хімічні елементи розташовуються в клітинках таблиці за зростанням атомних ваг, а в стовпцях – за кількістю валентних електронів.

Інформаційні моделі діляться на **описові та формальні.**

Описові інформаційні моделі – це моделі, створені природною мовою (тобто будь-якою мовою спілкування між людьми: англійською, українською, китайською, мальтійською тощо) в усній або письмовій формі.

Формальні інформаційні моделі – це моделі, створені формальною мовою (тобто науковою, професійною або спеціалізованою, наприклад мовою програмування). Приклади формальних моделей: усі види формул, таблиці, графи, карти, схеми тощо.

Ідеалізація об'єкта. Ідеалізація об'єкта є неодмінним етапом створення **інформаційної моделі.** Сутність ідеалізації полягає у визначенні, які саме риси та властивості об'єкта є суттєвими для розв'язання поставленої задачі, впливають на результат рішення і мають бути відтвореними в моделі, а які риси та властивості є несуттєвими і під час побудови моделі можуть не враховуватися.

3.5. Інформаційно-аналітична діяльність

Інформаційна аналітика виконує насамперед завдання якісно-змістовного перетворення первинної інформації, функціонально перетинаючись щодо цього з *науковою* та *управлінською* діяльністю. Характер функціональної взаємодії у системі «аналітика-наука-керування» визначається специфікою аналітики.

Аналітик, спираючись на *інформаційні моделі*, виявляє в них об'єктивні закономірності і тенденції, визначає рушійні механізми, причиново-наслідкові зв'язки. У цьому змісті аналітик створює нове знання про той фрагмент реальності, що перебуває в полі його професійного інтересу, виступаючи дослідником своєї предметної сфери.

Інформаційно-аналітична діяльність (далі – ІАД) включає в себе підбір і систематизацію фактів щодо певного питання, їх оцінку, відбір, тлумачення, чітке й продумане викладення в усній або письмовій формі.

Основна мета **інформаційно-аналітичної** діяльності полягає в отриманні максимальної користі від інформації, яка є в розпорядженні, для того щоб правильно зрозуміти й оцінити ситуацію, бачити її у перспективі, а в кінцевому підсумку – успішно діяти.

Цілі ІАД можна розподілити на: **стратегічні і тактичні.**

Стратегічні – на основі концепцій, методів, засобів зібрати, накопичити, обробити, проаналізувати такий обсяг необхідної і достатньої інформації, щоб надати споживачеві (замовнику) якомога більш якісний і повний інформаційно-аналітичний продукт.

Тактичні – залежно від завдання, яке ставить перед аналітиком споживач (замовник), визначаються конкретно в кожному випадку окремо. Це можуть бути, наприклад, питання збільшення

результативності та прибутковості певної ділянки бізнесової діяльності, зокрема пошук партнерів, ринків збуту, джерел фінансування, ступінь ризиків тощо.

Об'єктами інформаційно-аналітичної діяльності є:

– інформаційно-аналітичні продукти;

– бази даних;

– апаратні засоби;

– програмні засоби;

– інформаційні послуги;

– інформаційно-пошукові системні технології;

– імена, адреси в Internet тощо.

Суб'єктами інформаційно-аналітичної діяльності є:

– особистість;

– держава;

– виробники первинної інформації;

– виробники вторинної;

– компанії та підприємства оптової і роздрібної торгівлі інформаційними продуктами та послугами, серед них різноманітні служби пошуку інформації, бібліотеки, інформаційні центри, інститути;

– виробники (постачальники) технічних засобів обробки інформації;

– виробники (постачальники) програмних засобів обробки інформації;

– інформаційні посередники (брокерські фірми);

– інтернет-провайдери;

– виробники (розробники) технологій обробки інформації, засобів комунікацій тощо.

Інформаційно-аналітична робота – це процес, у результаті якого первинна інформація (сирі факти) перетворюються у вторинну, нову аналітичну інформацію, довершену продукцію, передбачену для передачі замовнику.

Будь-яка кількість фактів, оброблених найкращим чином, занесених у досьє, – не має жодної цінності, доки не розкриємо їх сенс, не зіставимо між собою, не покажемо перспективи розвитку ситуації, явища, процесу і не передамо у тому вигляді, в якому їх значення буде абсолютно ясним і зрозумілим споживачу.

Нараховується декілька сотень видів інформаційно-аналітичної діяльності залежно від сфери знань і практики застосування. З цієї кількості можна виділити найпоширеніші, найпопулярніші серед аналітиків види аналізів:

– *проблемний* – здійснення проблемного структурування, що вимагає виділення комплексу проблем ситуації, що склалася, їх типології, характеристик, наслідків, шляхів розв'язання;

– *системний* – визначення характеристик, структури ситуації, її функцій, взаємодії з навколишнім і внутрішнім середовищем;

– *причиново-наслідковий* – виявлення причин, які привели до виникнення цієї ситуації, наслідків її подальшого розвитку;

– *праксеологічний* – діагностика змісту діяльності в ситуації, що склалася, її моделювання й оптимізація;

– *аксіологічний* – побудова системи оцінок явищ, діяльності, процесів, ситуацій з позицій певної ціннісної системи;

– *ситуаційний* – моделювання ситуації, її складових, умов, наслідків, діючих осіб;

– *прогностичний* – підготовка прогнозів стосовно ймовірного, потенційного і бажаного майбутнього;

– *рекомендувальний* – вироблення рекомендацій відносно поведінки діючих осіб у відповідній ситуації;

– *програмно-цільовий* – розробка програм діяльності в ситуації, що виникла.

Схематично універсальна технологія аналітичної діяльності має такий вигляд:

1. Визначення об'єкта, предмета чи проблеми аналізу.
2. Побудова ідеальної моделі об'єкта чи предмета.
3. Побудова гіпотез.
4. Вибір типу (виду) аналізу.
5. Вибір або розробка методів аналізу.
6. Доведення гіпотез.
7. Формулювання аналітичних висновків.
8. Написання аналітичного документа.

Для розуміння сутності та механізмів організації інформаційної діяльності у сфері управління різними системами (підприємствами, установами, громадськими організаціями та ін.)

розглянемо ключову ланку цієї форми діяльності – інформаційно-аналітичний процес у сфері підготовки та прийняття управлінських рішень.

Інформаційно-аналітичний процес у сфері бізнес-аналітики, моделювання й управління являє собою процес пошуку, збору, переробки та подання інформації у формі, придатній для її використання під час ухвалення управлінських рішень і опису різних систем.

Першою і тому надзвичайно важливою процедурою інформаційно-аналітичної роботи є первинний аналіз (експрес-аналіз) і відбір релевантної інформації. Ця процедура слугує своєрідним фільтром, що відкидає непотрібне, і захищає аналітика від інформаційного шуму (надмірності). Зміст цієї процедури полягає перш за все у встановленні сутності, важливості, точності, повноти і значущості інформації на засадах її поділу (дроблення) і зіставлення.

Основне правило інформаційно-аналітичного процесу: аналітичний продукт (огляд, довідка, звіт, модель, прогноз тощо), створений у результаті інформаційно-аналітичного процесу, повинен бути придатним для використання замовником, корисним для нього вже в момент свого створення.

Сутність інформації полягає в сукупності відображених у ній ознак об’єктів, систем, явищ і процесів, виділеної з ширшого обсягу. Інформація завжди відображає ступінь зміни існуючого знання.

Інформація є важливою, якщо вона релевантна, тобто має зв’язок з вирішенням проблеми, і якщо її використання може внести вклад у діяльність (поточну або плановану).

Інформація може бути одночасно і важливою, і марною, оскільки її може виявитися недостатньо для розуміння сутності процесу, одиничного події або явища в цілому. Те, що конкурент змінює страхове агентство, інформація важлива, але в ізоляції від інших відомостей вона не має великої цінності. Однак якщо в організації існує звичай відправляти кожен ізольований факт у «кошик для сміття», то значно скоро може статися, що цей кошик буде «розумніше» будь-кого в такій установі. Ізольований факт –

це сигнал до початку активного пошуку зв'язків – тут «зависла в повітрі» інформація стає інструментом управління процесом інформаційно-аналітичної роботи (ІАР). Її потрібно не відкидати, а співвідносити з іншою інформацією, що може принести відчутну користь.

Основні принципи організації інформаційно-аналітичного процесу, сформульовані Вашингтоном Плетом: повідомляти достовірно, своєчасно та зрозуміло.

Достовірність повідомлення є *ключовою ознакою* під час оцінки будь-якого інформаційного документа, створеного в результаті аналітичної діяльності. Вона забезпечується завдяки правильному поєднанню ряду моментів, що визначають результативність роботи аналітика. Основними серед них є:

– глибоке розуміння дійсності тим, хто аналізує інформацію;

– правильний відбір фактів, які стосуються об'єкта аналізу;

– виділення на підставі аналізу фактів, основних моментів явищ і процесів, причиново-наслідкових зв'язків.

Вирішальну роль під час підготовки достовірних аналітичних матеріалів має рівень кваліфікації виконавців робіт – професійні знання, ерудиція, вміння орієнтуватись і робити висновки в нестандартній ситуації тощо.

Не завжди легко встановити, є інформація достовірною або помилковою, особливо якщо вона містить відомості про події, які ще не відбулися. Є критерії, за якими можна судити про достовірність інформації. До цих критеріїв належать:

– критерій обґрунтованості (наявність підтверджень отриманої інформації в ряді незалежних джерел);

– критерій несуперечності:

– відсутність протиріч між окремими твердженнями, викладеними в повідомленні;

– відсутність протиріч всередині групи повідомлень, що надійшли від одного і (або) групи джерел за певний проміжок часу;

– відсутність протиріч з наявними моделями інтерпретації і моделями предметної сфери;

– критерій авторитетності джерела і (або) ступеня захищеності носія (документа).

У загальному випадку ступінь достовірності інформації оцінюється відношенням кількості помилкових елементів інформації до загальної кількості всіх елементів інформації. Однак існують і інші – методики, засновані на експертних знаннях. Такі методики застосовуються тоді, коли не може бути оцінена несуперечливість повідомлення щодо деякої сукупності моделей.

Наприклад, достовірність інформації може оцінюватися і за такою шкалою:

– вільна бесіда з консультантом-фахівцем – 90–95 %;

– питально-відповідна форма опитування партнера – 40–70 %;

– вільна розповідь про події – 25–30 %.

Так само прийнято вважати, що достовірність повідомлень про події, отриманих від одного джерела інформації, не повинна оцінюватися вище 33 %. Двох – 66 %, а достовірність інформації, отриманої в повідомленнях від трьох джерел, міститься в межах 94–99 %.

Завжди потрібно враховувати можливість внесення добре підготовленої дезінформації, призначеної для порушення процесів управління і (або) зниження якості вироблених управлінських рішень. Така інформація використовується для навмисного відволікання уваги на помилкові напрями і, як наслідок, нанесення збитку.

Отримана інформація може бути розділена на два види:

1) факти як абсолютно достовірні твердження про реальне місце події, існування об’єктів / процесів тощо;

2) оцінку фактів, що повідомляються.

У письмових роботах аналітичного характеру (аналітичні довідки, записки, доповідні, прогнози тощо) посилання на джерела часто не наводяться, хоча майже для кожного матеріалу (факту), що викладається в них, можна було б вказати джерело (із ЗМІ або літератури). Однак посилання на одну або навіть дві публікації не може нічого ні підтвердити, ні спростувати. Справа в тому, що завжди можна знайти публікацію з досить точним описом будь-якої значимої події з життя країни, але точно так само можна знайти й іншу – з фактографічними помилками або неправильною інтерпретацією фактів. Так само, як і третю, де йдеться про подію, якої не було.

Аналітичні матеріали, що публікуються різними компаніями, також не завжди бездоганні: там може не бути помилок, але зустрічаються свідомі умовчання або неоднозначні формулювання. З цієї причини інформацію в ЗМІ можна вважати досить достовірною, тільки якщо вона неодноразово повторена, причому краще, якщо ця інформація була зазначена в ЗМІ різного виду та приналежності, наприклад у газеті, а через деякий час у діловому тижневику або фаховому журналі. Найбільш важливі факти потребують підтвердження, як мінімум, з джерел двох зазначених вище типів. Однак і цього може бути замало, оскільки цитування без посилання на первинне джерело – широко поширена практика в сучасній журналістиці.

Своєчасність отримання інформації визначає її цінність для користувача. Значення своєчасного подання інформації особам, які приймають рішення, може бути настільки великим, що заради нього варто навіть дещо поступитися «абсолютною достовірністю» повідомлень. Звичайно, йдеться не про суттєву зміну змісту повідомлення (інформації), а лише про певне зниження ступеня його точності, наявності певних допустимих похибок.

Ясність викладення матеріалу. Сутність реалізації цього принципу полягає в тому, щоб зробити доступними для інших результати аналітичної роботи. Коректне, правильне сприйняття інформації робить її не лише переконливою для споживача, а і більш корисною, оскільки це сприяє досить вдалому її застосуванню, а отже, забезпечує успіх справи. Крім того, інформаційно-аналітичний документ має бути: **лаконічним, логічним і переконливим.** Коротке повідомлення легше сприймається і краще запам’ятовується, ніж довге. Воно має переваги перед довгим (великим) і з огляду на те, що час на підготовку й прийняття управлінських рішень завжди обмежений. У середовищі аналітиків широко відомо **«правило п’яти сторінок»,** яке свідчить, що під час подання аналітичних висновків на вищий рівень управлінської ієрархії будь-якої організації обсяг аналітичної довідки повинен складати не більше п’яти сторінок.

Інформаційно-аналітичний документ можна вважати логічним, якщо його автор під час викладу матеріалу дотримується

законів логіки, а саме: закону тотожності, закону протиріччя, закону виключення третього та закону достатньої підстави. Щоб викладене у документі повідомлення було достатньо переконливим, воно має відображати суттєві причиново-наслідкові зв'язки у доступній і зрозумілій для споживача формі.

Далі наведено два базових підходи до здійснення ефективної аналітичної роботи. Ці два підходи, будучи застосовані послідовно, паралельно або в циклі, дозволяють сформувати думку про значущість інформації, її зв'язок з іншими знаннями, вбудувати її в процес управління.

Перший підхід – «циркуляція». Він складається в інтенсивному обміні здобутою інформацією в групі осіб, які до неї прямо стосуються (професіоналів, експертів). Це створює передумови до скорочення витрат часу на встановлення прямого або непрямого зв'язку розрізнених повідомлень. Наприклад, за наявності каналів інформаційної взаємодії між співробітниками фірми, що відповідають за різні ділянки її діяльності, можуть отримати непряме підтвердження чутки про можливе банкрутство конкурента в результаті їх зіставлення з такими зовні не пов'язаними фактами, як зміна аудитора і зміна статутного капіталу.

Другий підхід – «синтез». Він полягає у поєднанні на манер мозаїки всіх обривків інформації для створення максимально повного уявлення про подію, діяльність конкурента. Це не завжди легко, тому що часто відсутній ряд потрібних елементів або вони невідповідної форми. Іноді, навпаки, частини підходять, даючи чітке уявлення про ситуацію.

Формулюючи гіпотезу, аналітику часто доводиться пробляти складний шлях: використовувати і циркуляцію, і синтез, пов'язуючи результати з дослідженням блоків допоміжної інформації для того, щоб домогтися адекватного, чіткого і повного відображення об'єкта (ситуації).

Для пояснення подій завжди необхідно мати хоча б дві гіпотези. Причиною багатьох катастроф є не зневага інформацією, а помилкове її розуміння.

Перш за все для керівника важливе вміння нехтувати власними упередженнями. Якщо, наприклад, противник недооцінюється, то природним наслідком цього з'явиться спроба апріорно виключити певні гіпотези про нього. Так, західні підприємці недооцінили японський економічний наступ у 70-і роки.

У таких випадках, щоб уникнути помилки у разі прийняття складного управлінського рішення, можна запропонувати принцип незалежної експертної оцінки. Його сутність полягає в тому, що для оцінювання ситуації або пропонованого рішення варто запитати думку сторонніх експертів у цій галузі, не пов'язаних службовою, фінансовою або іншою залежністю із суб'єктом управління, замовником. Це дозволить виявити абсолютно різні гіпотези, оцінки і прогнози, звірити думки, отримати відомості за певними припущеннями. В ідеальному випадку кількість експертів має забезпечувати дієвість принципу мажоритарної оцінки (як основу процедури голосування), якщо інший спосіб обґрунтування вибору неможливий. Тут включається вже елемент відповідальності керівника за вибір рішення і втілення його в життя.

Якість аналізу та інтерпретації великою мірою залежить від професійної приналежності аналітика. Найбільш підходящою особою для інтерпретації отриманої інформації є людина, що є професіоналом у відповідній функціональній сфері (виробництво, маркетинг, фінанси, дослідження і розробки тощо).

Інформація стає особливо цінною тоді, коли вона прямим або непрямим чином бере участь у виробленні рішення вищим керівництвом організації. Мистецтво аналітичної роботи полягає в тому, щоб персонал організації своєчасно отримував інформацію, яка йому дійсно необхідна для вироблення адекватних управлінських рішень. Тільки в таких умовах організація існує і розвивається, не витрачаючи зусиль на переробку надлишкової інформації.

Для людини, що не володіє навичками інформаційно-аналітичної роботи, велика кількість інформації настільки ж згубна, як і відсутність її, це негативно позначається на якості рішень.

Оцінка й аналітична обробка інформації доводиться до керівників, відповідно до сфери їхньої діяльності і рівня компетенції,

або необмежено розповсюджується серед осіб, які приймають рішення, чи рядового персоналу (це визначається принципами функціонування організації). Якщо за своїм характером інформацію відносять до стратегічного планування, то її повинні отримувати особи, які складали плани і несуть відповідальність за прийняття стратегічних рішень. Варто чітко встановити правила поширення інформації з тим, щоб забезпечити необхідну обережність, але й не відмовляти в інформації тим, хто її потребує.

З часом були сформовані основні принципи інформаційно-аналітичної діяльності та основні принципи аналітичного методу, які в багатьох аспектах збігаються з діалектичними принципами. Основні принципи аналітичної діяльності перераховано й описано вище (цілеспрямованість, системність, актуальність, своєчасність, активність, ініціативність, достовірність, об'єктивність, повнота, безперервність, альтернативність, гнучкість, обґрунтованість). З цього комплексу основних принципів аналітичної діяльності випливають частинні принципи інформаційно-аналітичної діяльності як інструменту пізнання і впливу на дійсність. До них можна віднести:

– об'єктивність і незалежність підходу;

– всебічність розгляду, виявлення основних проблем і «больових точок», їх структурування;

– системність розгляду (емерджентність, комплексність, структурність, ієрархічність, функціональність);

– пізнання сутнісної сторони явищ і процесів;

– виявлення взаємозв'язків з іншими об'єктами;

– встановлення причиново-наслідкових зв'язків (принцип детермінізму);

– порівняння та аналогія;

– виявлення основних тенденцій розвитку (зміни);

– єдність кількісних і якісних характеристик;

– єдність форми і змісту та інші.

Твердження «істина – у відтінках» досить справедливе і для аналітики. Тут, як ніде, важливо «мистецтво дрібниць». Сутність його в тому, що іноді дрібні деталі можуть під час відповідної інтерпретації бути красномовними і надати досить важливу інформацію, бути сигналом зародження нових тенденцій і процесів.

3.6. Характеристика основних різновидів ІАД

Цільове призначення аналітичного документа визначається рішенням, яке споживач має прийняти на його основі. У зв'язку з цим інформаційні потреби для вирішення конкретних завдань можуть бути різноманітними, зокрема:

– отримати уявлення про зміст первинних документів з певної проблеми з метою відбору цінних джерел для поглибленого вивчення;

– оцінити стан і шляхи розвитку відповідної галузі науки, техніки, ринків, певних організацій та процесів з метою визначення місця проблеми, яка вирішується, серед інших проблем;

– використати новітні досягнення певних галузей науки, техніки та інформаційних технологій з метою підвищення ефективності діяльності організації;

– звернути увагу на проблему, яку вивчають з метою показу позитивних і негативних сторін досліджень, що проводяться;

– оцінити інформаційну ситуацію в конкретній сфері з метою визначення найперспективніших рішень поставленого завдання;

– визначити досягнутий рівень розвитку конкретної сфери розвитку суспільства чи організації й порівняти з ним рівень виконаної роботи;

– встановити значущість ідей, теорій, концепцій, гіпотез, практичних рішень, рекомендацій з метою використання їх у вирішенні конкретного завдання;

– виявити тенденції розвитку основної і суміжних галузей та встановити оптимальні технічні й економічні показники нових об'єктів, які створюють.

Анотація – це стисла характеристика твору друку (їх сукупності або частин) з точки зору змісту, призначення, виду, форми та інших особливостей. Вона, як правило, складається на видання в цілому. Метою анотації є інформування користувачів стосовно основного предметно-тематичного змісту документа, аспектів розгляду теми, матеріалу, на базі якого проведено дослідження, його географічних аспектів та хронологічних рамок. Середній

обсяг анотації – 600 др. знаків. У фаховій літературі існує безліч класифікацій анотацій залежно від:

– видів анотованої літератури;

– цілей та функцій анотацій;

– інформативності та глибини розкриття змісту документа.

Одна із спроб типізації анотацій базується на функціональному підході, що дає змогу відрізняти анотації з урахуванням їх суспільного призначення. Згідно з вищевказаним анотації поділяються на рекомендаційні та довідкові.

Рекомендаційна анотація – дає характеристику й оцінку документа з огляду на його цільове та читацьке призначення, слугує пропаганді та рекламі кращої літератури й покликана зацікавити, привернути увагу, активізувати процес читання. Як правило, вона має дидактичну спрямованість і може вміщувати педагогічні рекомендації, методичні поради тощо.

Довідкова анотація – стисло уточнює незрозумілий заголовок і повідомляє розширені відомості про автора, а також форму, зміст, жанр, призначення й інші особливості документа, відсутні в бібліографічному описі. Довідкові анотації не містять критичної оцінки документа.

У структурі інформаційно-аналітичної діяльності вагоме значення належить **оглядово-аналітичній діяльності.** Мета цього напряму – інформаційне забезпечення управлінських рішень і створення системи інформаційної підтримки базової діяльності користувачів інформації. Основним засобом виконання цих завдань є підготовка оглядової інформації, що дозволяє систематизовано та узагальнено оцінити стан певної галузі науки, техніки, виробництва, культури, освіти тощо.

Підготовка оглядової інформації базується на основних процесах обробки документальних джерел. Основними серед них є:

– витяг з документів і систематизація відомостей та кількісних даних, які характеризують різні аспекти стану розглянутого об’єкта;

– оцінка новизни, достовірності та взаємозалежності відібраних даних, їх доповнення й уточнення змісту;

– логічна переробка отриманих даних з метою одержання нової інформації щодо стану певної галузі суспільної діяльності, тобто визначення досягнутого рівня, тенденцій і перспектив розвитку.

За результатами переробки першоджерел готуються оглядово-аналітичні документи: різні огляди, щорічні доповіді про найважливіші досягнення у певній галузі, аналітичні довідки, інформаційні повідомлення тощо.

До змісту оглядово-аналітичних документів висувають такі, вимоги: актуальність, достовірність, об'єктивність, фактографічність, наявність висновків та їх обґрунтованість, стислість.

Аналітичний огляд – це вторинний синтезований текст, в якому подано зведену характеристику певного питання чи проблеми, що базується на використанні інформації, отриманої з ряду першоджерел за певний проміжок часу. Основним чинником, який визначає зміст і форму огляду, є мета його підготовки. Залежно від неї всі огляди можна поділити на дві групи: бібліографічні та огляди стану питання.

Бібліографічний огляд вміщує впорядковану сукупність бібліографічних описів джерел інформації; анотації, що розкривають їх зміст; текст, який пов'язує між собою описи окремих джерел. Мета підготовки такого огляду – допомогти фахівцям орієнтуватися в документальному потоці та вибирати найцікавіші джерела для вивчення.

Огляди стану питання містять інформацію із документів-першоджерел про результати досліджень, стан і шляхи розвитку певної галузі науки, техніки та виробництва. Залежно від цілей підготовки розрізняють два види таких оглядів: *реферативні та аналітичні.*

Реферативний огляд вміщує систематизовані та узагальнені відомості щодо стану і шляхів розвитку певної галузі науки, техніки, виробництва тощо. Ці дані отримано внаслідок аналізу першоджерел, але без їх критичного опрацювання. Збережено висновки та оцінки авторів першоджерел щодо фактів, теорій, методів, інструментів, про які йдеться у документах. Мета підготовки реферативного огляду – надання чіткого уявлення про стан робіт

і висвітлення основних результатів досліджень у певній сфері, орієнтування в сукупності фактів, отриманих фахівцями.

Найбільш складним і синтетичним видом вторинної інформації є науково-аналітичні огляди. Вони вміщують систематизовані й узагальнені відомості про стан, тенденції та перспективи розвитку певної галузі суспільної діяльності чи окремої проблеми.

Аналітичний огляд як інформаційний твір може бути не тільки узагальнюючою характеристикою вихідних джерел, а й вміщувати висновки та рекомендації для розв'язання проблеми. Як інформаційна основа прийняття управлінських рішень і результат інформаційно-аналітичної роботи, він широко застосовується в економічній та політичній сферах. Існує багато різновидів інформаційних документів, складених з використанням його методики. До них можна віднести аналітичні довідки, інформаційні довідки, аналітичні звіти, дайджести тощо.

Названі документи, що містять прогноз розвитку проблемної ситуації, є засобом інформаційної підтримки прийняття управлінських рішень. Огляди дозволяють зекономити час фахівців, позбавляють їх необхідності безпосереднього перегляду документів під час пошуку потрібних матеріалів, акцентують увагу на особливо цікавих документах.

За цільовим призначенням та методикою створення розрізняють такі види оглядових документів: експрес-інформація, бібліографічний огляд, реферативний огляд, огляд обґрунтування, щорічний огляд, щорічна доповідь, оглядова довідка тощо.

Інформаційний огляд складають шляхом логічної обробки документної інформації з метою одержання знання про стан, розвиток і можливі шляхи вирішення проблеми. Підготовка оглядової інформації – це не стільки опис, узагальнення документної інформації, скільки аналіз стану та розвитку проблеми на основі використання різноманітних джерел інформації.

Отримання оглядової інформації базується на таких процесах обробки документальних джерел інформації:

– вилучення з документів, відбір і систематизація системи даних і фактів, які характеризують основні аспекти стану об'єкта, що розглядається;

– доповнення їх, уточнення змісту, оцінка новизни і значущості, перевірка достовірності і встановлення взаємозалежності;

– логічне перетворення даних і фактів з метою отримання нової інформації про стан об'єкта, тобто знання про досягнутий об'єктом рівень, організаційно-економічну ситуацію і тенденції розвитку.

Інформаційний огляд містить систематизовані дані і факти, узагальнену інформацію про стан розробки питання без їх критичної (наукової, технічної, економічної) оцінки автором огляду. Основне призначення інформаційного огляду – орієнтування споживачів інформації на ознайомлення зі станом розробки проблеми, питання.

Огляд обґрунтування призначається для доведення необхідності й доцільності проведення конкретного дослідження або розробки, вибору оптимальних шляхів (напрямів, методики, принципового рішення), організації і планування роботи.

Оглядова доповідь містить відомості про окремі аспекти стану науково-практичних напрямів, проблем, об'єктів, організацій тощо. Вона має суттєве значення для оцінки поточних змін стану питання (проблеми, теми), що розглядається.

Бібліографічний огляд – це огляд, який містить пов'язану і послідовну бібліографічну характеристику сукупності документних джерел з певного питання (проблеми, теми), що вийшли за певний період, а також бібліографічні описи цих документів. Бібліографічний огляд містить зведену характеристику документів, виконану на основі документографічного аналізу. У бібліографічному огляді подають узагальнений опис змістових елементів, а також наводять формальні ознаки, характерні для цих видань.

Огляд стану питання – це огляд, що містить систематизовану й узагальнену інформацію про стан досліджуваної проблеми і практичної реалізації, тенденції розвитку питання, що розглядається. Огляд стану питання (проблеми, теми) складається з таких елементів: вихідні дані, реферат, вступ, аналітична частина, висновки, рекомендації, додатки, список використаних джерел, список скорочень, допоміжні покажчики, зміст.

Особливу роль відіграють оглядово-аналітичні документи в ухваленні виважених управлінських рішень.

Аналітичний прогноз містить аналіз інформації, яка відображає характер змін стану досліджуваного об'єкта (його структури, найважливіших показників і чинників, що визначають його розвиток), з метою виявлення закономірностей розвитку об'єкта, необхідних для проведення робіт з прогнозування. Цей огляд базується на аналізі первинних документів глибокої ретроспективи.

Прогнозний огляд складають для інформування про перспективи розвитку об'єкта на 5–10 років наперед і передбачають аналіз. Прогнозний огляд має таку структуру: вступ (підстави для складання огляду), аналітична частина (стан об'єкта, тенденції і прогноз розвитку), висновки, пропозиції.

Аналітичний звіт про діяльність розпочинається з анотації або реферата з коротким викладом завдання дослідження та отриманих результатів і зі вступу, в якому характеризуються вітчизняні й зарубіжні досягнення з досліджуваної проблеми. У самому тексті звіту викладають зміст завдання, формулюють технічне завдання, аналізують відомі методи і способи вирішення, розрахунки і результати експериментів. Завершується звіт висновками із зіставленням та аналізом результатів, отриманих у ході дослідження теоретичних та експериментальних даних, і, нарешті, висновками з оцінкою результатів і шляхів їх використання.

Модифікацією аналітичного документа за цільовим призначенням та методикою створення є такі інформаційні документи, як тематична підбірка, дайджест, прес-реліз, досьє та ін. Такі документи часто називають синтезованими.

Тематичні підбірки як вид аналітичних документів – це синтезована сукупність первинних документів, частин тексту, рефератів, анотацій, рисунків, схем, діаграм тощо, підібраних з певного питання (проблеми), їх роблять за двома умовами: у разі сталого споживацького інтересу до певної проблеми і коли немає узагальнених публікацій з неї, інформація розосереджена в широкому колі джерел.

Структура тематичної підбірки має, як правило, такий вигляд: титульна сторінка, на якій, крім теми, вказують хронологічні межі відбору документів або рік початку виявлення інформації,

якщо тематична підбірка поповнюється новими матеріалами; зміст; коротка передмова (від укладача); певним чином згруповані копії текстів документів, інколи ілюстрований матеріал; бібліографічний список усіх виявлених і включених до підбірки копій джерел; список переглянутих джерел.

Близьким за характером аналізу і подання інформації є дайджест. *Дайджест* – це документ, що становить добірку витягів з конкретного тексту, відібраних і згрупованих таким чином, щоб дати про нього загальне уявлення, чи добірку найцікавіших матеріалів, передрукованих з інших видань. Дайджести можуть бути неперіодичними, періодичними та продовжуваними. Дайджестом також називають короткий виклад будь-чого. Дайджест – це вторинний документ, результат аналітико-синтетичної обробки первинних документів, який містить новий текст «під завдання»: під рішення, яке приймає керівник; під навчальне завдання, запропоноване студентам; під потреби педагога, бізнесмена, які мають використати новий матеріал тощо. Дайджест має проблемно-тематичне спрямування. Структура дайджесту має такий вигляд: титульна сторінка; зміст; передмова (від укладача); основний текст, звично розділений на параграфи; список використаних джерел; додатки; ілюстративний матеріал (таблиці, графіки, діаграми, які розкривають основний текст); словник або глосарій основних термінів. Дайджести створюють у друкованій або електронній формі на замовлення споживачів.

Прикладом аналітичного вторинного документа є досьє. *Досьє* – це формалізований, тобто складений за певним планом вторинний документ, який містить розгорнуту характеристику об’єкта (організації, товару або послуги тощо). Досьє створюють шляхом аналітико-синтетичної переробки значної кількості первинних документів і постійно актуалізують, оскільки фактичні дані швидко змінюються.

Аналогічними за змістом досьє про організацію (установу) є прес-релізи. *Прес-реліз* – листівка, спеціальний бюлетень або брошура переважно з офіційним повідомленням для органів масової інформації.

До прес-релізів висувають такі інформаційні вимоги:

– зацікавленість споживачем (інформативність);

– змістовність (інформаційна місткість);

– яскравість у рекламному оточенні (контрасність);

– виразність основних елементів (первинна фіксація уваги);

– запам'ятовуваність (вторинна фіксація уваги – зворотна інформація);

– вчасність – необхідність для певних споживачів у потрібний час.

Аналітичні документи містять нову узагальнену інформацію, отриману в результаті всебічного, глибокого і критичного аналізу первинних документів, аргументовану оцінку стану і тенденцій розвитку проблеми, що розглядається. Їх створюють у процесі поглибленого аналізу і синтезу первинних документів з метою вилучення, оцінки, узагальнення і використання інформації, що в них міститься.

Однією з характерних рис аналітичного документа є наявність у ньому «аналітичної частини», яку становлять як концепції, погляди, ідеї, висновки, запозичені з первинного документа, так і власні думки, ідеї, судження автора вторинного документа.

Деякі аналітичні документи дають можливість уявити інформаційну модель проблеми. Це перш за все інформаційно-аналітична довідка, інформаційний звіт про діяльність, аналітичний огляд тощо. Для створення таких документів застосовують як загальнонаукові методи, так і інформаційно-прогностичні методики (моделювання, контент-аналіз аналогій та ін.).

3.7. Проблема інтерпретації фактів та типові помилки в аналітичних висновках

Мислити логічно – це означає мислити відповідно до правил і законів логіки. Аналізуючи інформаційні процеси і явища, застосовують такі основні логічні закони.

Закон тотожності. У межах цього закону предмет думки одного міркування, одного доказу, однієї теорії повинен залишатися незмінним. Закон тотожності вимагає, щоб у процесі міркування одне знання про предмет не підмінялося на інше. Якщо ця вимога порушується, то руйнуються логічні зв'язки.

Закон тотожності має практичне значення. Так, починаючи обговорювати якийсь предмет, необхідно бути впевненим, що обидві сторони вкладають у нього один і той самий зміст. Якщо співрозмовники не сходяться в розумінні термінів і по-різному трактують їх, обговорення стає беззмістовним. Закон тотожності спрямований і на тих осіб, які досить нечітко і неконкретно свідчать про якийсь факт чи ситуацію.

Закон протиріччя. Згідно з цим законом не можуть бути однаково правильними (істинними) обидва висловлювання (повідомлення), одне з яких будь-що стверджує, а друге – заперечує в той самий час. Наприклад, описуючи у звіті фінансово-виробничий стан за попередній календарний рік, автор повідомляє, що «організація досягла найвищих показників цього року. Не менш високих показників було досягнуто і минулого року». У цьому випадку перше твердження заперечується таким. Це – нонсенс, який завдає шкоди не тільки авторитету автора документа, а й самій організації, якщо документ призначений для виходу за її межі. Варто зазначити, що закон протиріччя не дає теоретичної відповіді на запитання, яке із суджень правильне. Достовірність вивіряється тільки практикою.

Закон виключення третього. Основний зміст цього закону логіки полягає в такому: якщо є два судження, які суперечать одне одному, то одне із суджень обов'язково правильне, а друге – неправильне. Отже, між ними не може бути середини – третього судження. Наприклад, у тексті документа може бути повідомлено, що «організація виконала план повністю, але недовиконаною залишилася ланка роботи...»

Закон достатньої підстави. Цей закон формулюється так: для того щоб визнати судження про предмет істинним, повинні бути вказані достатні підстави для цього. Не досить тільки стверджувати будь-що, необхідно думку (точку зору) обґрунтувати. Іншими словами, підтвердити її фактами, інакше висловлювання залишиться голослівним.

Мислення, в яких допущені логічні помилки внаслідок недотримання законів та правил логіки, називаються неправильними. Логічні помилки у процесі мислення можуть припускатися як ненавмисно (через незнання), так і навмисно, з метою введення в оману

опонента, обґрунтування неправдивого твердження, якоїсь нісенітниці тощо. У першому випадку логічна помилка називається **паралогізмом,** у другому – **софізмом.** Відмінною рисою софізму є те, що він, будучи неправильним мисленням, видається за правильне.

Типові помилки під час написання та оформлення аналітичної роботи:

1. Зміст роботи не відповідає плану роботи або не розкриває теми.

2. Сформульовані розділи (підрозділи) не відбивають реальну проблемну ситуацію, стан об'єкта.

3. Мета дослідження не пов'язана з проблемою, сформульована абстрактно і не відбиває специфіці об'єкта і предмета дослідження.

4. Автор не виявив самостійності.

5. Не зроблено глибокого і всебічного аналізу сучасних офіційних і нормативних документів, нової спеціальної літератури з теми дослідження.

6. Аналітичний огляд вітчизняних і зарубіжних публікацій з теми роботи має форму анотованого списку і не відбиває рівня дослідження проблеми.

7. Не розкриті зміст та організація експериментального дослідження (його сутність, тривалість, місце проведення, кількість обстежуваних, їхні характеристики), поверхово висвітлені практичні питання.

8. Кінцевий результат не відповідає меті дослідження, висновки не відповідають поставленим завданням.

9. У роботі немає посилань на першоджерела або вказані не ті, з яких запозичено матеріал.

10. Бібліографічний опис джерел у списку використаної літератури наведено довільно, без додержання вимог державного стандарту.

11. Обсяг та оформлення роботи не відповідають вимогам замовника, вона виконана неохайно, з помилками.

Якість аналізу і «правдивість» аналітичних висновків будуть залежати насамперед від якості джерел інформації. Помилки, неузгодженості та всі інші недоліки можуть відбитися на аналітичних розрахунках, спотворити результати аналізу. Тому попередньою

умовою аналізу є ретельна перевірка всіх залучених до аналізу матеріалів. Підготовка даних до аналізу полягає переважно у їхній перевірці, у перевірці якості джерел інформації.

Перевірку *джерел інформації* проводять з метою встановлення:

– правильності оформлення звітів, підрахунків і обчислень;

– повноти, якості і термінів проведення звітності, відображення її результатів в обліку і заходах, вжитих щодо виявлених недоліків;

– взаємного ув'язування показників, порівнянності звітних показників.

Документи цільової спрямованості, досліджувані аналітиком, будуть надійними у тому випадку, коли передбачені звичайні операції контролю:

– джерело інформації повинно бути незалежним;

– потрібно повторно звертатися до одного і того самого джерела даних (різні джерела мають бути узгоджені між собою або доповнювати один одного);

– домінуючою повинна бути актуальна інформація.

Думки, оцінки, погляди потенційно мають менший рівень вірогідності і надійності, ніж інформація про конкретну економічну ситуацію, підтверджену конкретними кількісними показниками. Саме конкретна досліджувана економічна ситуація є ключем до розкодування змісту поглядів, думок, що констатують економічний стан.

3.8. Інформаційно-аналітичні технології та методи організації ІАД

Будь-яка аналітична робота забезпечується комплексом технології обробки інформації, заснованих на цілісній методологічній базі. Інформаційні технології – це невід'ємна частина сучасних технологій у цілому, перш за все прогностичних, освітніх, технічних, комп'ютерних та інших. Однак усі ці технології залежать від людини-аналітика, здатної продукувати нові ідеї. Без такого суб'єкта – генератора і виробника ідей – уся інформаційна робота нічого не варта.

Рівень розвитку інформаційних технологій, безумовно, відіграє найважливішу роль у всіх сферах діяльності людини, перш за все у сфері управління, економіки, політики, технічних ресурсів, гуманітарних галузях науки.

Сутність інформаційних технологій полягає в тому, що вони покликані відповідним чином підготувати масиви даних для подальшого їх вивчення аналітиком. Сутність інформаційно-аналітичних технологій полягає в творчому, інтелектуальному змісті аналітичних процедур, які здійснює аналітик або особа, яка приймає рішення, після отримання набору даних.

Інформаційно-аналітичні технології (далі – ІАТ) можна визначити як систему знань, методів, операцій і правил, що дозволяють на основі залучення енергетичних, сировинних, технічних, інтелектуальних, кадрових, організаційних, інформаційних та інших ресурсів забезпечити найбільшу ефективність певного виду діяльності. Найважливішою особливістю ІАТ є їх міждисциплінарний характер, вони перебувають на «стику» ряду наукових дисциплін, часом слабопов'язаних між собою. До них можна віднести філософію, соціологію, логіку, математику, економічну науку, інформатику, управлінську науку, психологію та інші галузі науки.

Інформаційно-аналітичні технології – це не тільки плани і проєкти творчої спрямованості, вони можуть бути і потужною зброєю. Однією з різновидів ІАТ є і технології проведення інформаційних акцій деструктивної спрямованості. Інформаційно-аналітичні технології є і інструментом нападу, й інструментом протидії деструктивним інформаційним акціям. Можна зазначити, що «ІАТ – це і отрута, і протиотрута».

Важливим фактором особистості аналітика є його соціальний і особистий досвід. Як у жодній іншій сфері, в аналітиці надзвичайну роль відіграє суб'єктивний фактор. Знання, досвід та інтуїція аналітика, його професіоналізм і громадянська відповідальність слугують підставою для аналітичних оцінок і висновків.

Нині в літературі використовується термін «інтелектуальні технології». Фактично цей термін за своїм змістом рівнозначний терміну «інформаційно-аналітичні технології».

Інформаційно-аналітичні технології дозволяють:

1. Знизити навантаження на особу, яка приймає рішення, за рахунок здійснення первинної фільтрації потоку даних.

2. Звільнити аналітика від рутинної роботи із систематизації, реферування даних.

3. Підвищити системність сприйняття даних про обстановку за рахунок введення процедур інтеграції інформації.

4. Забезпечити перетворення даних про процес до виду, більш зручному для сприйняття аналітиком.

5. Автоматизувати пошук аргументації на підтвердження або спростування висунутих аналітиком гіпотез.

6. Забезпечити автоматичну індикацію логічно суперечних даних (за наявності відповідних процедур їх формалізації).

7. Забезпечити автоматичну індикацію очікуваних подій за рахунок постановки завдань автоматизованим системам.

8. Підвищити ефективність інформаційно-аналітичної роботи за рахунок алгоритмізації та стандартизації аналітичних процедур.

9. Забезпечити більш високу психологічну стійкість експерта-аналітика до стресових впливів за рахунок застосування спеціальних методик.

Як видно з переліку функцій, ІАТ – це не тільки і не стільки програмні й апаратні засоби обробки даних і сигналів, скільки специфічні методики, придатні як для реалізації їх із застосуванням засобами автоматизації, так і для використання поза комплексом засобів автоматизації. У першу чергу, ІАТ застосовуються в інтересах зниження напруженості і підвищення ефективності інтелектуальної праці, що часто досягається за рахунок алгоритмізації діяльності аналітика.

Тобто в рамках ІАТ можуть бути виділені два компонента: автоматизація діяльності (інструментальний напрям) і алгоритмізація діяльності аналітика (методологічно-організаційний напрям).

Завдяки цьому ІАТ дозволяють ефективно вирішувати різноманітні практичні завдання у найрізноманітніших сферах діяльності: у науці, техніці, економіці, соціальній сфері, політиці, гуманітарних галузях та інших. На їх основі проводяться: аналіз соціально-політичної ситуації, обґрунтування вибору політичних

або економічних рішень на різних рівнях владних повноважень, прогнозування, а також вирішується багато інших проблем, що полягають у площині соціально-економічних процесів.

У цьому сенсі ІАТ спрямовані на забезпечення діяльності різних осіб, які приймають рішення в умовах дефіциту часу, неповноти відомостей про досліджуваних процесах, нечіткості, суперечливості чи недостовірності інформації.

3.9. Основні методи аналітики та організації ІАД

Історичний метод передбачає розгляд об'єктивного процесу розвитку об'єкта, реальної його історії з усіма її поворотами, особливостями; це певний спосіб відтворення в мисленні історичного процесу в його хронологічній послідовності та конкретності. Завданням історичного дослідження є розкриття конкретних умов розвитку певних явищ.

Аналіз – поділ об'єкта (подумки чи реально) на складові з метою їх окремого вивчення. Такими частинами можуть бути якісь матеріальні елементи об'єкта або ж його властивості, ознаки, зв'язки й тощо. Аналіз – необхідний етап у пізнанні об'єкта.

Для розуміння об'єкта як єдиного цілого не можна обмежуватися вивченням лише його складових. У процесі пізнання необхідно розкривати об'єктивно існуючі зв'язки між ними, розглядати їх у єдності. Здійснити цей другий етап у процесі пізнання – перейти від вивчення окремих складових об'єкта до вивчення його як єдиного нерозривного цілого – можна тільки в тому випадку, якщо метод аналізу доповнюється іншим методом – *синтезом.* У процесі синтезу відбувається поєднання воєдино складових (властивостей, ознак, зв'язків тощо) досліджуваного об'єкта, розчленованих у результаті аналізу. На основі цього відбувається подальше вивчення об'єкта, але вже як єдиного цілого. При цьому синтез не означає простого механічного поєднання роз'єднаних елементів у єдину систему. Він розкриває місце й роль кожного елемента в цілісній системі, установлює взаємозв'язок і взаємозалежність між ними, тобто дозволяє зрозуміти справжню діалектичну єдність досліджуваного об'єкта.

Метод аналогії означає перенесення властивості (сукупності властивостей) одного явища на інше. У певному значенні аналогія є різновидом порівняння, оскільки вона спрямована на вивчення подібності та відмінності явищ. У багатьох випадках висновки за аналогією не можуть бути доказом, але вони мають силу припущення і дають підстави для висунення гіпотез.

Гіпотеза – це науково обґрунтоване припущення про можливі зв'язки, відносини, причини явищ.

Спрощеним методом аналогії є *відсотковий метод*. Виходячи з відомих даних, які характеризують сьогоднішнє становище (ВВП, податкові ставки, обсяги торгівлі тощо), можна визначити відповідні показники на майбутнє. Цей метод стає ефективним за умов того, що попередньо буде вивчено всю сукупність явищ, які зачіпаються. Лише після цього можна дійти висновку про те, що порівнювані явища такі подібні, що дають можливість вдатися до порівняння. Найкращі результати відсотковий метод дає під час обробки даних соціологічних опитувань. Так, у разі постійного моніторингу рейтингу політичних діячів або соціологічних опитувань з приводу певних подій можна розробити оптимальну політичну програму чи виявити вразливі місця в політиці держави.

Порівняння – це процес зіставлення предметів або явищ дійсності з метою встановлення подібності чи відмінності між ними, а також знаходження загального, притаманного, що може бути властивим двом або кільком об'єктам дослідження. Метод порівняння буде плідним, якщо під час його застосування виконуються такі вимоги:

– порівнюватись можуть тільки такі явища, між якими може існувати певна об'єктивна спільність;

– порівняння повинно здійснюватися за найважливішими, найсуттєвішими (щодо конкретного завдання) ознаками.

Узагальнення – логічний процес переходу від одиничного до загального чи від менш загального до більш загального знання, а також продукт розумової діяльності, форма відображення загальних ознак і якостей об'єктивних явищ.

Найпростіші узагальнення полягають в об'єднанні, групуванні об'єктів на основі окремої ознаки (синкретичні об'єднання). Складнішим є комплексне узагальнення, під час якого група об'єктів

з різними основами об'єднується в єдине ціле. Здійснюється узагальнення шляхом абстрагування від специфічних і виявлення загальних ознак (властивостей, відношень тощо), притаманних певним предметам.

Метод мережі зв'язків – для графічного зображення асоціацій між людьми й організаціями використовують такі аналітичні інструменти, як складання матриці асоціацій і схеми мережі зв'язків. Метод мережі зв'язків може бути корисний для встановлення оптимальних організаційних структур установ, урядів, у разі вибору ймовірних керівників вищої та середньої ланки.

Процес аналізу складається з таких етапів: підбір інформації, виділення даних, які стосуються індивідуумів і зв'язків, розробка матриці асоціацій та попередньої схеми зв'язків, внесення до схеми організаційних накладень і уточнення схеми.

Зміст повідомлень, тобто значення цінностей, які є предметом масової комунікації, досліджується, звичайно, за допомогою контент-аналізу. *Контент-аналіз* – це формалізований метод інтерпретації текстів.

Спостереження – це систематичне цілеспрямоване, спеціально організоване сприймання предметів і явищ об'єктивної дійсності, які виступають об'єктами дослідження. Як метод наукового пізнання, спостереження дає можливість одержувати первинну інформацію у вигляді сукупності емпіричних тверджень. Емпірична сукупність стає основою попередньої систематизації об'єктів реальності, роблячи їх вихідними об'єктами наукового дослідження.

Спостереження має відповідати таким вимогам:

– передбачуваності заздалегідь (спостереження проводиться для певного, чітко поставленого завдання);

– планомірності (виконується за планом, складеним відповідно до завдання спостереження);

– цілеспрямованості (спостерігаються лише певні аспекти явища, які викликають інтерес під час дослідження);

– вибірковості (спостерігач активно шукає потрібні об'єкти, риси, явища);

– системності (спостереження ведеться безперервно або за певною системою).

Метод виключень ґрунтується на поступовому виключенні значної групи складових на підставі закону виключення третього, значення якого полягає в тому, що справа йде так, як описується у висловлюванні, або так, як зазначає заперечення, і третього не дано.

Метод пошуку закономірності є пошуком стійкої і невипадкової характеристики або зв'язків між явищами.

Метод класифікацій – впорядкування об'єктів за істотними ознаками певних класів.

Метод моделювання – один із найважливіших в аналітиці. Передбачає побудову моделі, яка відображає істотні аспекти об'єкта, що аналізується, спрощує його і заміняє реальний об'єкт аналізу.

Системний метод – приведення даних у деяку систему, що дозволяє їх пояснити з позицій системного підходу.

Метод екстраполяції – поширення висновків, отриманих зі спостереження за однією частиною явища на іншу його частину.

Метод ідеалізації – уявна процедура, пов'язана з уявленням будь-чого як ідеалу і подальшим порівнянням реального об'єкта з ідеальним.

Метод емпатії – входження аналітика в образ аналізованого об'єкта, уявлення себе тим, хто вивчається, осмислення з його позицій дій, що ним здійснюються.

Методи організації інформаційно-аналітичної діяльності ототожнюються з іншими науковими методами, проте вони мають і свою специфіку – використовуються, як правило, під час вирішення певної проблемної ситуації, часто складних і неструктурованих завдань.

Метод експертних оцінок – узагальнена думка групи експертів приймається як розв'язання проблеми. Виділяють індивідуальні та колективні методи експертної оцінки.

Індивідуальні експертні методи засновані на використанні думок незалежних експертів-фахівців відповідного профілю.

Інтерв'ю – бесіда аналітика з експертом, у ході якої аналітик, відповідно до заздалегідь розробленої програми, ставить перед експертом питання щодо перспектив розвитку прогнозованого об'єкта.

Аналітичні експертні оцінки передбачають тривалу та ретельну самостійну роботу експерта над аналізом тенденцій, оцінкою стану й шляхів розвитку об'єкта прогнозу.

Колективні експертні оцінки ґрунтуються на принципах виявлення колективної думки експертів про перспективи розвитку об'єкта аналізу. Нині поширення набули експертні методи, засновані на роботі спеціальних комісій, – обговорення групою експертів за круглим столом проблеми для узгодження думок та генерування єдиної думки. Груповий підхід підвищує ймовірність вибору найбільш правильного рішення.

Вимоги до учасників колективної експертизи:

– високий рівень загальної ерудиції;

– глибокі спеціальні знання в галузі, що досліджується;

– наявність психологічної настанови на майбутнє;

– наявність виробничого та (чи) дослідницького досвіду в галузі, що досліджується.

Під час статистичної обробки результатів експертних оцінок у вигляді кількісних даних, що містяться в анкетах, визначаються статистичні оцінки прогнозованих характеристик, статистичні оцінки погодженості думок експертів. На підставі розрахункових значень за відповідними таблицями визначається рівень значимості здобутих даних. У разі обробки результатів експертних оцінок за відносною важливістю напрямів середнє значення та коефіцієнт варіації обчислюються для кожної альтернативи.

Метод мозкового штурму – це сукупність прийомів отримання нових ідей завдяки творчій співпраці організованої групи фахівців. Кращі результати досягаються тоді, коли група працює в складі п'яти-десяти осіб упродовж обмеженого часу (одна-дві години). Керує «мозковим штурмом» фахівець із досвідом проведення подібних заходів.

Різновиди методу

Пряма «мозкова атака» – метод колективного генерування ідей щодо розв'язання творчого завдання, мета якого – добір ідей.

Правила для учасників:

1) учасники сідають за стіл обличчям один до одного;

2) заборонені суперечки, критика, оцінки того, що висловлюються;

3) час виступу для учасника становить 1–2 хв;

4) висловлюються будь-які ідеї, навіть абсурдні;

5) кількість ідей важливіша за їх якість.

Зустрічна «мозкова атака» – припускає не генерацію нових ідей, а критику вже наявних.

Правила для учасників:

1) критика викладається лаконічно, позитивно;

2) ідеї, що вимагають тривалого часу для обговорення, розглядаються пізніше; виступати кожному можна багато разів, але краще по колу;

3) тривалість виступу – 15–20 хв.

Метод «подвійної мозкової атаки» передбачає як генерацію нових ідей, так і критику вже наявних.

Правила для учасників:

1) кількість учасників: 30–40 осіб;

2) тривалість роботи 2,5–3 год, включаючи 45-хвилинну перерву;

3) всім учасникам розсилаються письмові запрошення зі вказівкою мети «мозкової атаки» та роз'ясненням їхньої ролі у вирішенні завдання;

4) у перерві, яка є складовою роботи, ідеї уточнюються й обговорюються, відбувається генерування нових ідей;

5) виступ усіх без винятку учасників суворо регламентується;

6) ідеї оцінюються експертами та рекомендуються до впровадження.

Метод колективної генерації спрямований на одержання великої кількості ідей, у тому числі і від осіб, що, володіючи досить високим ступенем ерудиції, звичайно, утримуються від висловлень.

До найважливіших недоліків методу колективної генерації ідей відносять значний рівень інформаційного шуму, створюваного тривіальними ідеями, спонтанний і стихійний характер генерації ідей.

Метод дискусії передбачає підготовку рішень із залученням широкого кола учасників (не більше 20 осіб). Дискусія проводиться як відкрите колективне обговорення проблеми, основним завданням якого є всебічний аналіз усіх факторів, позитивних і негативних наслідків, з'ясування позицій учасників, узгодження та інтеграція. У ході дискусії дозволяється критика.

Метод ключових запитань доцільно застосовувати для збору додаткової інформації в умовах проблемної ситуації. Запитання, що ставляться, слугують стимулом у формуванні стратегії й тактики вирішення завдання, розвивають інтуїцію, формують алгоритми мислення, наводять на ідею вирішення, спонукають до правильних відповідей.

Метод аналогії заснований на багаторазовому використанні зафіксованого в банку даних організації досвіду розв'язання певних ситуацій, що мали місце на фірмі, та прийнятих за ними рішень.

Прогнозування методом аналогії ґрунтується на подібності між аналізованими об'єктами. Застосування методу аналогії потребує всебічного аналізу місцевих умов (економічних, природно-кліматичних, демографічних) для того, щоб запобігти механічного запозичення. Його застосовують також у разі, якщо немає ніякої інформації про проблему, що досліджується.

Метод Дельфі. Експертам (5–20 осіб) пропонуються питання та формулювання відповідей без аргументації. Для подання пропозицій встановлюється часовий ліміт – приблизно 14 днів. Після того, як усі відповіді надійшли, пропозиції, що містяться в них, підсумовуються, ті, що повторюються, – відсіваються. Метод найбільш складний і тривалий у часі. Основна перевага методу – незалежність думки експертів, що перебувають у просторовому віддаленні один від одного. Під час опитування зберігається анонімність відповідей експертів.

Методи прогнозування економічних процесів. Методи економічного прогнозування – це сукупність способів і прийомів розробки прогнозів, які дозволяють на основі аналізу даних ретроспективного періоду, зовнішніх і внутрішніх факторів впливу, а також їх кількісних змін здійснити переконливі передбачення стосовно майбутнього розвитку економіки чи суспільства в цілому. Для прогнозування використовується статистична інформація, яка описує процеси за минулі роки, та проводиться експертна оцінка тенденцій змін макроекономічних показників. Об'єктами прогнозування є економіка, її галузі, регіони, форми власності тощо. Для вибору методу прогнозування варто визначити мету й завдання прогнозу та період, на який він формується, врахувати

специфіку об'єкта прогнозування, види, повноту та вірогідність вхідної інформації, а також ряд інших факторів.

Методи прогнозування мають відповідати таким вимогам: поєднання суб'єктивної цінності й об'єктивної значущості оцінок; чітке застосування оцінок, яке не допускає різних тлумачень щодо вибору методів; створення можливості нагромадження статистичної інформації та її використання для прогнозування. За ступенем формалізації методи економічного прогнозування можна розділити на інтуїтивні (експертні) та формалізовані.

Запитання до параграфа 3

1. Що таке інформація?
2. Які властивості інформації?
3. Які особливості інформації як товару?
4. Які є параметри інформації?
5. Що є результатом інформаційно-аналітичної діяльності?
6. Які є види інформації?
7. Які є джерела інформації?
8. Які є базові моделі, що застосовуються під час аналізу інформаційних явищ?
9. Що таке інформаційно-аналітична діяльність?
10. Які цілі інформаційно-аналітичної діяльності?
11. Що є об'єктами інформаційно-аналітичної діяльності?
12. Які є суб'єкти інформаційно-аналітичної діяльності?
13. Яка схема універсальної технології аналітичної діяльності?
14. Які основні принципи організації інформаційно-аналітичного процесу?
15. Які є критерії достовірності інформації?
16. Які є частинні принципи інформаційно-аналітичної діяльності?
17. Які є види інформаційних потреб?
18. Що таке анотація?
19. Що таке аналітичний огляд?
20. Яка сутність інформаційного огляду?
21. Що таке огляд стану питання?
22. Що таке аналітичний документ?
23. Які є види аналітичних документів?

24. Які є проблеми в інтерпритації фактів?

25. Які є закони логіки?

26. Які існують типові помилки під час написання та оформлення аналітичної роботи?

27. Що таке інформаційно-аналітичні технології?

28. Які цілі інформаційно-аналітичних технологій?

29. Які існують основні методи організації інформаційно-аналітичної діяльності?

§ 4. Основи моделювання економічних процесів

4.1. Сутність понять «модель» і «моделювання»

Функціонування економічного механізму на практиці досить складне й багатопланове, до того ж економічна ситуація на практиці ніколи повністю не повторюється, а тому неможливо повністю застосувати одну й ту ж стратегію для двох, на перший погляд, нібито подібних ситуацій. Тому для розгляду основних принципів і виявлення певних закономірностей, які допомогли б під час аналізу складних економічних реалій, економісти використовують спрощені описи економічних явищ – економічні моделі.

Моделювання є однією з основних категорій теорії пізнання, воно давно і широко застосовується у процесі дослідження ринків праці: національних, галузевих тощо. Сутність моделювання полягає в дослідженні будь-яких явищ, процесів або систем об'єктів не безпосередньо, а шляхом побудови і вивчення їхніх моделей[13]. У широкому розумінні модель – це штучна система, яка відображає основні властивості досліджуваного об'єкта – оригінала, відображення в зручній формі багаточисельної інформації про досліджуваний об'єкт[14]. Побудова моделі досліджуваного об'єкта є опосередкованим методом його наукового дослідження, коли його

---

[13] Тимохин В. М. Теоретико-методологічні засади моделювання економічної динаміки : автореф. дис. на здобуття наук. ступеня доктора економ. наук : спец. 08.00.11 «Математичні методи, моделі та інформаційні технології в економіці». Д., 2007. 36 с.

[14] Сиденко В. М., Грушко И. М. Основы научных исследований. Харьков : Выща школа, 1977. 200 с.

безпосереднє вивчення з певних причин неможливе або недоцільне. Очевидним є той факт, що економічні об'єкти практично завжди вимагають непрямого дослідження, а тому значення моделювання тут важко переоцінити. Причини використання модельного підходу в економіці різні, найголовнішими серед них є[15]:

– недоступність об'єкта-оригіналу або небезпечність безпосереднього його дослідження;

– неможливість інколи сподіватись на повторення експерименту;

– можливість вибирати або будувати для вивчення такі моделі, які були б значно простіші, ніж досліджуваний об'єкт;

– можливість виявити найсуттєвіші фактори, які «відповідають» за певні властивості цього об'єкта, оскільки в моделі відображаються лише деякі найбільш важливі сторони і зв'язки досліджуваного об'єкта;

– можливість за допомогою економічної моделі вивчити особливості функціонування економічного об'єкта і на основі цього здійснювати прогноз стосовно подальшого його розвитку.

Модель об'єкта і сам об'єкт перебувають у відношенні подібності, а висновки, які зроблені під час вивчення моделі, переносяться на оригінал на основі використання аналогії і теорії подібності. У науковій літературі можна зустріти різні визначення поняття «модель»[16]. Найбільш широким є таке: «Модель – це певний стан процесу створення з існуючих про об'єкт фактів системи, дослідження елементів якої дає знання про елементи об'єкта, що досліджується, без взаємодії суб'єкта з об'єктом»[17]. Спільним у різних визначеннях є те, що модель – це певний об'єкт-замінник, дослідження якого є засобом отримання нової інформації про інший досліджуваний об'єкт-оригінал (модельований оригінал).

---

[15] Иванилов Ю. П., Лотов А. В. Математические модели в экономике. М. : Наука, Главная редакция физико-математической литературы, 1979. 304 с.

[16] Тимохин В. М. Теоретико-методологічні засади моделювання економічної динаміки : автореф. дис. на здобуття наук. ступеня доктора економ. наук : спец. 08.00.11 «Математичні методи, моделі та інформаційні технології в економіці». Д., 2007. 36 с.

[17] Фаренік С. А. Логіка і методологія наукового пізнання : наук.-метод. посібник. К. : Вид-во УАДУ, 2000. 340 с.

Однією з основних особливостей моделі є те, що вона не може бути цілком адекватна об'єкту моделювання, та з практичної точки зору це не потрібно. Модель має бути настільки простою, наскільки це дозволяє мета дослідження.

Модель – це абстрактне зображення об'єкта дослідження, що відображає його характеристики, які є істотно важливими для мети дослідження. Ті фактори, що на об'єкт впливають відносно слабо в цікавому для дослідження аспекті, в моделі не враховуються. У зв'язку з цим моделі поділяють на ізоморфні, в яких є місце чіткої відповідності оригіналу, і гомоморфні, які відображають лише певні властивості оригіналу. Варто зауважити, що в сукупності неврахована фактори можуть викликати не тільки значне відхилення в поведінці об'єкта, але й корінним чином змінити характер цієї поведінки. Тому склад врахованих у моделі факторів та її структура у ході удосконалення моделі можуть бути уточнені. Тобто у процесі побудови та дослідження моделей аналітик отримує нове знання, яке може бути використане для уточнення моделі та побудови її нової більш досконалої версії. У цьому розумінні процес моделювання є циклічним.

В економічних дослідженнях ринків праці, де існує можливість кількісного вимірювання величин, широко використовують метод математичного моделювання, який є достатньо ефективним засобом дослідження економічних об'єктів.

Математична модель економічного об'єкта – це її гомоморфне відображення, яке представлене у вигляді сукупності рівнянь, нерівностей, графіків, логічних відношень.

Економіко-математична модель як деяка математична структура, що описує певну економічну систему чи явище, будується аксіоматично. Як аксіоми, виступають принципи або гіпотези. Залежно від того, які гіпотези були покладені автором моделі в її основу, він отримає певну модель. При цьому незначна відмінність в аксіоматиці моделей може призвести до суттєвих відмінностей у висновках, отриманих на основі аналізу цих моделей.

У загальному зазначимо, що модель, зокрема математична, щоб бути ефективним інструментом вивчення процесів в економіці, має відповідати ряду вимог:

– будуватися на основі економічної теорії й відбивати об'єктивні закономірності процесів;

– правильно відтворювати функцію та (чи) структуру реальної економічної системи;

– відповідати певним математичним умовам (мати розв'язок, узгодженні розмірності тощо).

Варто зауважити, що в останні роки в моделюванні економічних об'єктів почали себе проявляти такі негативні тенденції. По-перше, є спроби занадто широкого застосування математичних моделей для економічних досліджень, коли при цьому не здійснюється глибокий якісний аналіз об'єкта або явища, що моделюються. Як результат дослідження, подається модель у вигляді деякої формули або ж математичний розв'язок деякої економічної задачі без пояснення економічної сутності та економічної природи як моделі, так і розв'язку. По-друге, спостерігаються спроби «сліпого» переносу в економічну науку з інших природничих наук методів і моделей, які не можуть узагальнено описувати економічні процеси й об'єкти. Це призводить до порушення одного з основних методологічних принципів застосування математики в економіці – спиратися на фундамент економічної теорії. Отримані таким чином результати у більшості випадків не мають ніякого практичного значення.

Проте математичні моделі як метод пізнання залишається методом, що розвиває і поглиблює знання про закономірності розвитку економічних об'єктів. У цілому зауважимо, що порушення границь застосування математичних методів в економіці призводить до двох полярних явищ:

– математичні викладки занадто абстрактні, тоді їхнє використання не має ніякої цінності з практичної точки зору навіть для ілюстрації;

– з досить простих економічних передумов, які, як правило, не повністю охоплюють досліджуваний об'єкт (навіть його основні прояви), будуються інколи досить складні математичні викладки, які неадекватно відображають ті явища, для дослідження яких вони були застосовані.

У таких випадках математика «пригнічує» змістовний аспект досліджуваної проблеми. Тобто основним недоліком економістів математичної школи є те, що в них математика інколи «відривається» від економіки.

Незважаючи на вищезазначене, використання математичних моделей в економіці дозволяє:

– виділити та формально описати найбільш важливі суттєві зв'язки економічних факторів і об'єктів;

– індуктивним шляхом отримати нові знання про об'єкт: оцінити форму і параметри залежностей, які найбільшою мірою відповідають наявним спостереженням;

– з чітко сформульованих вихідних даних і співвідношень методом дедукції можна отримати висновки, які будуть адекватні досліджуваному об'єкту такою самою мірою, що і зроблені припущення;

– чітко і компактно висвітлити положення економічної теорії, формулювати її поняття і висновки;

– виразити в точній формі не лише результати поточної діяльності окремих економічних об'єктів, а й прогнозувати їх подальший розвиток.

Економіко-математична модель має певну математичну структуру і її змістовну інтерпретацію. Так, різні за природою економічні явища можуть описуватись однаковим математичним виразом, хоча економічна інтерпретація моделі та результати розрахунків будуть різними. Тобто одні і ті самі математичні моделі і методи можуть бути використані для розв'язання абсолютно різних економічних задач. З іншого боку, для однієї і тієї ж економічної задачі можуть будуватись будь-які за формою (математичною структурою) моделі, які використовують різний аналітичний апарат. Тоді постає проблема порівняння цих моделей між собою, що є досить часто складним теоретичним питанням.

Із загальнотеоретичних позицій економіко-математичні моделі повинні відображати тільки суттєві зв'язки, які необхідні для розв'язання економічної задачі, повинні бути наочними, мати добре ілюстровану структуру, надавати нову інформацію про досліджуваний об'єкт чи процес, впливати на спосіб збору інформації.

У зв'язку із вищевикладеним як показники придатності математичних моделей відображати досліджуваний об'єкт чи процес вважається доцільним використовувати показники адекватності,

стійкості, точності, ефективності. У загальному випадку якість економіко-математичних моделей ринку праці становлять такі характеристики, як ефективність і адекватність, які взаємно доповнюють одна одну[18]. Поняття адекватності моделі базується безпосередньо на визначенні об'єкта моделювання. Поняття ефективності ґрунтується на меті моделювання. Узагальнені фактори, що визначають рівень цих характеристик, наведено на рис. 1.12.

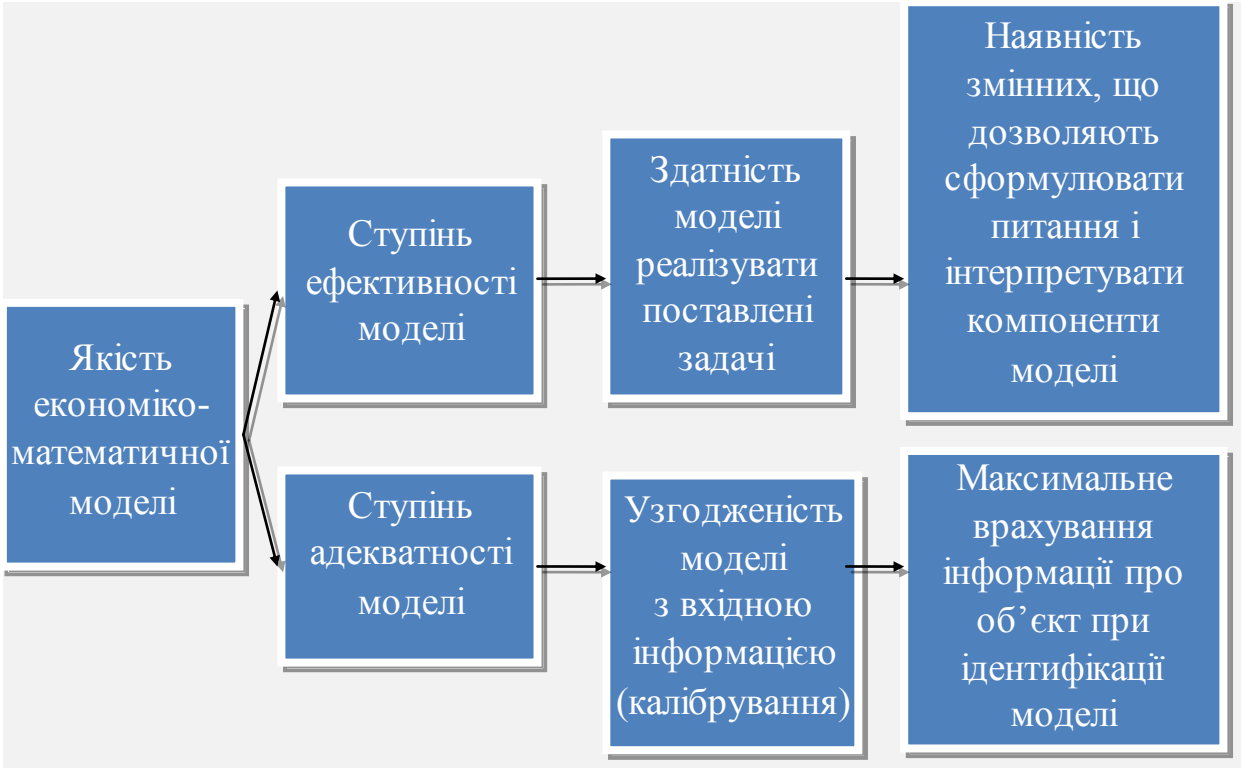


Рис. 1.12. Фактори, що визначають якість економіко-математичних моделей

З точки зору методології застосування математики в економіці, з одного боку, вимагає наявність ґрунтовних знань про економічну дійсність, а з іншого – широке використання математики дозволяє поглибити економічні знання й отримати більш чітко обґрунтовану базу для їхнього подальшого розвитку. При цьому окремі економічні закони не проявляються у «чистому» вигляді, процеси, що

[18] Клейнер Г. Б., Смоляк С. А. Эконометрические зависимости: принципы и методы построения. М. : Наука, 2000. 104 с.
Клейнер Г. Б. Экономико-математическое моделирование как связующее звено в триаде «Экономическая теория – политика – практика». Экономика и математические методы. 2001. Т. 37, № 3. С. 116–126.

вони описують, перебувають під впливом різних об'єктів і явищ. Зокрема, на них впливає і діяльність людини. До того ж, не останнє місце займає спосіб збору статистичної інформації.

У зв'язку з цим ще на початку побудови економіко-математичної моделі постає потреба у глибокому, змістовному обґрунтуванні усіх математичних структур і властивостей з метою, щоб математична за формою модель була б коректною за змістом і всіма своїми елементами та відношеннями між ними, найбільш повно відображала об'єкт моделювання.

Поняття «коректність» розглядається в двох аспектах[19]. У першому коректність розглядається як поняття математичне (з моменту його виникнення). Якщо малі зміни (збурення, похибки) в початкових умовах математичної задачі, у тому числі оптимізаційної, не є причинами різкої зміни результатів розрахунків або ці зміни не є суттєвими, то такі задачі (або моделі) є коректними. Як правило, така коректність пов'язана із вхідною інформацією.

У другому випадку коректність пов'язують з економічною змістовністю задачі. Постановка моделі вважається коректною в економічному сенсі, якщо економічний зміст мають не тільки вона сама і її структурні елементи, але й будь-які її інваріанти з їх структурними елементами. Так, у коректній моделі економічний зміст мають як прямі, так і двоїсті розв'язки.

Процес побудови економіко-математичної моделі має етапи:

– загальне знайомство та початковий аналіз досліджуваного об'єкта;

– визначення понять, що будуть використані;

– формування предмета і мети дослідження (досягнення якої вимагає залучення моделі), введення обмежень і припущень;

– збір і аналіз даних, оцінка їхньої точності;

– виокремлення структурних елементів та вибір факторів, істотних з погляду мети дослідження (специфікація змінних);

– якісний опис взаємозв'язків між елементами моделі;

---

[19] Точилин В. А. Коректность экономико-математических моделей. Киев : Наук. думка, 1989. 176 с.

– введення символічних позначень для відповідних характеристик досліджуваного об’єкта та формалізація, наскільки можливо, взаємозв’язків між ними; таким чином, формалізується (описується мовою математики) математична модель;

– побудова алгоритму розв’язання математичної задачі;

– виконання розрахунків за математичною моделлю;

– перевірка отриманих результатів відповідно до відомих принципів та законів й експериментальних даних (аналіз отриманих результатів);

– визначення сфери застосування і точності отриманого результату; перспектива використання в практичних і теоретичних цілях.

4.2. Принципи побудови моделей

Під час побудови економіко-математичної моделі варто спиратися на основні принципи, на яких базується побудова економіко-математичної моделі будь-якого економічного об’єкта чи процесу.

1. Концептуальна основа моделювання. Вона базується на детальному дослідженні основних складових частин ринку праці в освітній сфері і виявленні елементарних, найбільш суттєвих взаємозв’язків між ними, які впливають на результат. Ці логічні схеми в умовах недосконалої інформації застосовують для дослідження даних. Складається абстрактна схема, елементами якої є уявні об’єкти, що характеризуються елементарними законами і відповідними їхніми перетвореннями.

2. Повнота опису (дослідження). Кожна абстракція (математична структура), яка виникає у системі під час моделювання, повинна бути чітко вираженою і підпорядкованою певним діям. При цьому повнота опису може не враховувати неважливі обставини, які можуть виникати внаслідок певних явищ.

3. Принцип розвитку моделі. Дослідження починається з простих математичних моделей. Далі відбувається поступове їхнє ускладнення. Реалізація цього принципу сприяє аналізу автономних підмоделей (блоків, модулів) частини системи. Така структура, зокрема, дозволяє додавати до моделі блоки, які враховують нові чинники.

4. Ієрархія моделей. Об'єкти моделювання низького рівня є елементами більш вищих рівнів.

5. Системний підхід. Одним з основних етапів дослідження є побудова структурно-функціональної схеми взаємодії економічних факторів. Основою системного підходу є дослідження економічного процесу чи об'єкта як системи, складеної з блоків, деякі властивості яких залежать від характеристики складових блоків. Для побудови економіко-математичної моделі важливо визначити інтенсивність взаємодії складових потоків усередині системи. Для цього визначають параметри коефіцієнтів пропорційності між інтенсивністю потоків законами, що описують керування цією системою, та кількісними характеристиками.

6. Формування математичної моделі. Цей процес повинен базуватися на основі побудованої структурно-функціональної схеми.

7. Інформаційне забезпечення. Після того, як досліджено схему взаємодії економічних структур і на цій підставі сформовано математичну задачу, окреслюються вихідні дані, які необхідні для проведення розрахунків на основі побудованої економіко-математичної моделі. Вихідні дані характеризують початковий стан системи та інтенсивність протікання певних окремих процесів.

Економіко-математичні моделі можна розділити на класи за порядком ознак, які належать до особливостей об'єкта моделювання, мети моделювання і використаного матеріалу[20].

За загальним цільовим призначенням моделі ринку праці поділяються на теоретико-аналітичні та прикладні. Теоретико-аналітичні дозволяють вивчати загальні властивості ринку праці і його характерних елементів шляхом дедукційного аналізу і висновків, що отримані на основі формальних передумов. У теоретичних моделях об'єктом моделювання є сама теорія, тобто деяка конструкція з назв змінних (економічних показників), аксіом, логічних формул, правил. Теоретичні моделі потрібні як засіб організації і використання емпіричних даних. Вони є своєрідним «містком», який пов'язує економічну теорію з прикладними моделями.

---

[20] Панасюк Б. Я. Прогнозування та регулювання розвитку економіки. К. : Поліграфкнига, 1998. 304 с.

Саме тому побудові економіко-математичної моделі, основаної на статистичних даних, повинна передувати теоретико-аналітична модель. На основі аналізу цієї моделі обґрунтовуються: перелік факторів, що враховуються в моделі, і причини неврахування інших, вибір виду зв'язку між факторами, спосіб збору емпіричних даних тощо. Прикладні моделі дають можливість оцінити параметри конкретного економічного об'єкта, сформулювати рекомендації з метою прийняття практичного рішення.

За ступенем агрегування моделі поділяються на макроекономічні і мікроекономічні. Макроекономічні описують ринок праці в цілому. У загальному мікроекономічні моделі описують взаємодію структурних і функціональних складових економіки, а також поведінку окремої складової в ринковому середовищі.

За ознакою врахування фактора часу економіко-математичні моделі ринку праці поділяються на статичні і динамічні. Статичні моделі описують не процес, а стан економічного об'єкта, окремі співвідношення між його параметрами в конкретний момент часу або відрізок часу. Динамічні моделі включають фактор часу, який відіграє роль параметра, що дозволяє описати процес функціонування і розвитку об'єкта, що досліджується.

4.3. Моделювання бізнес-процесів

Побудова моделі бізнес-процесів – складне завдання, яке вимагає певного набору методів і засобів для його розв'язання. Існують різноманітні способи опису бізнес-процесів та подання їх у вигляді схем і діаграм. На сьогодні методології, методи та інструменти моделювання (опису) бізнес-процесів є суттєвим напрямом наукових досліджень. Спектр методів моделювання досить широкий: від найпростіших графічних нотацій, що використовуються для побудови блок-схем та алгоритмів, і таких математичних апаратів, як мережі Петрі, до об'єктно-орієнтованих мов моделювання, наприклад UML, і спеціально розроблених для опису бізнес-систем мов моделювання, наприклад XPDL і BPEL.

У першу чергу, варто зазначити, що поняття «методологія» має два основних значення: по-перше, це – система певних

правил, принципів й операцій, що застосовуються у певній сфері діяльності; по-друге, вчення про цю систему, загальна теорія методу. Структурні компоненти методології[21]:

– підстави методології (філософія, психологія, системний аналіз, наукознавство, етика, естетика);

– характеристики діяльності (особливості, принципи, умови, норми діяльності);

– логічна структура діяльності (предмет, суб'єкт, об'єкт, форми, засоби, методи, результат діяльності);

– часова структура діяльності (фази, стадії, етапи).

Отже, під *методологією* створення моделі бізнесу-процесу розуміють сукупність способів, за допомогою яких бізнес-процеси або окремі операції та зв'язки між ними зображуються у вигляді моделі. Основна задача методології полягає у наданні аналітику практичної послідовності кроків, що приводять до заданого результату. Методологія визначає стандарт і підхід до опису бізнес-процесів організації.

Методологія моделювання складається з чотирьох компонентів:

– теоретичні засади методології моделювання;

– характеристики моделювання (особливості, принципи й умови);

– логічна структура моделювання (предмет, суб'єкт, об'єкт, засоби, методи, нотації, результати тощо);

– часова структура моделювання (фази, стадії, етапи проєкту моделювання тощо).

Методологію опису бізнес-процесів варто розглядати як сукупність нотацій, методів та правил їхнього використання, які застосовуються для моделювання бізнес-процесів та об'єднані загальним підходом.

Метод, у свою чергу, є сукупністю певних правил, прийомів, способів, норм пізнання та дій. Він є системою приписів, принципів, вимог, що орієнтують суб'єкт у вирішенні певної задачі або досягненні певного результату. Метод створення схеми (моделі) бізнес-процесів – найважливіша складова методології опису

[21] Новиков А. М., Новиков Д. А. Методологи. М. : СИНТЕГ, 2007. 668 с.

бізнес-процесів підприємства. Згідно з визначенням будь-який метод є способом досягнення певної мети, розв'язання конкретної задачі. Отже, *метод* – це сукупність практичних та теоретичних прийомів, які дозволяють отримати розв'язок поставленої задачі. Кожний метод є мовою опису об'єктів реального світу за допомогою спеціально розробленого синтаксису, який використовує ряд графічних символів, що відображають реальні об'єкти та зв'язки між ними, а також спосіб опису діяльності підприємства.

Нерозривно з методами пов'язані нотації моделювання бізнес-процесів, які є сукупністю графічних елементів. Тобто нотація – це синтаксис графічної мови моделювання.

Підсумовуючи вищевикладене та ґрунтуючись на проведеному аналізі наукових джерел, вважаємо, що як методології моделювання бізнес-процесів доцільно виділити такі:

– створення моделей структурного типу;

– моделювання потоків робіт;

– моделювання потоків даних;

– симуляція бізнес-процесів.

Серед підходів до побудови та відображення моделей бізнес-процесів виділяють такі моделювання:

– функціональне;

– об'єктно-орієнтоване;

– імітаційне.

За *функціональним підходом* основним елементом є функція (бізнес-функція, дія, операція), система зображується у вигляді ієрархії взаємозалежних функцій. Моделювання бізнес-процесів зводиться до побудови їхньої схеми у вигляді послідовності кроків, а саме виконання бізнес-функцій, з якими пов'язані матеріальні й інформаційні об'єкти, ресурси, що використовуються, та організаційні одиниці. Перевагою такого підходу є наочність послідовності та логіки операцій у бізнес-процесах підприємства, а недоліком – деяка суб'єктивність у деталізації операцій.

За *об'єктно-орієнтованим підходом* система розбивається на множину об'єктів, що тотожні до об'єктів реального світу та взаємодіють між собою за допомогою передачі інформації (певних сигналів). Об'єктами в моделюванні бізнес-процесів є певні предмети або реальні сутності, наприклад клієнт, замовлення,

послуга тощо. Кожен об'єкт характеризується набором атрибутів, значення яких описують його стан, а також набором операцій для перевірки та зміни цього стану. Об'єктно-орієнтований підхід передбачає спочатку виділення об'єктів, а потім визначення тих дій, в яких вони беруть участь. При цьому розрізняють пасивні об'єкти (матеріали, документи, обладнання), над якими виконуються дії, та активні об'єкти (організаційні одиниці, виконавці, програмне забезпечення), які здійснюють дії. Такий підхід дозволяє більш об'єктивно виділити операції над об'єктами і вирішити завдання щодо доцільності використання цих об'єктів. Недолік об'єктно-орієнтованого підходу полягає у меншій наочності.

Розглянуті підходи до моделювання бізнес-процесів можливо об'єднати у комплексний підхід *аналітичного (статичного) моделювання.*

Імітаційне моделювання передбачає побудову моделей бізнес-процесів за допомогою програмних засобів для імітації їхнього виконання у часі. Такий підхід дозволяє імітувати виконання бізнес-процесів з урахуванням графіків робочого часу та наявності необхідної кількості ресурсів, що дозволить проаналізувати особливості виконання бізнес-процесів в умовах невизначеності внутрішнього та зовнішнього середовища, оцінити реальний час їхнього виконання. Імітаційна модель включає у себе структурну модель процесу та сукупність аналітичних моделей, пов'язаних між собою. Математичною основою цих моделей є нелінійні диференційні рівняння, ймовірнісні методи, методи автоматного моделювання тощо.

Розрізняють активні моделі, які передбачають широку взаємодію з користувачами та легкість внесення в них певних змін, і пасивні, тобто такі, що не взаємодіють з користувачами.

До основних цілей, які спонукають до побудови моделей бізнес-процесів, а саме: опис і вивчення, вдосконалення, автоматизація, – варто додати моніторинг і контроль бізнес-процесів, оскільки саме завдяки їм формується зворотний зв'язок у системі управління бізнес-процесами організацій. Також під час вибору певного методу моделювання можуть братися до уваги й інші критерії, такі як складність моделювання, зовнішня «привабливість» моделі тощо.

Тести до розділу 1

1. Бізнес-аналітика – це:

а) інструменти і додатки для пошуку, аналізу, моделювання і доставки інформації, необхідної для прийняття рішень;

б) мультидисциплінарна сфера, що перебуває на межі таких дисциплін, як: інформаційні технології, бази даних, алгоритми інтелектуальної обробки інформації, математична статистика, методи візуалізації даних тощо;

в) дисципліна, яка використовується для виявлення ділових потреб і знаходження рішень ділових проблем;

г) усі відповіді правильні.

2. Одним із завдань бізнес-аналітика є:

а) планування процесу розробки вимог, визначення пріоритетів вимог і управління змінами;

б) оцінка і перевірка правильності рішення;

в) збір і дослідження неструктурованої інформації;

г) збір та аналіз вимог.

3. Стратег – це:

а) ІТ-відділи, знаючи про бізнес-проблеми, звертаються до бізнес-аналітика, щоб краще зрозуміти і визначити вимоги для їхніх ІТ-систем;

б) бізнес-аналітик, який задовольняє потребу організації у вирішенні стратегічних питань, є досвідченим в аналізі стратегічної мети організації та її середовища, консультуючи вище керівництво з питань належної політики організації та ефектів від політичних рішень;

в) бізнес-аналітики, які аналізують цілі, процеси і ресурси, і пропонують шляхи, якими могла бути зроблена модернізація (реорганізація бізнес-процесів) або удосконалення (модернізація бізнес-процесів);

г) правильної відповіді немає.

4. Найбільша професійна організація у сфері бізнес-аналітики:

а) Міжнародний інститут бізнес-аналітики (IIBA);

б) Британське комп'ютерне суспільство (BCS);

в) Інститут аналітиків і програмістів (IAP);

г) Австралійська асоціація бізнес-аналітики (ABAA).

5. BABOK – це:

а) діяльність, яка дозволяє впроваджувати зміни в компанії шляхом визначення потреб, і рекомендації рішень, які забезпечують цінність для зацікавлених осіб;

б) платформа, яка об'єднує всі точки зору, що описують завдання бізнес-аналізу, які дозволяють належним чином проаналізувати зміни або оцінити їхню необхідність;

в) програма, в якій бізнес-аналітики можуть виконувати різнопланову роботу в рамках компанії;

г) інструменти і додатки для пошуку, аналізу, моделювання та доставки інформації, необхідної для прийняття рішень.

6. Предметом бізнес-аналітики є:

а) одна з функцій управління, що зводиться до аналітичного забезпечення даних рішень;

б) діяльність компаній, спрямована на здійснення змін у бізнесі виходячи з вимог стейкхолдерів;

в) одна з функцій управління, яка відображає технологічний етап процесу прийняття управлінських рішень щодо змін у бізнесі під впливом внутрішніх і зовнішніх факторів;

г) одна з функцій управління, яка відображає технологічний етап процесу прийняття управлінських рішень щодо змін у бізнесі під впливом внутрішніх і зовнішніх факторів і зводиться до аналітичного забезпечення цих рішень.

7. Основні вимоги, без врахування яких отриманий результат не можна розглядати як повний:

а) комунікація;

б) нормативність;

в) адаптивність;

г) правильна відповідь б) і в).

8. Основними цілями сучасного управлінського обліку можна визначити:

а) новий стиль управління, управління бюджетом;

б) максимізацію прибутку, виробничі технології;

в) максимізацію прибутку, контроль витрат;

г) контроль витрат, виробничі технології.

9. Основні принципи аналітичної роботи:

а) гнучкість, цілеспрямованість, адаптивність;

б) активність, об'єктивність, адекватність;

в) цілеспрямованість, повнота, безперервність;

г) усі відповіді правильні.

10. Об'єкт наукової дисципліни – це:

а) те, на що направлене вивчення, вибрана для вивчення частина внутрішньої реальності;

б) те, на що направлене вивчення, вибрана для вивчення частина зовнішньої реальності;

в) те, на що направлене дослідження, вибрана для аналізу частина зовнішньої реальності;

г) правильної відповіді немає.

11. Системний аналіз (СА) – це:

а) сукупність методів і засобів, які використовуються під час дослідження складних технічних, економічних, організаційних структур;

б) напрям дослідження і вивчення світу, в основі якого полягає розгляд об'єктів як системи, орієнтований на розкриття цілісності об'єкта, виявлення різноманітності зв'язків у ньому і приведення їх до єдиної теоретичної картини;

в) сукупність методів і засобів, які використовуються під час дослідження та проєктування складних технічних, економічних, організаційних структур;

г) сукупність методів і засобів, які використовуються у процесі проєктування складних технічних, економічних, організаційних структур.

12. Основні принципи системного підходу:

а) взаємозв'язку, ієрархічності;

б) багатоплановості, достовірності;

в) ієрархічності, нормативності;

г) динамічності, адаптивності.

13. Принцип різнопорядковості:

а) система розглядається як складна структура з різними рівнями, між якими встановлюються певні зв'язки;

б) система вивчається як частина певної макросистеми;

в) система розглядається в русі й розвитку;

г) полягає у тому, що різні ієрархічні рівні системи породжують закономірності різного порядку.

14. До етапів СА входить етап формулювання загальної мети та критеріїв системи, що включає в себе:

а) визначення цілей; виділення елементів; обчислення оцінок за критерієм;

б) визначення цілей; визначення критеріїв; декомпозиція критеріїв за підсистемами;

в) формулювання конкретних завдань виходячи з проблеми; декомпозиція критеріїв за підсистемами;

г) визначення критеріїв; виділення підсистем, визначення цілей.

15. Графічно «дерево цілей» повернуте «кроною»:

а) донизу;

б) доверху;

в) ліворуч;

г) праворуч.

16. Важливі властивості, які мають алгоритми:

а) дискретність, визначеність, нескінченність;

б) скінченність, вихідні дані, невизначеність;

в) дискретність, визначеність, нескінченність;

г) скінченність, вхідні дані, ефективність.

17. Метод ділових ігор:

а) полягає в опитуванні групи фахівців з метою з'ясування їхньої думки стосовно досліджуваної проблеми;

б) імітаційне моделювання реальних ситуацій, за якого учасники гри поводять себе так, як би вони діяли в реальній ситуації;

в) одержання та зіставлення анонімних суджень про питання, яке становить для нас інтерес, через послідовне розсилання анкет, що перемежовується з обробленням отриманої інформації;

г) метод посилення творчого підходу стимулюванням генерування ідей у процесі їхнього обговорення групою людей, під час якого забороняється критик.

18. Структуризація – це:

а) метод наближеного подання замкнутих нелінійних систем, у процесі якого дослідження нелінійної системи замінюється аналізом лінійної системи, в деякому розумінні еквівалентної початкової;

б) процес подання інформації про об'єкт, процес, явище у структуризованому вигляді;

в) метод відображення певної сфери у вигляді струкрної системи, коли форма виділяється в особливий предмет дослідження незалежно від змісту;

г) розподіл системи на ієрархічні підсистеми і компоненти та встановлення між ними зв'язків і відносин, що дозволяє здійснювати управління такою системою.

19. За складністю системи поділяють на:

а) відкриті, закриті;

б) живі, неживі;

в) матеріальні, рефлексивні;

г) малі, великі.

20. Система управлінського обліку складається з:

а) обліку управлінської інформації;

б) управлінської звітності;

в) ключових показників;

г) усі відповіді правильні.

21. Послідовність виконання побудови повної бізнес-моделі складається з таких етапів, як:

а) побудова стратегічної моделі; розбудова організаційної та функціональної моделей;

б) побудова фінансової моделі; моделювання бізнес-процесів;

в) побудова інформаційної моделі; побудова стратегічної моделі;

г) повної відповіді немає.

22. Стратегічна модель будується на основі:

а) дерева цілей та стратегічних карт;

б) дерева цілей;

в) стратегічних карт;

г) немає правильної відповіді.

23. Метою моделювання бізнес-процесів є:

а) систематизація інформації про компанію і її бізнес-процеси в графічній або табличній формі;

б) прогнозування доходів підприємства;

в) визначення складу і структури різних документів, облікових реєстрів, звітів тощо;

г) уніфікація збору первісної інформації.

24. До кількості найбільш значущих, що визначають сутність аналітики, належать такі процеси:

а) аналіз цілей управління і формулювання завдання аналітичної роботи;

б) адаптивне управління збором інформації в інтересах вирішення управлінських завдань в умовах мінливої ситуації;

в) аналіз і оцінювання отриманої інформації в контексті цілей управління, виявлення сутності процесів і явищ;

г) побудова моделі предметної сфери досліджень, об’єкта досліджень і середовища його функціонування, перевірка адекватності моделі та її корекція;

д) усі відповіді правильні.

25. Специфічні методологічні підходи, які, в свою чергу, можуть розглядатися як самостійні аналітичні системи (мета-системи):

а) канонічний математичний аналіз; теоретико-імовірнісний аналіз;

б) статистичний аналіз; логічний аналіз;

в) експертний аналіз; лінгвістичний аналіз;

г) концептуальний аналіз;

д) усі відповіді правильні.

26. Реляційний системний параметр – це:

а) набір відносин таких, що будь-які системи перебувають у будь-якому відношенні з цього набору;

б) набір таких властивостей, одним з яких володіє будь-яка система;

в) обидві відповіді правильні.

27. Системний аналіз вивчає такі питання:

а) утворення цілого;

б) побудова цілого;

в) зростання і розвиток цілого;

г) усі відповіді правильні.

28. До аналізу проблеми належить:

а) точне формулювання проблеми;

б) аналіз структури проблеми;

в) зовнішні зв’язки проблеми;

г) усі відповіді правильні.

29. До аналізу структури системи відносять:

а) визначення цілей;

б) визначення критеріїв;

в) декомпозицію критеріїв за підсистемами;

г) немає правильної відповіді.

30. До формулювання загальної мети та критеріїв системи належить:

а) визначення рівнів ієрархії;

б) виділення підсистем;

в) визначення функціональних і структурних зв'язків;

г) правильної відповіді немає.

31. У процесі побудови дерева цілей необхідно:

а) чітко визначити цілі, досягнення яких сприяє вирішенню виявленої проблеми;

б) виявити інформацію про параметри системи та зовнішнього середовища, які необхідно враховувати;

в) визначити сукупність допущень та обмежень, у рамках яких буде розв'язуватися проблема;

г) усі відповіді правильні.

32. Формалізація – це:

а) процес подання інформації про об'єкт, процес, явище у формалізованому вигляді;

б) метод відображення певної сфери у вигляді формальної системи, коли форма виділяється в особливий предмет дослідження незалежно від змісту;

в) обидві відповіді правильні.

33. Лінеаризація – це:

а) процес подання інформації про об'єкт, процес, явище у формалізованому вигляді;

б) метод відображення певної сфери у вигляді формальної системи, коли форма виділяється в особливий предмет дослідження незалежно від змісту;

в) метод наближеного подання замкнутих нелінійних систем;

г) розподіл системи на ієрархічні підсистеми і компоненти та встановлення між ними зв'язків і відносин.

34. Структуризація – це:

а) процес подання інформації про об’єкт, процес, явище у формалізованому вигляді;

б) метод відображення певної сфери у вигляді формальної системи, коли форма виділяється в особливий предмет дослідження незалежно від змісту;

в) метод наближеного подання замкнутих нелінійних систем;

г) розподіл системи на ієрархічні підсистеми і компоненти та встановлення між ними зв’язків і відносин.

35. Ознаками системи є:

а) цілісність; якісна визначеність; відмежованість стосовно середовища;

б) гетерогенність і структурованість; взаємодія частин системи між собою; взаємодія і зв’язок з навколишнім середовищем;

в) наявність інтегральних характеристик; емерджентність;

г) наявність цілей та їхньої сукупності, цілеспрямованість;

д) усі відповіді правильні.

РОЗДІЛ 2
ПІДГОТОВКА ДАНИХ ДО АНАЛІЗУ

§ 5. Консолідація даних

5.1. Системи підтримки прийняття рішень

Використання комп'ютерів привело до завдань, пов'язаних з аналізом накопиченої інформації для вилучення нових (латентних) знань. Тому виникла потреба у створенні сховищ даних і систем підтримки прийняття рішень, заснованих у тому числі і на методах теорії штучного інтелекту. Управління підприємством, банком, різними сферами бізнесу, зокрема електронного, сьогодні неможливе без процесів накопичення, аналізу певної інформації, виявлення певних закономірностей і залежностей, прогнозування тенденцій і ризиків.

У сучасному світі обчислювальні системи і комп'ютерні мережі дозволяють накопичувати великі масиви даних для вирішення завдань обробки й аналізу статистичної інформації. Проте сама по собі машинна форма подання даних містить інформацію в прихованому вигляді, і для її вилучення потрібно використовувати спеціальні методи аналізу даних. Значний обсяг інформації, з одного боку, дозволяє отримати більш точні розрахунки, аналіз, прогнози, висновки, з іншого – перетворює пошук рішень у складну задачу. У результаті з'явився клас програмних систем, покликаних полегшити роботу аналітиків. Такі системи прийнято називати системами підтримки прийняття рішень – СППР. Для виконання аналізу СППР повинна накопичувати інформацію, володіючи засобами її введення і зберігання.

Виділяють три основні завдання СППР:

– введення даних;

– зберігання даних;

– аналіз даних.

Таким чином, СППР – це системи, що володіють засобами введення, зберігання й аналізу даних, які стосуються певної предметної сфери, з метою пошуку оптимальних рішень. Основне

завдання СППР – надати аналітикам інструмент для виконання аналізу даних. Сама по собі система не генерує правильні рішення, а тільки надає аналітику дані у відповідному вигляді (звіти, таблиці, графіки тощо) для вивчення та аналізу. Тому для ефективного використання СППР її користувач-аналітик повинен мати відповідну кваліфікацію.

За ступенем «інтелектуальності» обробки даних під час аналізу виділяють три класи завдань:

– інформаційно-пошуковий – СППР здійснює пошук необхідних даних. Характерною рисою такого аналізу є виконання заздалегідь визначених запитів;

– оперативно-аналітичний – СППР виробляє групування й узагальнення даних у будь-якому вигляді, необхідному аналітику. На відміну від інформаційно-пошукового аналізу в цьому випадку неможливо заздалегідь передбачити необхідні аналітику запити;

– інтелектуальний – СППР здійснює пошук функціональних і логічних закономірностей у накопичених даних, побудову моделей і правил, які пояснюють знайдені закономірності і (або) прогнозують розвиток деяких процесів (з певною ймовірністю).

Цінність і достовірність знань, отриманих у результаті аналізу бізнес-даних, залежить не тільки від ефективності використовуваних аналітичних методів і алгоритмів, але і від того, наскільки правильно підібрані і підготовлені вихідні дані для аналізу. Як правило, аналітикам доводиться мати справу із ситуацією, яка проявляється в такому. По-перше, дані в організації розташовані в різних джерелах найрізноманітніших форматів і типів – в окремих файлах офісних документів (Excel, Word тощо), в облікових системах (1 С: Підприємство, Парус та ін.), у базах даних (Oracle, Access, dBase тощо). По-друге, дані можуть бути надлишковими або, навпаки, недостатніми. По-третє, дані є «брудними», тобто містять фактори, що заважають їх правильній обробці та аналізу (пропуски, аномальні значення, дублікати і протиріччя). Тому перш ніж приступати до безпосередньо аналітичної роботи, необхідно виконати ряд процедур, метою яких є приведення даних до прийнятного рівня якості та інформативності, а також організувати їхнє інтегроване зберігання в структурах, що забезпечують їхню цілісність, несуперечність, високу швидкість і гнучкість виконання аналітичних запитів.

Консолідація даних – комплекс методів і процедур, спрямованих на вилучення даних з різних джерел, забезпечення необхідного рівня їхньої інформативності та якості, перетворення в єдиний формат, в якому вони можуть бути завантажені в аналітичну систему.

Консолідація даних є початковим етапом реалізації будь-якої аналітичної діяльності. В основі консолідації полягає процес збору та організації зберігання даних у вигляді, оптимальному з точки зору їхньої обробки на конкретній аналітичній платформі або вирішення конкретної аналітичної задачі. Супутніми завданнями консолідації є оцінка якості даних і їхнього збагачення.

Основні критерії оптимальності з точки зору консолідації даних:

– забезпечення високої швидкості доступу до даних;

– компактність зберігання;

– автоматична підтримка цілісності структури даних;

– контроль несуперечності даних.

Ключовим поняттям консолідації є джерело даних – об'єкт, що містить структуровані дані, які можуть виявитися корисними для вирішення аналітичної задачі. Необхідно, щоб аналітична платформа могла здійснювати доступ до даних з цього об'єкта безпосередньо або після їхнього перетворення в інший формат. Аналітичні додатки, як правило, не містять розширених і гнучких засобів введення і редагування даних, а працюють з уже сформованими вибірками.

У процесі консолідації даних вирішуються такі завдання:

– вибір джерел даних;

– розробка стратегії консолідації;

– оцінка якості даних;

– збагачення;

– очищення;

– перенесення в сховище даних.

На початковому етапі здійснюється вибір джерел, що містять дані, які можуть мати відношення до розв'язуваної аналітиком задачі, потім визначаються тип джерел і методика організації доступу до них. У зв'язку з цим можна виділити три основні підходи до організації зберігання даних.

1. Дані, що зберігаються в окремих файлах. Таким джерелом може бути будь-який файл, дані в якому організовані у вигляді стовпців і записів. Стовпці мають бути типізовані, тобто містити дані одного типу (тільки текстові або числові). Перевага таких джерел у тому, що вони можуть створюватися і редагуватися за допомогою простих офісних додатків, робота з якими не вимагає від персоналу спеціальної підготовки. Недоліки далеко не завжди оптимальні з точки зору швидкості доступу до них, компактності представлення даних і підтримки їхньої структурної цілісності.

2. Бази даних (далі – БД) різних СУБД, таких як: Oracle, SQL Server, Firebird, dBase, MySQL, Access тощо. Файли БД краще підтримують цілісність структури даних, оскільки тип і властивості їхніх полів жорстко задаються під час побудови таблиць. Однак для створення й адміністрування БД потрібні фахівці з більш високим рівнем підготовки, ніж для роботи з популярними офісними додатками.

3. Спеціалізовані сховища даних (далі – СД), їх структура і функціонування спеціально оптимізуються для роботи з аналітичною платформою. Більшість СД забезпечують високу швидкість обміну даними з аналітичними додатками, автоматично підтримують цілісність і несуперечність даних. Основна перевага СД перед іншими типами джерел даних – наявність семантичного шару, який дає користувачеві можливість оперувати термінами предметної сфери для формування аналітичних запитів до сховища.

У процесі розробки стратегії консолідації даних необхідно враховувати характер розташування джерел даних – локальний, коли вони розміщені на тому самому ПК, що й аналітичний додаток або віддалений. Характер розташування джерел даних може істотно вплинути на якість зібраних даних (втрата фрагментів, неузгодженість у часі їхнього поновлення, суперечливість тощо).

Іншим важливим завданням є оцінка якості даних з точки зору їхньої придатності для обробки за допомогою різних аналітичних алгоритмів і методів. У більшості випадків вихідні дані є «брудними», тобто містять фактори, які не дозволяють їх коректно аналізувати, виявляти приховані структури та закономірності,

встановлювати зв'язки між елементами даних і виконувати інші дії, під час може виникнути необхідність для отримання аналітикного рішення. До таких факторів належать помилки введення, пропуски, аномальні значення, шуми, протиріччя тощо. Тому перед тим, як приступити до аналізу даних, потрібно оцінити їхню якість і відповідність вимогам, що пред'являються аналітичною платформою. Якщо в процесі оцінки якості будуть виявлені фактори, які не дозволяють коректно застосувати до даних певні аналітичні методи, необхідно виконати відповідну очистку даних.

Очищення даних – комплекс методів і процедур, спрямованих на усунення причин, що заважають коректній обробці аномалій, пропусків, дублікатів, протиріч, шумів тощо.

Ще однією операцією, у процесі якої може виникнути необхідність під час консолідації даних, є їхнє збагачення. Збагачення – процес доповнення даних деякою інформацією, що дозволяє підвищити ефективність вирішення аналітичних задач.

Збагачення дозволяє більш ефективно використовувати консолідовані дані в тих випадках, коли дані містять недостатньо інформації для якісного вирішення певної задачі. Збагачення даних дозволяє підвищити їхню інформаційну насиченість і, як наслідок, значимість для аналітичної задачі.

Місце консолідації в загальному процесі аналізу даних може бути представлено у вигляді структурної схеми (рис. 2.1).

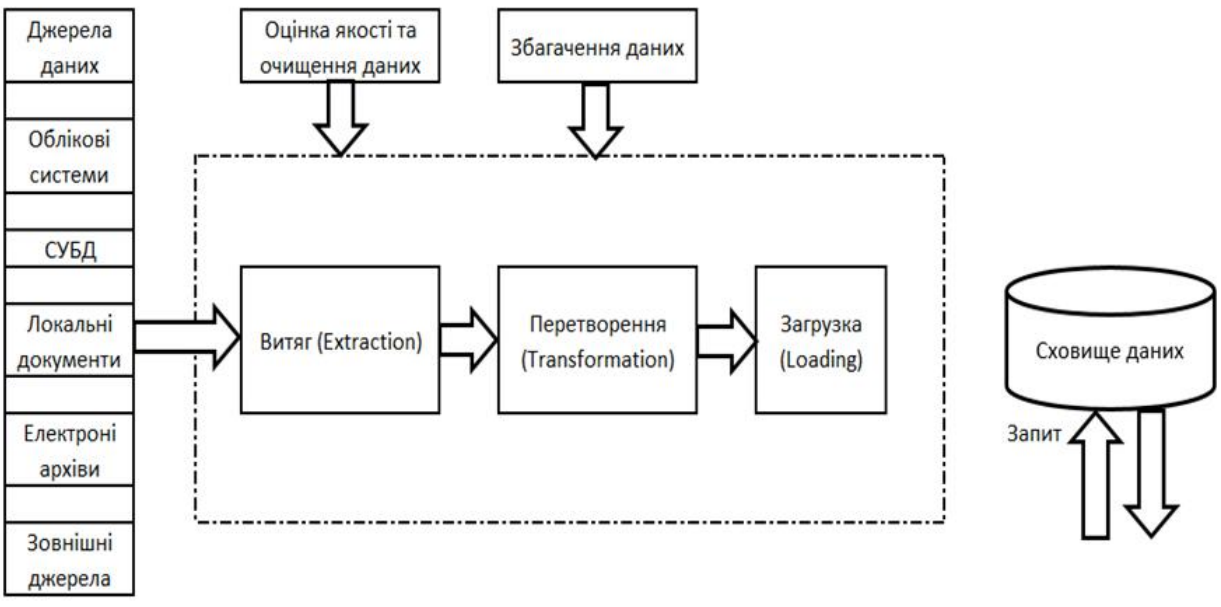


Рис. 2.1. Процес консолідації даних

В основі процедури консолідації даних полягає процес ETL (extraction, transformation, loading). Процес ETL вирішує завдання отримання даних з різних за типом джерел, їх перетворення до вигляду, придатного для зберігання в певній структурі, а також завантаження до відповідної БД або СД. Якщо в аналітика виникають сумніви в якості й інформативності вихідних даних, то за необхідності він може задіяти процедури оцінки їхньої якості, очищення або збагачення, які також є складовими частинами процесу консолідації даних.

5.2. Введення в сховища даних

З появою персональних комп'ютерів почався період бурхливого розвитку інформаційних систем для організації збору і зберігання великих масивів різної ділової та службової інформації, такі системи стали доступними для множини дрібних і середніх фірм, підприємств і організацій. Системи оперативної обробки інформації отримали назву OLTP (On-Line Transaction Processing – оперативна, тобто в режимі реального часу, обробка транзакцій).

Транзакція – деякий набір операцій над базою даних, який розглядають як єдину завершену, з точки зору користувача, дію над деякою інформацією, зазвичай пов'язану зі зверненням до бази даних.

Загальна структура системи OLTP представлена на рис. 2.2.

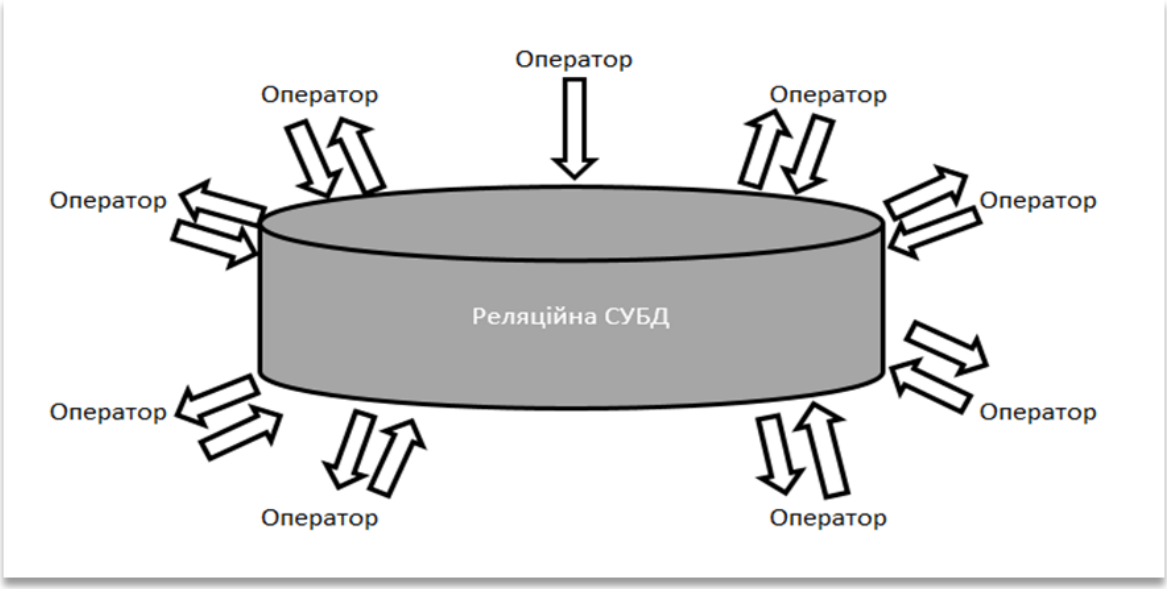


Рис. 2.2. OLTP-система

Типовим прикладом застосування OLTP-систем є масове обслуговування клієнтів (оплата послуг, товарів через Інтернет тощо). Ці ситуації мають дві загальні властивості: досить значна кількість клієнтів і безперервне надходження інформації. Так, для прикладу, під час бронювання авіаквитків з численних пунктів продажу безперервно збігається інформація про вже продані квитки, яку вводять зі своїх робочих місць оператори-продавці. У тій самій БД формується інформація про вільні місця. З точки зору цього завдання транзакція включає в себе набір таких дій, як:

– запит оператора про наявність вільних місць на певний рейс;

– відгук БД з наданням відповідної інформації;

– введення оператором інформації про клієнта, номер замовленого місця і сплачену суму (можливо, буде присутня ще якась службова допоміжна інформація);

– передача нової інформації в базу даних і внесення до неї відповідних змін;

– передача оператору підтвердження про те, що операція пройшла успішно.

Такі транзакції виконуються тисячі разів на день у сотнях пунктів продажів. Очевидно, що основним пріоритетом у цьому випадку є забезпечення мінімального часу відгуку під час максимального завантаження системи.

Характерні риси цього процесу, властиві певною мірою всім OLTP-системам, полягають у такому. Запити і звіти повністю регламентовані. Оператор не може сформувати власний запит, щоб уточнити або проаналізувати будь-яку інформацію. Як тільки обслуговування завершено, інформація про цього клієнта втрачає сенс, стає неактуальною і підлягає видаленню після певного часу.

Операції проводяться над даними з максимальним рівнем деталізації, тобто щодо кожного клієнта окремо. Ситуація змінюється, коли, наприклад, керівництво авіакомпанії ухвалює рішення про вивчення пасажиропотоків з метою їхньої оптимізації. Такі дослідження можуть бути реакцією на інформацію про те, що останнім часом у багатьох пунктах продажів почастішали випадки

нестачі квитків на певні маршрути, що дозволяє зробити припущення про доцільність організації додаткових рейсів. Однак для проведення таких досліджень необхідні як мінімум три речі:

– дані про продажі квитків за досить тривалий період (кілька місяців або років);

– дані не повинні містити протиріч, пропусків, аномальних значень та інших чинників, які не дозволять виконати коректний аналіз;

– необхідна додаткова інформація про бізнес-середовище: про конкурентів, ринкові тенденції, ціни на паливо тощо.

Очевидно, що типова OLTP-система не може це забезпечити. Саме з розумінням цих проблем приходить усвідомлення необхідності використання більш розвинених систем зберігання даних, орієнтованих на аналіз. Зібрана інформація може виявитися досить корисною в процесі управління організацією, пошуку шляхів удосконалення діяльності та отримання за допомогою цих конкурентних переваг. Але для цього потрібні системи, які дозволяли б виконувати не тільки найпростіші дії над даними: підраховувати суми, середні, максимальні та мінімальні значення. З'явилася потреба в інформаційних системах, які дозволяли б проводити глибоку аналітичну обробку, для чого необхідно вирішувати такі завдання, як пошук прихованих структур і закономірностей у масивах даних, висновок з цих правил, яким підпорядковується ця предметна сфера, стратегічне й оперативне планування, формування нерегламентованих запитів, прийняття рішень і прогнозування їхніх наслідків.

Розуміння переваг, які здатні дати інтелектуальний аналіз, призвело до появи нового класу систем – інформаційних систем підтримки прийняття рішень (інформаційних СППР), орієнтованих на аналітичну обробку даних з метою отримання знань, необхідних для розробки рішень у галузі управління.

Узагальнена структурна схема інформаційної СППР представлена на рис. 2.3.

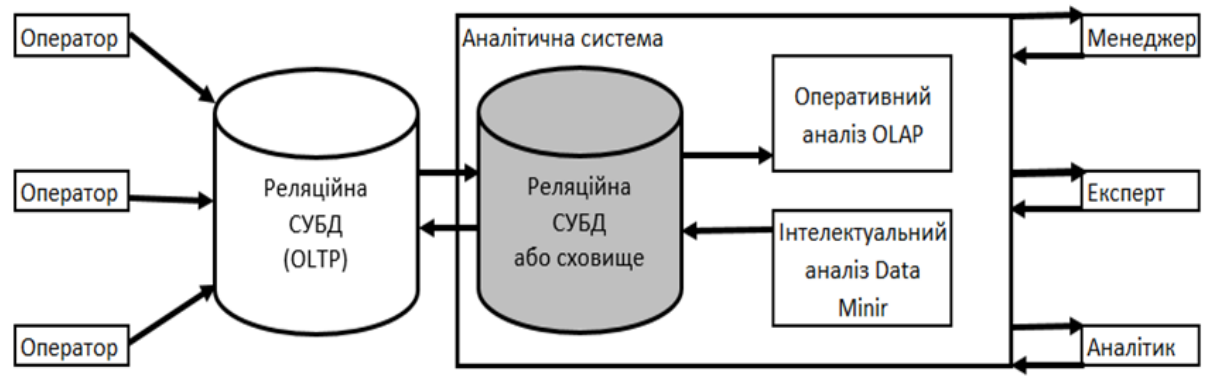


Рис. 2.3. Структура інформаційної СППР

В основі роботи з такою СППР полягають запити, з якими до неї звертається аналітик. При цьому запити, допустимі в традиційних системах оперативної обробки даних, досить примітивні. Водночас аналітична система може відповісти на багато більш складні запити.

У процесі розробки систем аналізу інформації та методології їхнього застосування виявилося, що для ефективного функціонування такі системи повинні бути організовані іншим способом, ніж той, який застосовується в OLTP-системах. Це обумовлено такими причинами:

– для виконання складних аналітичних запитів необхідна обробка великих масивів даних з різноманітних джерел;

– для виконання запитів, пов'язаних з аналізом тенденцій, прогнозуванням протяжних у часі процесів, необхідні історичні дані, накопичені за досить тривалий період, що не забезпечується звичайними OLTP-системами;

– дані, що використовуються для цілей аналізу й обслуговування аналітичних запитів, відрізняються від використовуваних у звичайних OLTP-системах.

У зв'язку з цим виділяють ряд принципових відмінностей між СППР і OLTP-системами (табл. 2.1).

Таблиця 2.1

Відмінності СППР і OLTP-систем

Властивість	OLTP-система	СППР
Цілі використання даних	Швидкий пошук, найпростіші алгоритми обробки	Аналітична обробка з метою пошуку прихованих закономірностей, побудови прогнозів і моделей та ін.
Рівень узагальнення (деталізації) даних	Деталізовані	Як деталізовані, так і узагальнені (агреговані)
Вимоги до якості даних	Можливі некоректні дані (помилки реєстрації, введення тощо)	Помилки в даних не допускаються, оскільки можуть призвести до некоректної роботи аналітичних алгоритмів
Формат зберігання даних	Дані можуть зберігатися в різних форматах залежно від програми, в який вони були створені	Дані зберігаються й обробляються в єдиному форматі
Час зберігання даних	Як правило, не більше року (в межах звітного періоду)	Роки, десятиліття
Зміна даних	Дані можуть додаватися, змінюватися і віддалятися	Допускається тільки поповнення; раніше додані дані змінюватися не повинні, що дозволяє забезпечити їхню хронологію
Періодичність поновлення	Часто, але в невеликих обсягах	Рідко, але в більших обсягах
Доступ до даних	Повинен бути забезпечений доступ до всіх поточних (оперативних) даних	Повинен бути забезпечений доступ до історичних (тобто накопичених за досить тривалий період часу) даних з дотриманням їхньої хронології
Характер виконуваних запитів	Стандартні, налаштовані заздалегідь	Нерегламентовані, що формуються аналітиком «на льоту» залежно від необхідного аналізу
Час виконання запиту	Декілька секунд	До декількох хвилин

Як видно, вимоги до СППР і OLTP-систем істотно відрізняються. Тому в СППР використовуються спеціалізовані бази даних, які називаються сховищами даних (далі – СД). Сховища даних орієнтовані на аналітичну обробку та задовольняють вимогам, що пред'являються до СППР.

Отже:

1. Для спрощення розробки прикладних програм, що використовують БД, створюються системи управління базами даних (далі – СУБД) – програмне забезпечення для управління даними, їх зберігання та безпеки даних.

2. У СУБД розвинений механізм управління транзакціями, що зробило їх основним засобом створення систем оперативної обробки транзакцій (OLTP-систем). До таких систем належать перші СППР, що вирішували задачі інформаційно-пошукового аналізу – ІСР.

3. OLTP-системи не можуть ефективно використовуватися для вирішення завдань оперативно-аналітичного та інтелектуального аналізу інформації. Основна причина полягає в суперечливості вимог до OLTP-систем і СППР.

4. Нині для об'єднання в рамках однієї системи OLTP-підсистем і підсистем аналізу використовується концепція сховищ даних. Загальна ідея полягає у виділенні БД для OLTP-підсистем і БД для виконання аналізу.

В основі концепції СД полягає ідея поділу даних, що використовуються для оперативної обробки і вирішення завдань аналізу. Це дозволяє застосовувати структури даних, які задовольняють вимогам їхнього зберігання з урахуванням використання в OLTP-системах і системах аналізу.

Сховище даних – предметно-орієнтований, інтегрований, незмінний, що підтримує хронологію, набір даних. Він організований з метою підтримки прийняття певних рішень.

Під предметною орієнтованістю в цьому випадку мається на увазі, що СД має розроблятися з урахуванням специфіки конкретної предметної сфери, а не аналітичних додатків, з якими його передбачається використовувати. Структура СД повинна відображати уявлення аналітика про інформацію, з якою йому доводиться працювати.

Інтегрованість означає, що повинна бути забезпечена можливість завантаження в СД інформації з джерел, що підтримують різні формати даних і створених у різних. При цьому дані, що допускають будь-який формат, у процесі завантаження повинні бути перетворені в єдиний. Важливо перевірити завантажувальні дані на цілісність і несуперечливість, забезпечити необхідний їх рівень. Обсяг даних у сховищі повинен бути достатнім для ефективного вирішення аналітичних задач, тому в СД може накопичуватися інформація за кілька років і навіть десятиліть.

Принцип незмінності передбачає, що в СД дані після завантаження не повинні піддаватися будь-яким змінам, за винятком додавання нових даних. Підтримка хронології означає дотримання порядку проходження записів, для чого в структуру СД вводяться атрибути «Дата» і «Час». Якщо фізично впорядкувати записи в хронологічному порядку, то можна зменшити час виконання аналітичних запитів.

5.3. Основні особливості концепції СД

Сьогодні однозначного визначення СД не існує. Узагальнюючи вимоги, що пред'являються до СППР, можна дати таке визначення СД.

Сховище даних – різновид систем зберігання, який орієнтований на підтримку процесу аналізу даних, що забезпечує цілісність, несуперечність і хронологію даних, високу швидкість виконання аналітичних запитів.

Найважливішим елементом СД є семантичний шар – механізм, що дозволяє аналітику оперувати даними за допомогою бізнес-термінів предметної сфери. Семантичний шар дає користувачеві можливість зосередитися на аналізі і не замислюватися про механізми отримання даних.

Типове СД істотно відрізняється від звичайних систем зберігання даних. Основною відмінністю є цілі використання. Інша важлива відмінність полягає в динаміці зміни даних. Бази даних в OLTP-системах характеризуються досить високою динамікою зміни записів через повсякденну роботу значної кількості користувачів. Що стосується СД, то дані з нього не видаляються, а поповнення відбувається відповідно до визначеного регламенту.

Основні вимоги до СД:

– висока швидкість отримання даних зі сховища;

– автоматична підтримка внутрішньої несуперечності даних;

– можливість отримання і порівняння зрізів даних;

– наявність зручних засобів для перегляду даних у сховищі;

– забезпечення цілісності та достовірності даних, що зберігаються.

Щоб усі перераховані вимоги були дотриманими для побудови і роботи СД, використовується не один додаток, а система, в яку входить декілька програмних продуктів. Одні з них являють собою, власне, систему зберігання даних, інші – програми їхнього перегляду, вилучення, завантаження тощо.

Основні концепції сховищ даних

В основі концепції СД полягають такі положення:

– інтеграція й узгодження даних з різних джерел, таких як звичайні системи оперативної обробки, бази даних, облікові системи, офісні документи, електронні архіви, розташовані як усередині підприємства, так і в зовнішньому оточенні;

– поділ наборів даних, що використовуються системами виконання транзакцій і СППР.

Використання концепції СД у СППР та аналізі даних сприяє досягненню таких цілей, як:

– своєчасне забезпечення аналітиків і керівників усією інформацією, необхідною для вироблення обґрунтованих і якісних управлінських рішень;

– створення єдиної моделі представлення даних в організації;

– створення інтегрованого джерела даних, що надає зручний доступ до різних видів інформації та гарантує отримання однакових відповідей на однакові запити з різних аналітичних додатків.

Процес розробки СД досить трудомісткий, деякі організації витрачають на нього декілька місяців і навіть років, а також вкладають значні фінансові кошти. Основними завданнями, які потрібно вирішити в процесі розробки СД, є:

– вибір структури зберігання даних, що забезпечує високу швидкість виконання запитів і мінімізацію обсягу оперативної пам’яті;

– початкове заповнення і подальше поповнення сховища;

– забезпечення єдиної методики роботи з різнорідними даними і створення зручного інтерфейсу користувача.

Коло завдань інтелектуального аналізу даних досить широке, а самі завдання істотно різняться за рівнем складності. Тому залежно від специфіки вирішуваних завдань і рівня їхньої складності архітектура СД і моделі даних, що використовуються для їхньої побудови, можуть відрізнятися. Узагальнена концептуальна схема СД представлена на рис. 2.4.

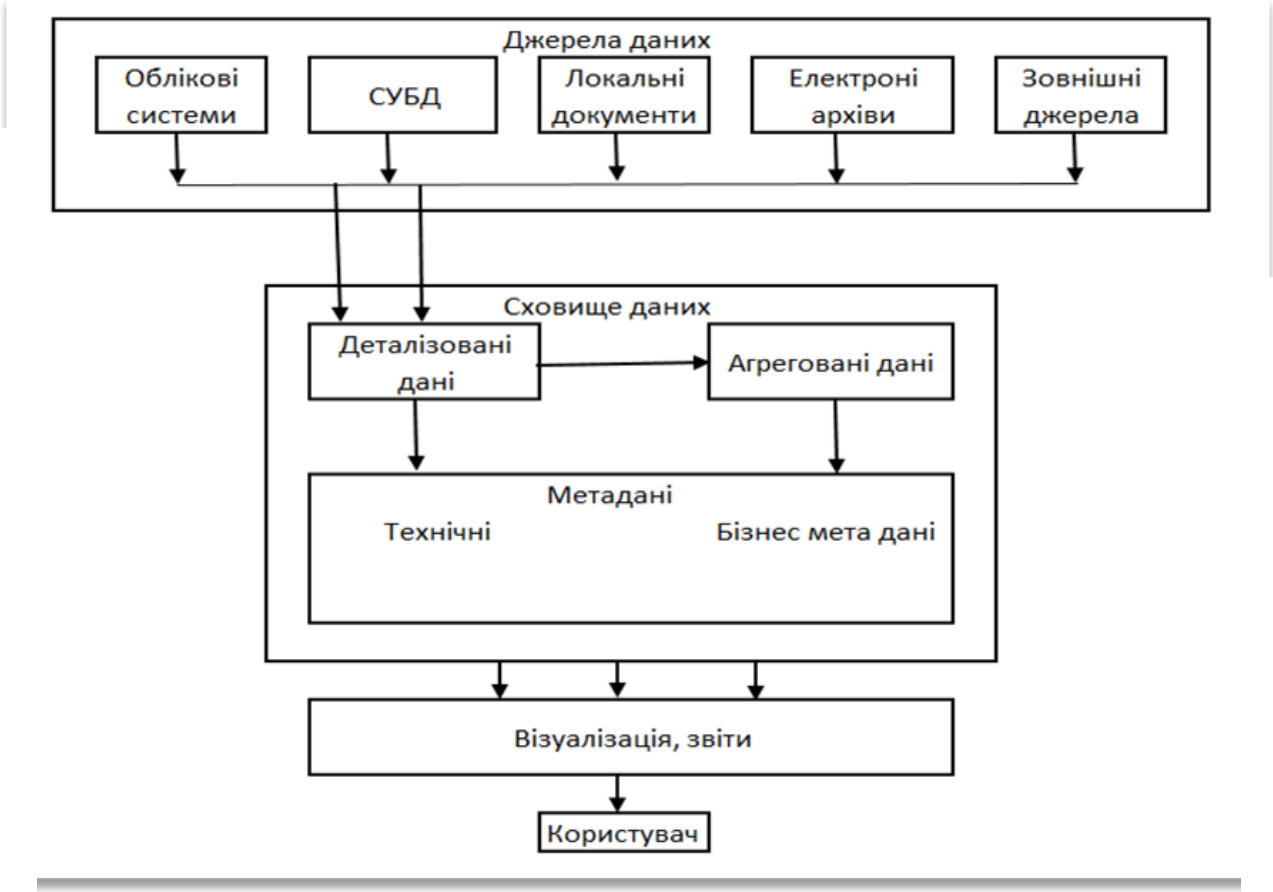


Рис. 2.4. Концептуальна схема СД

Згідно зі схемою дані вилучаються з різних джерел і завантажуються в СД, яке містить як, власне, дані, подані відповідно до деякої моделі, так і метадані.

Дані в СД зберігаються як у деталізованому, так і агрегованому вигляді. Дані в деталізованому вигляді надходять безпосередньо з джерел даних і відповідають елементарним подіям. Це неподільні значення, спроба додаткової деталізації позбавляє їх логічного сенсу.

Багато задач аналізу (наприклад, прогнозування) вимагають використання даних певною мірою узагальнених. Процес узагальнення деталізованих даних називається агрегуванням. Зазвичай агрегуванню піддаються числові дані, вони обчислюються і містяться в СД разом з деталізованими даними. Оскільки один і той самий набір деталізованих даних може породити декілька наборів агрегованих даних з різним ступенем узагальнення, обсяг СД зростає, іноді істотно.

Для опису значення і властивостей інформації з метою кращого її розуміння, використання й управління нею використовують метадані. З точки зору ІТ-технологій метадані – будь-яка інформація, необхідна для аналізу, проєктування, побудови, впровадження та застосування комп'ютерної інформаційної системи. Одне з основних призначень метаданих – підвищення ефективності пошуку. Пошукові запити, які використовують метадані, здійснюють можливим виконання складних операцій з фільтрації та відбору даних.

Якщо розглядати поняття «метадані» в контексті технології СД, то його можна визначити таким чином. Метадані – високорівневі засоби відображення інформаційної моделі та опису структури даних, використовуваної у СД. Метадані повинні містити опис структури даних сховища і структури імпортованих даних з інших джерел. Метадані зберігаються окремо від даних у репозитарії метаданих. Метадані містять усю інформацію, необхідну для отримання, перетворення і завантаження даних з різних джерел, а також для подальшого використання та інтерпретації даних, що містяться в СД.

Виділяють два рівня метаданих – технічний і бізнес-рівень. Технічний рівень містить метадані, необхідні для забезпечення функціонування сховища (статистика завантаження даних і їх використання, опис моделі даних тощо). Бізнес-метадані забезпечують користувачеві можливість концентруватися на процесі аналізу, а не на технічних аспектах роботи зі сховищем. Вони включають бізнес-терміни та визначення, якими звик оперувати користувач. Фактично бізнес-метадані являють собою опис предметної сфери, для роботи в якій створюється аналітична система або СД. Бізнес-метадані описують об'єкти предметної сфери,

інформація про які міститься в СД: атрибути об'єктів і їхні можливі значення, відповідні поля в таблицях тощо. Бізнес-метадані утворюють так званий семантичний шар. Аналітик оперує близькими йому термінами предметної сфери: товар, клієнт продажу, покупки тощо, а семантичний шар транслює бізнес-терміни в низькорівневі запити до даних у сховищі.

За допомогою аналітичного додатка, що використовується спільно з СД, можна формувати запити й отримувати з них дані зі сховища. Дані можуть візуалізуватися безпосередньо або підлягати обробці засобами аналітичного додатка, тоді візуалізуються результати цієї обробки.

Виділяють три основні підходи до використання СД:

– регулярні звіти – підготовка звітів стандартних форм, одержуваних багаторазово з певною періодичністю;

– нерегламентовані запити – можливість отримувати відповіді на нестандартні, сформовані «на вимогу» питання;

– інтелектуальний аналіз даних – підтримка процесу інтелектуального аналізу великих масивів даних з метою виявлення прихованих закономірностей, структур і об'єктів, побудови моделей, прогнозів тощо.

5.4. Огляд архітектур СД

Щоб наблизити СД до умов і специфіки конкретної організації, на сьогодні розроблено декілька архітектур сховищ: реляційні, багатовимірні, гібридні і віртуальні.

Реляційні СД використовують класичну реляційну модель. Дані зберігаються в реляційних таблицях, але утворюють спеціальні структури, адаптовані багатовимірному представленню даних. Така технологія позначається абревіатурою.

Багатовимірні СД реалізують багатовимірне представлення даних на фізичному рівні у вигляді багатовимірних кубів. Ця технологія отримала назву MOLAP.

Гібридні СД поєднують у собі властивості як реляційної, так і багатовимірної моделей даних. У гібридних СД деталізовані дані зберігаються в реляційних таблицях, а агрегати – в багатовимірних кубах. Така технологія побудови СД називається HOLAP.

Віртуальні СД не є сховищами даних у звичному розумінні. У таких системах робота ведеться з окремими джерелами даних, але при цьому емулюється робота звичайного СД. Іншими словами, дані не консолідуються фізично, а збираються безпосередньо в процесі виконання запиту.

Основне призначення OLAP-систем – підтримка аналітичної діяльності, довільних запитів користувачів-аналітиків. Мета OLAP-аналізу – перевірка гіпотез.

Дванадцять вимог, яким повинні задовольняти продукти, що дозволяють виконувати оперативну аналітичну обробку, викладених Коддом:

1. Багатомірність. OLAP-система на концептуальному рівні повинна представляти дані у вигляді багатовимірної моделі, що спрощує процеси аналізу і сприйняття інформації.

2. Прозорість. OLAP-система повинна приховувати від користувача реальну реалізацію багатовимірної моделі, спосіб організації, джерела, засоби обробки і зберігання.

3. Доступність. OLAP-система повинна надавати користувачеві єдину, узгоджену і цілісну модель даних, забезпечуючи доступ до даних незалежно від того, як і де вони зберігаються.

4. Постійна продуктивність під час розробки звітів. Продуктивність OLAP-систем не повинна значно зменшуватися у разі збільшення кількості вимірювань, за якими виконується аналіз.

5. Клієнт-серверна архітектура. OLAP-система повинна бути здатна працювати в середовищі «клієнт-сервер», оскільки більшість даних, які сьогодні потрібно піддавати оперативній аналітичній обробці, зберігаються розподілено. Основною ідеєю тут є те, що серверний компонент інструменту OLAP повинен бути досить інтелектуальним і дозволяти будувати загальну концептуальну схему на основі узагальнення та консолідації різних логічних і фізичних схем корпоративних БД для забезпечення ефекту прозорості.

6. Рівноправність вимірювань. OLAP-система повинна підтримувати багатовимірну модель, в якій усі вимірювання рівноправні. За необхідності додаткові характеристики можуть бути надані окремим вимірам, але така можливість повинна бути у будь-якого вимірювання.

7. Динамічне управління розрідженими матрицями. OLAP-система повинна забезпечувати оптимальну обробку розріджених матриць. Швидкість доступу має зберігатися незалежно від розташування осередків даних і бути постійною величиною для моделей, що мають різні кількість вимірювань і ступінь розрідженості даних.

8. Підтримка розрахованого на багатокористувацький режим. OLAP-система повинна надавати можливість кільком користувачам працювати спільно з однією аналітичною моделлю. При цьому можливі як читання, так і запис даних, тому система має забезпечувати їхню цілісність і безпеку.

9. Необмежені перехресні операції. OLAP-система повинна забезпечувати збереження функціональних відносин, описаних за допомогою певної формальної мови між осередками гіперкуба під час виконання будь-яких операцій зрізу, обертання, консолідації або деталізації. Система повинна самостійно (автоматично) виконувати перетворення встановлених відносин, не вимагаючи від користувача їхнього перевизначення.

10. Інтуїтивна маніпуляція даними. OLAP-система має надавати спосіб виконання операцій зрізу, обертання, консолідації та деталізації над гіперкубом без необхідності користувачеві здійснювати безліч дій з інтерфейсом. Вимірювання, визначені в аналітичній моделі, повинні містити всю необхідну інформацію для виконання вищевказаних операцій.

11. Гнучкі можливості отримання звітів. OLAP-система має підтримувати різні способи візуалізації даних. Рядки, стовпці або сторінки повинні показувати одночасно від 0 до N вимірювань, де N – кількість вимірювань всієї аналітичної моделі.

12. Необмежена розмірність і кількість рівнів агрегації. Дослідження про можливу кількість необхідних вимірювань, потрібних в аналітичній моделі, показало, що одночасно можуть використовуватися до 19 вимірювань. Звідси випливає нагальна рекомендація, щоб аналітичний інструмент міг одночасно надати хоча б 15, а краще – і 20 вимірювань. Більш того, кожне із загальних вимірювань не повинно бути обмежене за кількістю визначених користувачем-аналітиком рівнів агрегації і шляхів консолідації.

1995 р. Кодд до наведеного переліку додав такі шість правил:

1. Пакетний. Вилучення проти інтерпретації. OLAP-система повинна рівною мірою ефективно забезпечувати доступ як до власних, так і до зовнішніх даних.

2. Підтримка всіх моделей OLAP-аналізу. OLAP-система має підтримувати чотири моделі аналізу даних.

3. Обробка ненормалізованих даних. OLAP-система повинна бути інтегрована з ненормалізованими джерелами даних. Модифікації даних, виконані в середовищі OLAP, не повинні приводити до змін даних, що зберігаються у вихідних зовнішніх системах.

4. Збереження результатів OLAP: зберігання їх окремо від вихідних даних. OLAP-система, що працює в режимі читання-запису, після модифікації вихідних даних має зберігати результати окремо. Іншими словами, повинна забезпечуватися безпека вихідних даних.

5. Виключення відсутніх значень. OLAP-система, представляючи дані користувачеві, має відкидати всі відсутні значення. Іншими словами, відсутні значення повинні відрізнятися від нульових значень.

6. Обробка відсутніх значень. OLAP-система має ігнорувати всі відсутні значення без урахування їхнього джерела. Ця особливість пов’язана з 17-м правилом.

Крім того, Е. Кодд розбив усі 18 правил на такі чотири групи, назвавши їх особливостями. Ці групи отримали назви B, S, R і D.

Основні особливості (B) включають такі правила:

– багатовимірне концептуальне уявлення даних (правило 1);

– інтуїтивне маніпулювання даними (правило 10);

– доступність (правило 3);

– пакетне вилучення проти інтерпретації (правило 13);

– підтримка всіх моделей OLAP-аналізу (правило 14);

– архітектура «клієнт-сервер» (правило 5);

– прозорість (правило 2);

– розрахована на багато користувачів підтримка (правило 8).

Спеціальні особливості (S):

– обробка ненормалізованих даних (правило 15);

– збереження результатів OLAP: зберігання їх окремо від вихідних даних (правило 16);

– виняток відсутніх значень (правило 17);

– обробка відсутніх значень (правило 18).

Особливості подання звітів (R):

– гнучкість формування звітів (правило 11);

– постійна продуктивність звітів (правило 4);

– автоматична настройка фізичного рівня (змінене оригінальне правило 7).

Управління вимірами (D):

– універсальність вимірів (правило 6);

– необмежена кількість вимірювань і рівнів агрегації (правило 12);

– необмежені операції між розмірностями (правило 9).

5.5. Тест FASMI

Крім розглянутих раніше особливостей, існує також тест FASMI (Fast of Analysis Shared Multidimensional Information), створений на основі аналізу правил Кодда. У цьому контексті акцент зроблений на швидкість обробки, розрахований на багато користувачів доступ, релевантність інформації, наявність засобів статистичного аналізу і багатовимірність, тобто на представлення аналізованих фактів як функцій від значної кількості їхніх характеризуючих параметрів. Таким чином, OLAP визначили такими п'ятьма ключовими словами: FAST (Швидкий), ANALYSIS (Аналіз), SHARED (Розділяти), MULTIDIMENSIONAL (Багатовимірність), INFORMATION (Інформація).

FAST (Швидкий). OLAP-система повинна забезпечувати видачу більшості відповідей користувачам у межах приблизно 5 секунд. Найпростіші запити обробляються упродовж 1 секунди і далеко не всі – більше 20 секунд. Нещодавнє дослідження в Нідерландах показало, що останні користувачі сприймають процес невдалим, якщо результати не отримані після закінчення 30 секунд. Навіть якщо система попередить, що процес триватиме значно довше, користувачі можуть відволіктися і втратити думку, при цьому якість аналізу страждає. Такої швидкості нелегко досягти з

великою кількістю даних, особливо якщо потрібні спеціальні обчислення «на льоту». Для досягнення цієї мети використовуються різні методи, включно із застосуванням апаратних платформ з більшою продуктивністю.

ANALYSIS (Аналіз). OLAP-система повинна справлятися з будь-яким логічним і статистичним аналізом, характерним для цього додатка, і забезпечувати його збереження у вигляді, доступному для останнього користувача. Природно, система повинна дозволяти користувачеві робити нові спеціальні обчислення як частину аналізу і формувати звіти будь-яким бажаним способом без необхідності програмування. Усі необхідні функціональні можливості аналізу повинні забезпечуватися зрозумілим для останніх користувачів способом.

SHARED (Розділяти). OLAP-система повинна виконувати всі вимоги захисту конфіденційності. Якщо для запису є необхідний множинний доступ, забезпечується блокування модифікацій на відповідному рівні. Обробка множинних модифікацій повинна виконуватися своєчасно і безпечним способом.

MULTIDIMENSIONAL (Багатовимірні). OLAP-система має забезпечити багатовимірне концептуальне представлення даних, включаючи повну підтримку для ієрархій і множинних ієрархій, що забезпечують найбільш логічний спосіб аналізу. Ця вимога не встановлює мінімальну кількість вимірювань, які повинні бути оброблені, оскільки цей показник залежить від програми. Воно також не визначає використовувану технологію БД, якщо користувач дійсно отримує багатовимірне концептуальне представлення інформації.

INFORMATION (Інформація). OLAP-система повинна забезпечувати отримання необхідної інформації в умовах реального часу. Потужність різних систем вимірюється не обсягом інформації, що зберігається, а кількістю вхідних даних, які можуть бути оброблені.

5.6. Багатовимірні сховища даних

Основне призначення багатовимірних сховищ даних (далі – МСД) – підтримка систем, орієнтованих на аналітичну обробку даних, оскільки такі сховища краще справляються з виконанням складних нерегламентованих запитів.

Багатовимірна модель даних, що полягає в основі побудови багатовимірних сховищ даних, спирається на концепцію багатовимірних кубів, або гіперкубів. Вони являють собою впорядковані багатовимірні масиви, які також часто називають OLAP-кубами (On-Line Analytical Processing – оперативна аналітична обробка).

Сутність багатовимірного представлення даних полягає в такому. Більшість реальних бізнес-процесів описується безліччю показників, властивостей, атрибутів тощо. Наприклад, для опису процесу продажів може виникнути необхідність відомості про найменування товарів або їхніх груп, постачальника і покупця, місто, де проводилися продажі, а також про ціни, кількості проданих товарів і загальні суми. Крім того, для відстеження процесу в часі повинен бути введений у розгляд такий атрибут, як дата. Якщо зібрати всю цю інформацію в таблицю, то вона виявиться складною для візуального аналізу й осмислення. Більш того, вона може виявитися надмірною, якщо, наприклад, один і той самий товар продавався одного і того самого часу у різних містах, тому доведеться кілька разів повторити одну й ту саму відповідність «місто – товар» із зазначенням різних сум і кількості. Зазначені проблеми виникають з однієї простої причини: в плоскій таблиці зберігаються багатовимірні дані.

Можна використати інший метод – виконати декомпозицію інформації і перейти до реляційної моделі, яку використовують класичні бази даних. Однак реляційна модель не є оптимальною з точки зору завдань аналізу, оскільки передбачає високу ступінь нормалізації, в результаті чого знижується швидкість виконання запитів. Тому розробка багатовимірної моделі представлення даних, яка реалізується за допомогою багатовимірних кубів, стала природним кроком.

В основі багатовимірного представлення даних полягає їх поділ на дві групи – вимірювання і факти.

Вимірювання – це категоріальні атрибути, найменування і властивості об’єктів, що беруть участь у деякому бізнес-процесі. Значеннями вимірювань є найменування товарів, назви фірм-постачальників і покупців, ПІП людей, назви міст тощо. Вимірювання можуть бути і числовими, якщо будь-якій категорії відповідає числовий код, але в будь-якому випадку це дані дискретні, тобто приймають значення з обмеженого набору. Вимірювання якісно описують досліджуваний бізнес-процес.

Факти – це дані, які кількісно описують бізнес-процес, безперервні за своїм характером, тобто вони можуть приймати безліч значень (ціна товару, їхня кількість, сума продажів або закупівель, зарплата співробітників тощо).

Багатовимірний куб можна розглядати як систему координат, осями якої є вимірювання. За осями будуть відкладатися значення вимірювань: дати, найменування товарів, назви фірм тощо.

У такій системі кожному набору значень вимірів відповідатиме осередок, в якому можна розмістити числові показники, пов’язані з цим набором. Таким чином, між об’єктами бізнес-процесу і їхніми числовими характеристиками буде встановлений однозначний зв’язок.

Принцип організації багатовимірного куба показано на рис. 2.5.

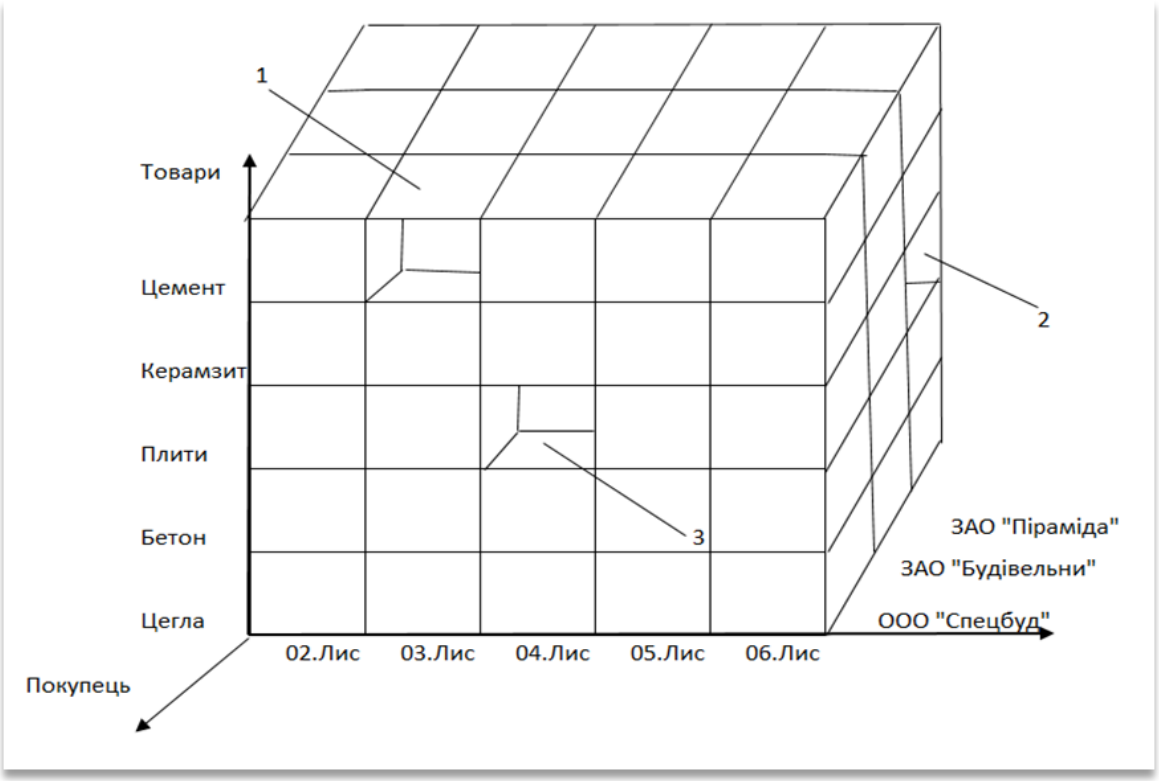


Рис. 2.5. Принцип організації багатовимірного куба

Багатовимірний погляд на вимірювання представлений на рис. 2.6. Фактами в цьому випадку можуть бути Ціна, Кількість, Сума. Тоді виділений сегмент міститме інформацію про те, скільки плит, на яку суму і за якою ціною придбала фірма ЗАТ «Будівельник» 3 листопада.

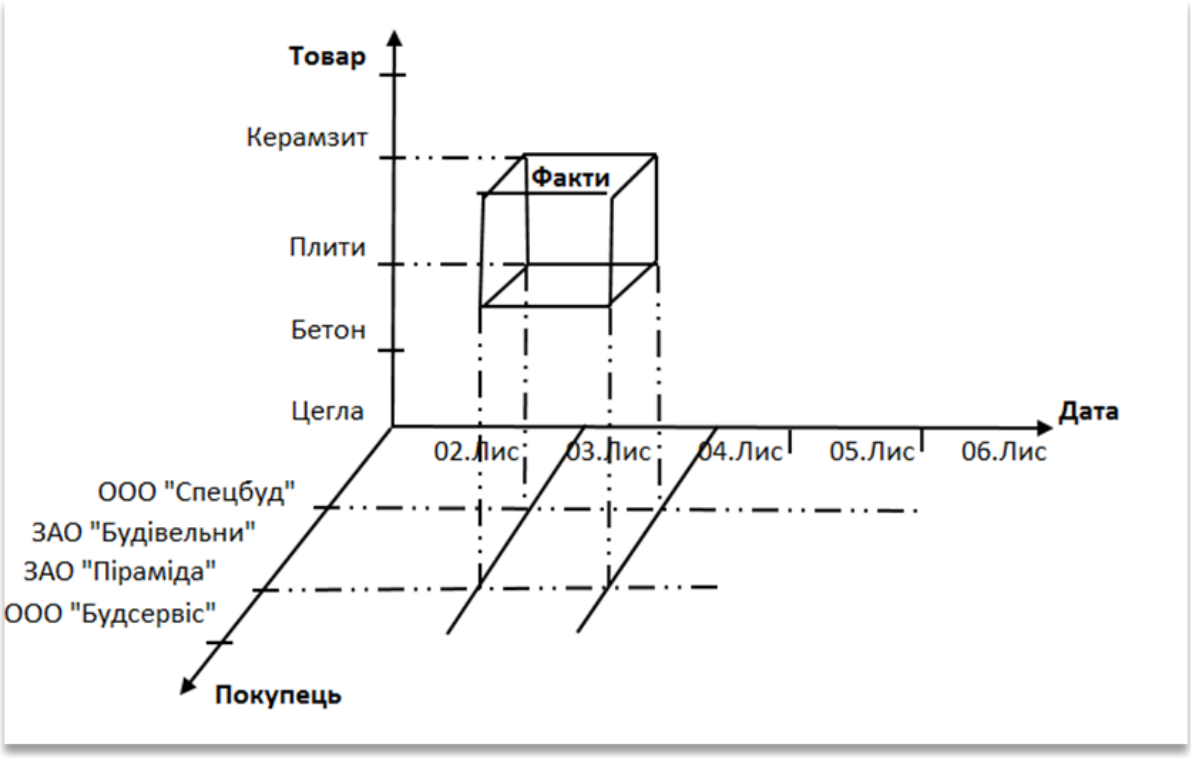


Рис. 2.6. Вимірювання і факти в багатовимірному кубі

Отже, інформація в багатовимірному сховищі даних є логічно цілісною. Це вже не просто набори строкових і числових значень, які в разі реляційної моделі потрібно отримувати з різних таблиць, а цілісні структури типу «кому, що і в якій кількості було продано на цей момент часу».

Представлення даних у вигляді багатовимірних кубів наочніші, ніж сукупність нормалізованих таблиць реляційної моделі, структуру якої представляє тільки адміністратор БД.

Можливості побудови аналітичних запитів до системи, що використовує БСД, більш широкі.

У деяких випадках використання багатовимірної моделі дозволяє значно зменшити тривалість пошуку в БСД, забезпечуючи виконання аналітичних запитів практично в режимі реального часу. Це пов’язано з тим, що агреговані дані обчислюються попередньо і зберігаються в багатовимірних кубах разом з деталізованими, тому

витрачати час на обчислення агрегатів під час виконання запиту вже не потрібно.

OLAP-куб може бути реалізований і за допомогою звичайної реляційної моделі. У цьому випадку має місце емуляція багатовимірного представлення сукупністю плоских таблиць. Такі системи отримали назву ROLAP – Relational OLAP.

Застосування систем зберігання, в основі яких полягає багатовимірне представлення даних, доцільно тільки в тих випадках, коли обсяг використовуваних даних порівняно невеликий, а сама багатовимірна модель має стабільний набір вимірювань.

У процесі пошуку і вилучення з гіперкуба потрібної інформації над його вимірами проводиться ряд дій, найбільш типовими з яких є:

– перетин (зріз);

– транспонування;

– згортка;

– деталізація.

Перетин полягає у виділенні підмножини комірок гіперкуба під час фіксованого значення одного або декількох вимірювань. У результаті розтину виходить зріз або кілька зрізів, кожен з яких містить інформацію, пов’язану зі значенням виміру, за яким він був побудований. Маніпулюючи перетинами гіперкуба, користувач завжди може отримати інформацію в потрібному розрізі. Потім на основі побудованих зрізів може бути сформована крос-таблиця і за її допомогою досить швидко отриманий необхідний звіт.

На рис. 2.7 схематично представлені перетини гіперкуба.

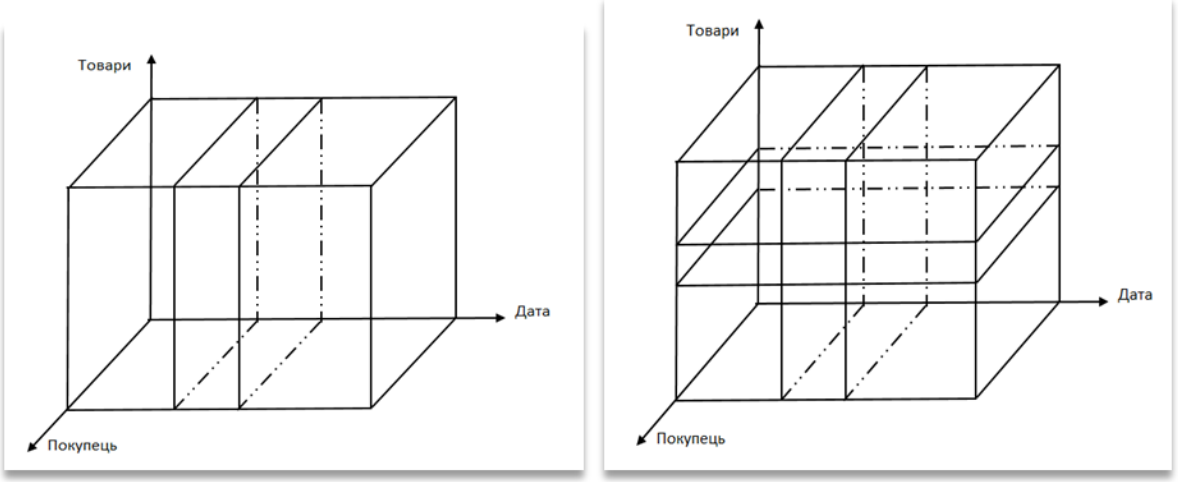


Рис. 2.7. Перетини гіперкуба

Операції згортки (групування) і деталізації (декомпозиції) можливі тільки тоді, коли має місце ієрархічна підпорядкованість значень вимірів. Під час згортки одне або кілька підлеглих значень вимірів замінюються тими значеннями, яким вони підпорядковані. При цьому рівень узагальнення даних зменшується. У табл. 2.2 представлена вихідна таблиця, а в табл. 2.3 – результат її згортки з вимірювання «Товар».

Таблиця 2.2

Вихідна таблиця

Група	Товар	Сума
Будматеріали	Цегла	22 000
	Цемент	12 000
	Керамзит	4 500
	Дошка	7 400
Інструменти	Викрутка	1 200
	Електропилка	7 600
	Точило	2 450
	Шпатель	780

Таблиця 2.3

Результат згортки вихідної таблиці щодо вимірювання «Товар»

Група	Сума
Будівельні матеріали	45 900
Інструмент	12 030

Деталізація – це процедура, зворотна згортці. Рівень узагальнення даних зменшується. При цьому значення вимірювань вищого ієрархічного рівня замінюються одним або декількома значеннями нижчого рівня, тобто замість найменувань груп товарів відображаються найменування окремих товарів.

5.7. Реляційні сховища даних

Реляційна база даних – сукупність відносин, що містять усю інформацію, яка повинна зберігатися в базі. Фізично це виражається в тому, що інформація зберігається у вигляді двовимірних таблиць, пов'язаних за допомогою ключових полів. Застосування реляційної моделі під час створення СД у ряді випадків дозволяє отримати переваги над багатовимірною технологією, особливо в частині ефективності роботи з великими масивами даних і використання пам'яті комп'ютера.

В основі технології РСД полягає принцип, відповідно до якого вимірювання зберігаються в плоских таблицях так само, як і звичайних реляційних СУБД, а факти – в окремих спеціальних таблицях цієї ж бази даних. При цьому таблиця фактів є основою для пов'язаних з нею таблиць вимірів. Вона містить кількісні характеристики об'єктів і подій, сукупність яких передбачається в подальшому аналізувати.

На логічному рівні розрізняють дві схеми побудови РСД – «зірка» та «сніжинка».

У процесі використання схеми «зірка» центральною є таблиця фактів, з якою пов'язані всі таблиці вимірювань. Таким чином, інформація про кожен вимір розташовується в окремій таблиці (рис. 2.8).

Однак розміщення всієї інформації про вимірювання в одній таблиці виявляється не завжди виправданим. Наприклад, якщо товари, що продаються об'єднані в групи, то доведеться в певний спосіб показати, до якої групи належить кожен товар, що призведе до багаторазового повторення назв груп. Це не тільки викличе зростання надмірності, але і підвищить імовірність виникнення протиріч.

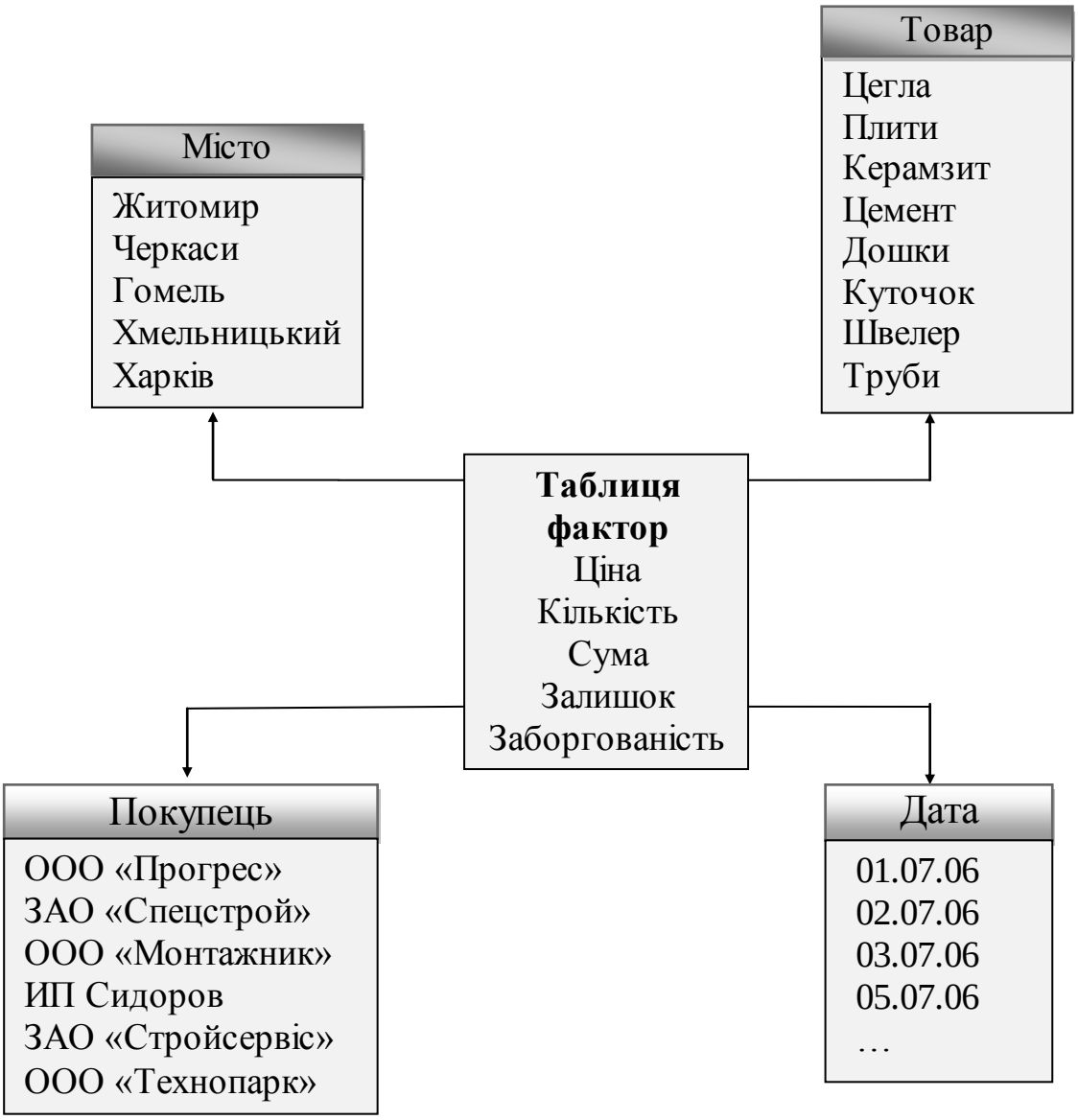


Рис. 2.8. Схема побудови РСД «зірка»

Для більш ефективної роботи з ієрархічними вимірами була розроблена модифікація схеми «зірка», яка отримала назву «сніжинка». Основною особливістю схеми «сніжинка» є те, що інформація про один вимір може зберігатися в декількох пов'язаних таблицях. Тобто якщо хоча б одна з таблиць вимірів має одну або декілька пов'язаних з нею інших таблиць вимірів, у цьому випадку застосовуватиметься схема «сніжинка» (рис. 2.9).

Основна функціональна відмінність схеми «сніжинка» від схеми «зірка» – це можливість роботи з ієрархічними рівнями, що визначають ступінь деталізації даних. У наведеному прикладі схема «сніжинка» дозволяє працювати з даними на рівні максимальної

деталізації, наприклад з кожним товаром окремо, або використовувати узагальнене уявлення за групами товарів з відповідною агрегацією фактів.

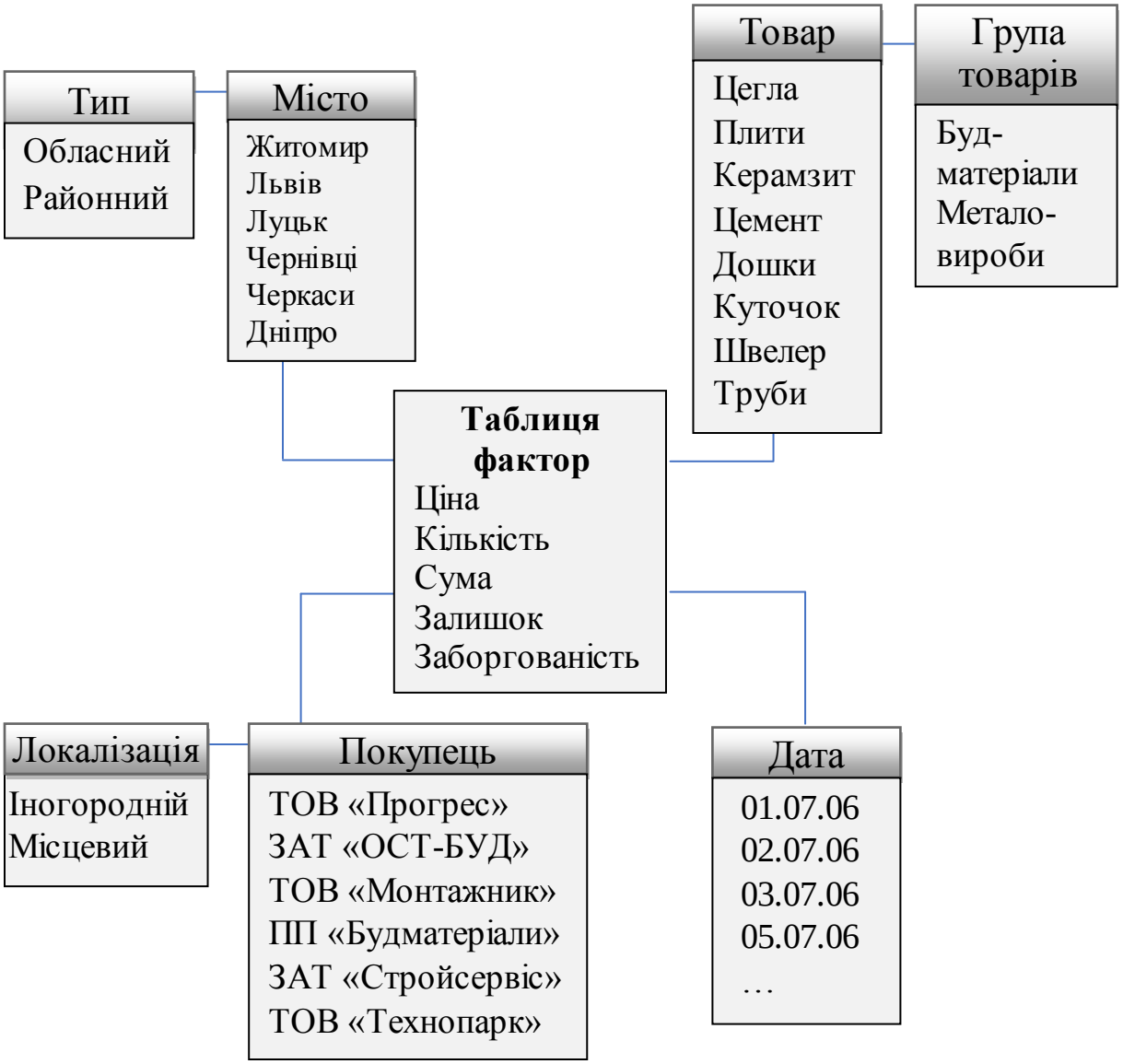


Рис. 2.9. Схема побудови РСД «сніжинка»

Вибір схеми для побудови РСД залежить від використовуваних механізмів збору та обробки даних. Кожна зі схем має свої переваги і недоліки, які, однак, можуть проявлятися більшою чи меншою мірою залежно від особливостей функціонування СД у цілому.

До переваг схеми «зірка» можна віднести:

– простоту і логічну прозорість моделі;

– простішу процедуру поповнення вимірювань, оскільки доводиться працювати тільки з однією таблицею.

Недоліками схеми «зірка» є:

– повільна обробка вимірювань, оскільки одні й ті самі значення вимірювань можуть зустрічатися декілька разів в одній і тій самій таблиці;

– висока ймовірність виникнення невідповідностей у даних (зокрема, суперечностей), наприклад через помилки введення.

Перевагами схеми «сніжинка» є:

– вона ближче до представлення даних у багатовимірній моделі;

– процедура завантаження з РСД у багатовимірній структурі більш ефективна і проста, оскільки завантаження проводиться з окремих таблиць;

– набагато нижча ймовірність появи помилок, невідповідності даних;

– велика, порівняно зі схемою «зірка», компактність представлення даних, оскільки всі значення вимірювань згадуються тільки один раз.

Недоліки схеми «сніжинка»:

– досить складна для реалізації і розуміння структура даних;

– ускладнена процедура додавання значень вимірів.

Крім того, існує ряд технічних особливостей, які можуть визначити вибір розробників РСД під час вибору схеми його побудови.

Основні переваги РСД такі:

– практично необмежений обсяг збережених даних;

– оскільки реляційні СУБД полягають в основі побудови багатьох систем оперативної обробки (OLTP), які зазвичай є основними джерелами даних для СД, використання реляційної моделі дозволяє спростити процедуру завантаження й інтеграції даних у сховищі;

– у разі додавання нових вимірів даних немає необхідності виконувати складну фізичну реорганізацію сховища на відміну, наприклад, від багатовимірних СД;

– забезпечуються високий рівень захисту даних і широкі можливості розмежування прав доступу.

Основний недолік РСД полягає в тому, що під час використання високого рівня узагальнення даних та ієрархічності вимірювань у таких сховищах починають «розмножуватися» таблиці агрегатів. У результаті швидкість виконання запитів реляційних сховищ сповільнюється.

Водночас у багатовимірних сховищах, де дані зберігаються у вигляді багатовимірних кубів, ця проблема практично не виникає і в більшості випадків вдається досягти більш високої швидкості виконання запитів.

Таким чином, вибір реляційної моделі у процесі побудови СД доцільний у таких випадках:

1. Значним є обсяг збережених даних (багатовимірні СД стають неефективними).

2. Ієрархія вимірювань нескладна (небагато агрегованих даних).

3. Потрібна часта зміна розмірності даних. Під час використання реляційної моделі можна обмежитися додаванням нових таблиць, а для багатовимірної моделі доведеться виконувати складну перебудову фізичної структури сховища.

Логічно було б використовувати таку модель СД, яка являла б собою комбінацію реляційної і багатовимірної моделей і дозволяла б поєднувати високу продуктивність, характерну для багатовимірної моделі, і можливість зберігати скільки завгодно великі масиви даних, властиві реляційної моделі. Така модель, що поєднує в собі принципи реляційної і багатовимірної моделей, отримала назву гібридної, або HOLAP (Hybrid OLAP).

Сховища даних, побудовані на основі HOLAP, називаються гібридними сховищами даних (далі – ГСД) (рис. 2.10).

Основним принципом побудови ГСД є те, що деталізовані дані зберігаються в реляційній структурі (ROLAP), яка дозволяє зберігати великі обсяги даних, а агреговані – в багатовимірній (MOLAP), яка дозволяє збільшити швидкість виконання запитів (оскільки під час виконання аналітичних запитів уже не потрібно обчислювати агрегати).

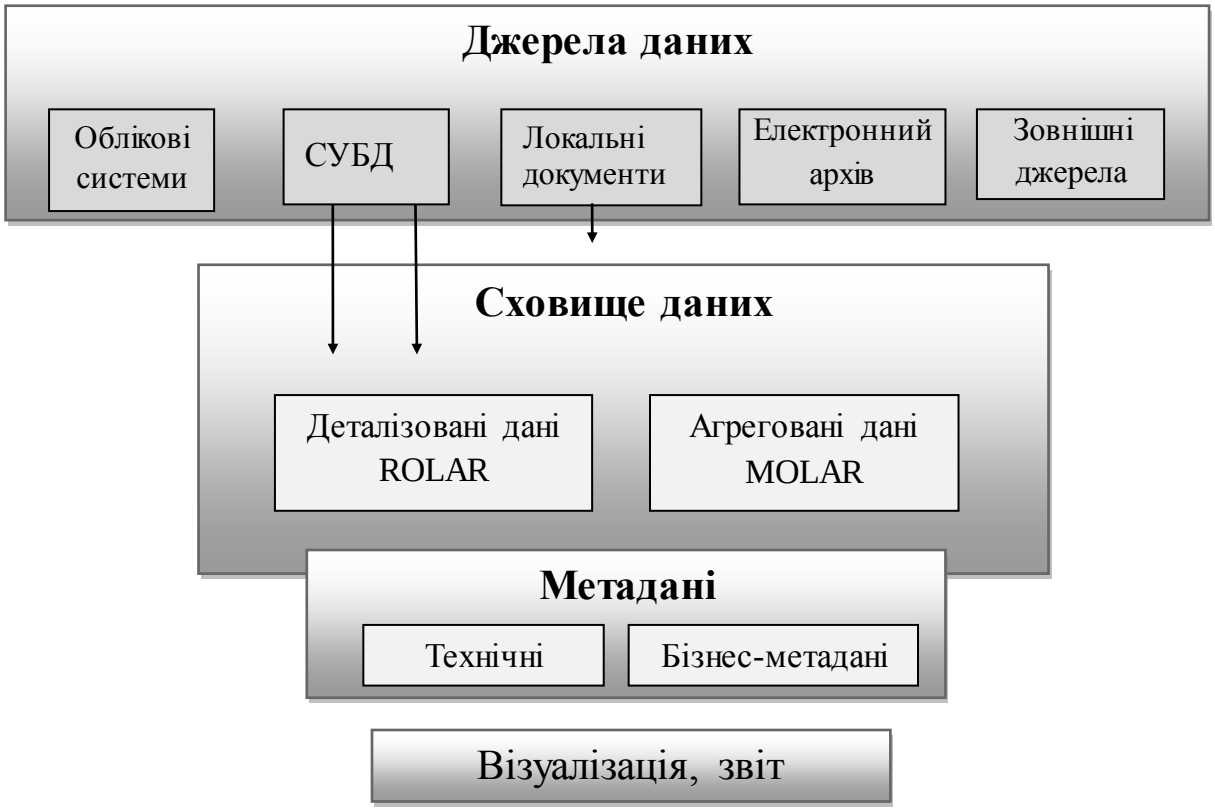


Рис. 2.10. Гібридне СД

Недоліком гібридної моделі є ускладнення адміністрування СД через більш складний регламент його поповнення, оскільки при цьому необхідно узгоджувати зміни в реляційній і багатовимірній структурах. Однак існує і ряд переваг.

1. Зберігання даних у реляційній структурі робить їх більшою мірою системно незалежними, що особливо важливо під час використання в управлінні підприємством економічної інформації (показників).

2. Реляційна структура формує стійкі і несуперечливі опорні точки для багатовимірного сховища.

3. Оскільки реляційне сховище підтримує актуальність і коректність даних, воно забезпечує досить надійний транспортний рівень для доставки інформації в багатовимірне сховище.

З гібридної архітектури СД зазвичай пов’язують ще одне важливе поняття – вітрини даних. Ситуація, коли для аналізу необхідна вся інформація, що міститься в СД, виникає значно рідко. У більшості випадків підрозділи організації використовують профільну інформацію, що стосується тільки того напряму діяльності, який вони обслуговують. Як правило, обсяг такої тематичної інформації

невеликий порівняно із загальним обсягом сховища і цілком ефективно можуть обслуговуватися MOLAP-системою.

Концепція вітрини даних полягає у виділенні профільних даних, найчастіше використовуваних за певним напрямом діяльності, в окремий набір, а в організації його зберігання – в окрему багатовимірну БД, підключену до централізованого РСД. Найчастіше для побудови вітрин даних використовується багатовимірна модель, оскільки за невеликих обсягів даних вона забезпечує більш швидкий відгук на запити, ніж реляційна, хоча в деяких випадках використовується і реляційна модель.

Вітрина даних – спеціалізоване локальне тематичне сховище, підключене до централізованого СД і обслуговуюче окремий підрозділ організації або певний напрям її діяльності.

На рис. 2.11 зображена система консолідації з використанням вітрин даних.

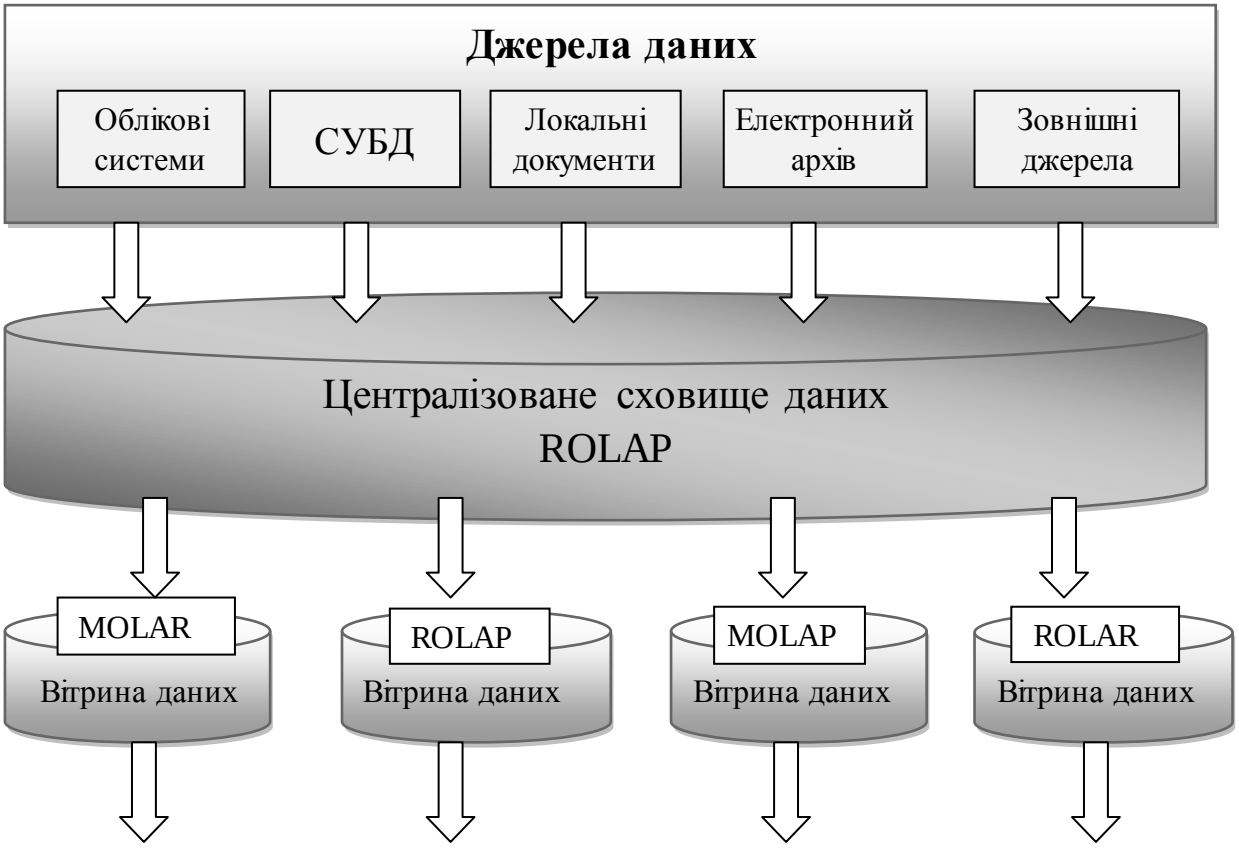


Рис. 2.11. Консолідація з використанням вітрин даних

Використання вітрин даних має такі переваги, які роблять їх ближче і доступніше останньому користувачеві:

– зміст даних, тематично орієнтованих на конкретного користувача;

– відносно невеликий обсяг збережених даних, на організацію і підтримку яких не потрібно значних витрат;

– поліпшені можливості в розмежуванні прав доступу користувачів, оскільки кожен з них працює тільки зі своєю вітриною і має доступ тільки до інформації, що належить до певного напряму діяльності.

Використання вітрин даних найбільш ефективно у великих організаціях з великою кількістю незалежних підрозділів, кожний з яких вирішує власні аналітичні завдання. У цьому випадку вітрини даних можуть застосовуватися як самостійно, так і разом з централізованим СД. Однак використання самостійних вітрин даних пов'язане з такою проблемою, як багаторазове дублювання даних у різних вітринах, що в кінцевому підсумку може призвести до суперечливості даних.

Використання вітрин даних разом з централізованими СД дозволяє підвищити достовірність даних, одержуваних користувачами вітрин, оскільки вітрини «харчуються» даними зі сховищ, де автоматично підтримується цілісність і несуперечність даних та проводиться їхнє очищення. Крім того, корпоративна інформаційна система може ефективно нарощуватися за рахунок додавання нових вітрин даних. І нарешті, використання вітрин дозволяє знизити навантаження на централізоване СД.

Прикладом організації централізованого СД з вітринами даних для організації з великою кількістю підрозділів може слугувати схема, представлена на рис. 2.12.

За всіма позитивними сторонами сховища даних як окремого консолідованого джерела зустрічаються ситуації, коли ця ідея не працює. Справа в тому, що існує практика нічного завантаження зібраних за день даних. Такий регламент дозволяє зменшити навантаження на OLTP-систему впродовж робочого дня, тобто в період її активного використання. Частіше ніж один раз на добу поповнювати СД сенсу немає. Однак такий стан речей не забезпечує можливості аналізувати інформацію впродовж робочого дня відразу за мірою її надходження. Іноді це досить критично, і доводиться шукати альтернативу традиційному СД.

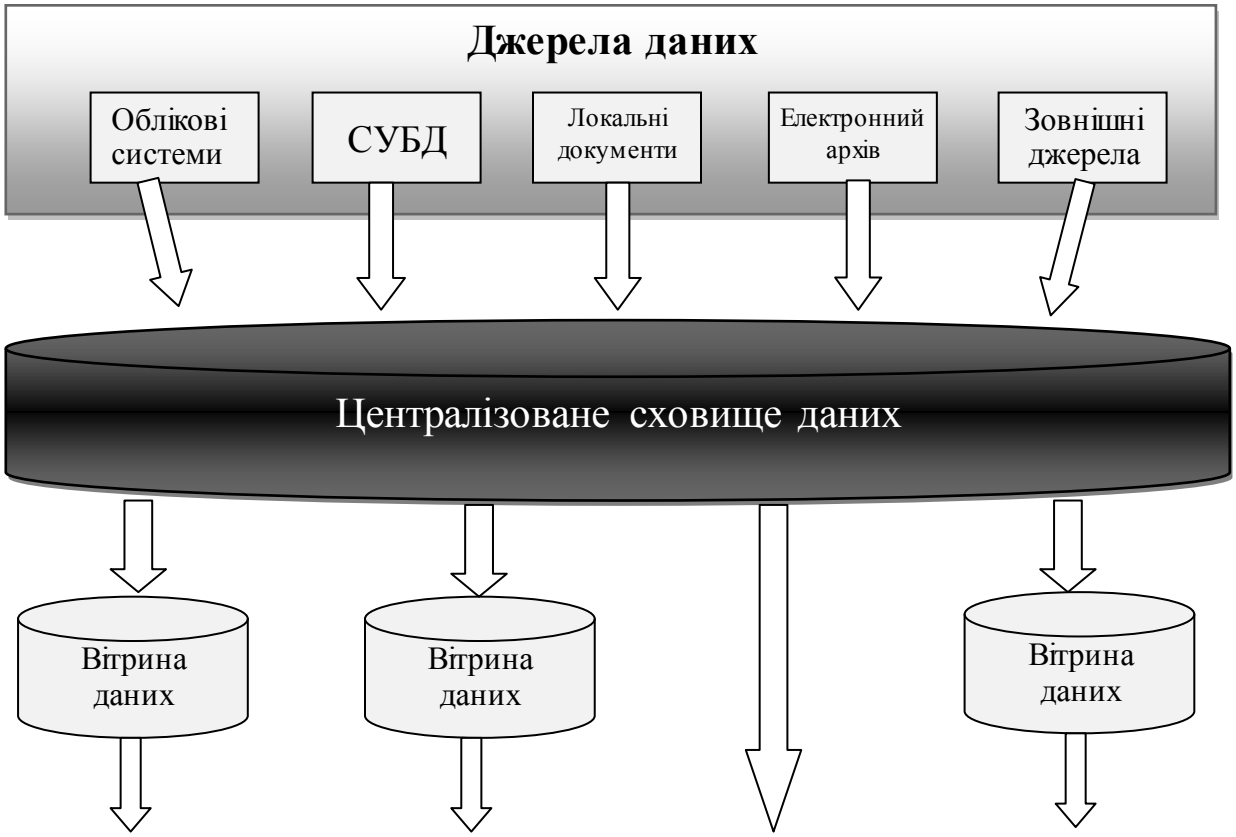


Рис. 2.12. Централізоване СД з вітринами даних

Крім того, неминучою проблемою під час використання СД у бізнес-аналітиці є надмірність. Надмірність певною мірою характерна як для реляційних, так і багатовимірних сховищ. Ситуація ускладнюється ще й тим, що СД зберігають історичну інформацію і реалізують принцип незмінності даних. Тобто на відміну від звичайних систем оперативної обробки, де зберігаються лише актуальні дані, СД можуть тільки поповнюватися новими даними, а видалення історичних даних не проводиться. Крім того, часто потрібно зберігати великі обсяги агрегованих даних. Подолати перераховані вище проблеми дозволяє концепція віртуального сховища даних (далі – ВСД). У її основі полягає принцип, відповідно до якого дані з локальних джерел, зовнішнього оточення, баз даних і облікових систем не консолідуються в єдине СД фізично, а вилучаються, перетворюються й інтегруються безпосередньо під час виконання запиту в оперативній пам'яті комп'ютера. Віртуальним сховищем даних називається система, яка працює з розрізненими джерелами даних і емулює роботу звичайного сховища даних, вилучаючи, перетворюючи й інтегруючи дані безпосередньо в процесі виконання запиту.

До переваг такого підходу належить:

– можливість аналізу даних в OLTP-системі відразу після їхнього надходження без очікування завантаження в сховище;

– мінімізується обсяг необхідної дискової та оперативної пам'яті, оскільки відсутня необхідність зберігання історичних даних і численних агрегованих даних для різних рівнів узагальнення інформації;

– наявність у ВСД розвиненого семантичного шару, який дозволяє аналітику повністю абстрагуватися від проблем, пов'язаних з процесом отримання даних з різноманітних джерел, і зосередитися на вирішенні завдань аналізу даних.

Під час роботи з ВСД користувач має справу з «ілюзією» сховища даних (рис. 2.13). Віртуальність передбачає, що ВСД існує тільки доти, доки працює відповідна програма. Як тільки вона завершує роботу, віртуальне сховище припиняє існування.

Концепція ВСД має ряд недоліків порівняно з СД, де інформація консолідується фізично.

1. Джерела даних, інформація з яких запитується у ВСД, можуть виявитися недоступними, якщо доступ до них здійснюється мережею або якщо змінилося місце їхньої локалізації.

2. Відсутня автоматична підтримка цілісності і несуперечності даних, можуть бути втрачені окремі фрагменти документів тощо.

3. Дані в джерелах зберігаються в різних форматах і кодуваннях, що може призвести до помилок за їхньою обробкою і спотворення інформації, отриманої у відповідь на запит.

4. Через можливу неузгодженість моментів поповнення джерел даних і відсутність підтримки в них хронології за одним і тим самим запитом у різні моменти часу можуть бути отримані будь-які дані.

5. Практично неможлива робота з даними, накопиченими за тривалий період часу, оскільки у ВСД доступні тільки ті дані, які перебувають у джерелах у конкретний момент часу.

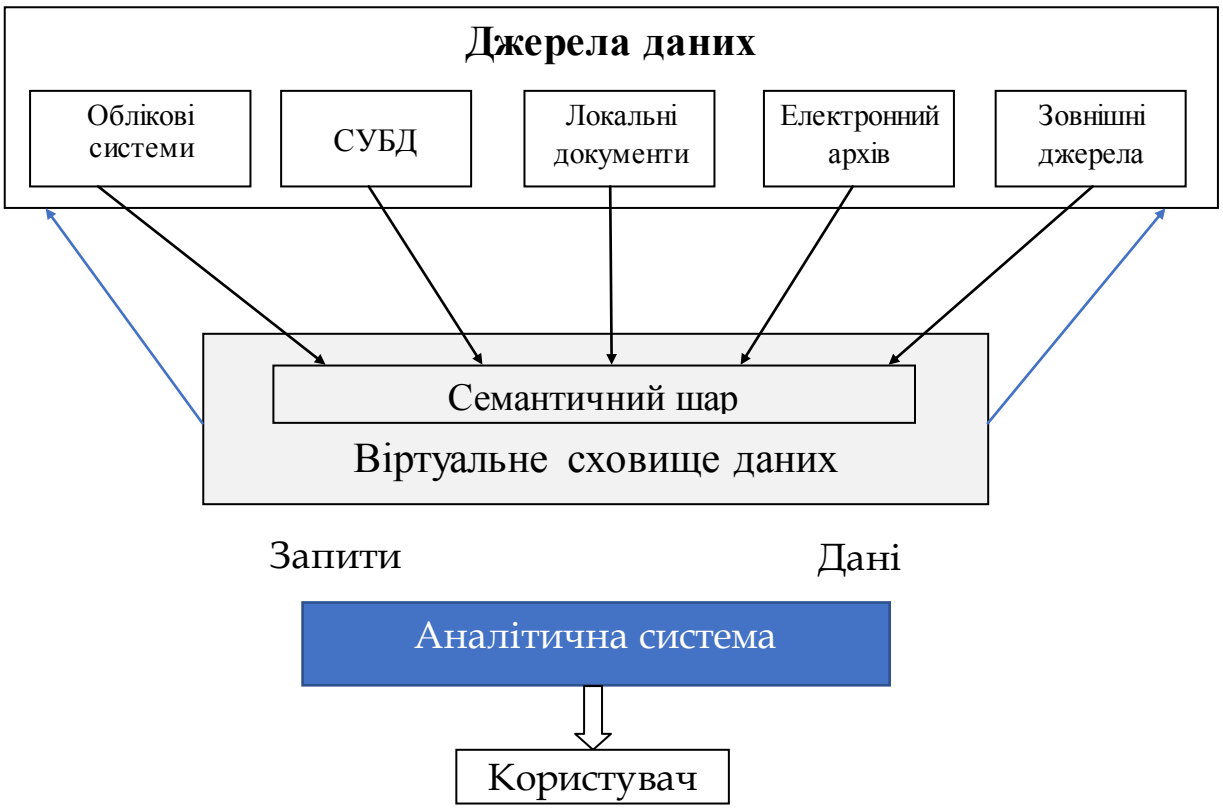


Рис. 2.13. Віртуальне СД

Найважливішою особливістю ВСД є те, що вони, працюючи безпосередньо з джерелами, що містять дані оперативного обліку, мають справу з даними в межах деякого періоду актуальності. Це пов'язано з тим, що OLTP-системи не зберігають історичні дані. Тому якщо історичні дані відіграють важливу роль під час аналізу, то потрібно застосовувати різновиди СД з фізичною консолідацією даних. А ВСД варто використовувати в системах, орієнтованих на аналіз оперативної інформації, актуальної тільки впродовж обмеженого періоду.

Більш стабільна інформація, наприклад про клієнтів або постачальників, зберігається в обліковій системі, підтримуваної адміністратором. За необхідності проаналізувати, наприклад, ефективність роботи з клієнтами на основі їхньої купівельної активності система самостійно звертається до відповідних джерел, збираючи потрібну інформацію. При цьому користувач працює з ними як з єдиним джерелом інформації, незважаючи на те, що дані отримані з декількох джерел різних типів. Така ситуація схематично проілюстрована на рис. 2.14.

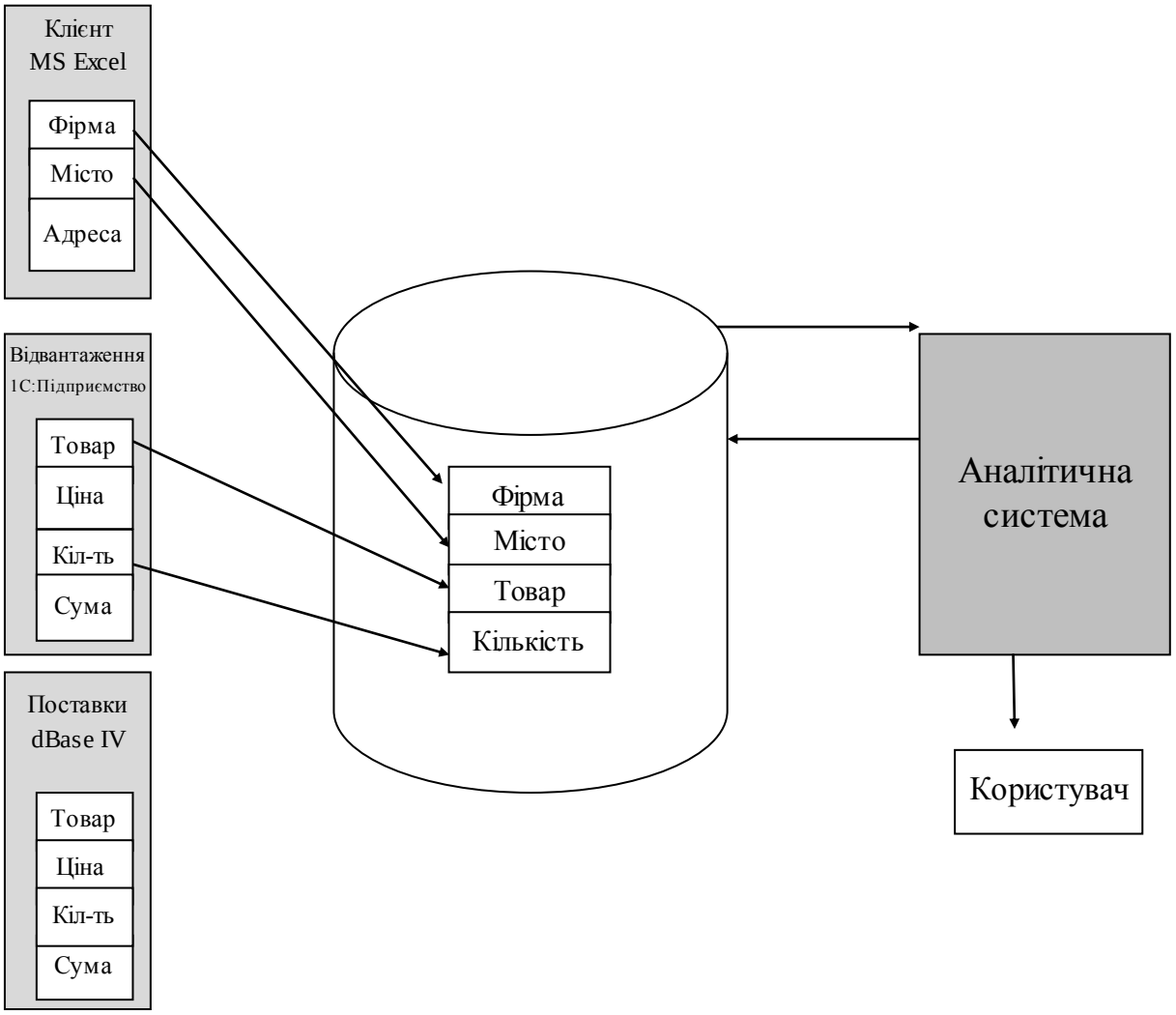


Рис. 2.14. Варіант організації ВСД

Навіть коли одна і та сама фірма виступає в ролі і постачальника, і покупця, не відбувається дублювання або протиріччя. В одному випадку запит пов'язує фірму з даними з таблиці «Поставки», а в іншому – з даними з таблиці «Відвантаження». Таким чином, віртуальне сховище, формуючи запит «нальоту», дозволяє мінімізувати надмірність і більш ефективно використовує дисковий простір.

Таким чином, застосування ВСД виявляється корисним для підприємств, які не мають технічних засобів і кваліфікованого персоналу для підтримки фізичних СД. Особливо великі переваги ВСД коли виникає необхідність аналізувати найсвіжішу інформацію. У ВСД відсутній етап завантаження даних, тому часовий інтервал між появою інформації в OLTP-системі і її готовністю до аналізу даних мінімальний.

Запитання до параграфа 5

1. Які основні завдання систем підтримки прийняття рішень?
2. Що таке системи підтримки прийняття рішень?
3. Що таке консолідація даних?
4. Що таке очищення даних?
5. Які завдання вирішуються в процесі консолідації даних?
6. Яка загальна схема процесу консолідації даних?
7. Що таке транзакція?
8. Що таке OLTP-система?
9. Яка структура інформаційної системи підтримки прийняття рішень?
10. Які відмінності OLTP-систем і СППР?
11. Які основні вимоги до СППР і OLTP-систем?
12. Що таке сховище даних?
13. Які основні особливості концепції сховища даних?
14. Які основні вимоги до СД?
15. Які основні положення полягають у концепції СД?
16. Яка концептуальна схема СД?
17. Які основні завдання вирішуються в процесі розробки СД?
18. Що таке метадані?
19. Які є основні підходи до використання СД?
20. Які є вимоги, яким повинні задовольняти продукти, що дозволяють виконувати оперативну аналітичну обробку, викладених Коддом?
21. Що таке тест FASMI?
22. Яке основне призначення багатовимірних сховищ даних?
23. Що полягає в основі багатовимірного представлення даних?
24. Який принцип організації багатовимірного куба?
25. Що таке реляційні сховища даних?
26. Які є схеми побудови РСД?
27. Які переваги схеми «зірка»?
28. Які основні переваги РСД наступні?
29. Яка загальна схема гібридного СД?
30. Яка загальна схема консолідації з використанням вітрин даних?
31. Що таке віртуальні СД?

§ 6. Введення в ETL

6.1. Поняття ETL

Вилучення даних з різнотипних джерел і перенесення їх у сховище даних з метою подальшої аналітичної обробки пов'язані з рядом проблем, основними з яких є такі:

1. Вихідні дані розташовані в джерелах найрізноманітніших типів і форматів, створених у різних додатках, і, крім того, можуть використовувати різне кодування, в той час, як для вирішення завдань аналізу даних повинні бути перетворені в єдиний універсальний формат, який підтримується СД і аналітичним додатком.

2. Дані в джерелах зазвичай надмірно деталізовані, тоді як для вирішення завдань аналізу в більшості випадків потрібні «великі цифри».

3. Вихідні дані, як правило, є «брудними», тобто містять різні фактори, які заважають їх коректному аналізу.

Тому для перенесення вихідних даних з різних джерел у СД використовують спеціальний інструментарій, який повинен отримувати дані з джерел різного формату, перетворювати їх в єдиний формат, підтримуваний СД, а за необхідності проводити очищення даних від факторів, що заважають коректно виконувати їхню аналітичну обробку. Такий комплекс програмних засобів отримав узагальнену назву ETL (extraction, transformation, loading – вилучення, перетворення, завантаження). Сам процес перенесення даних і пов'язані з ним дії називаються ETL-процесом, а відповідні програмні засоби – ETL-системами.

ETL – комплекс методів, що реалізує процес перенесення вихідних даних з різних джерел в аналітичний додаток або підтримує його сховище даних.

6.2. Основні цілі і завдання процесу ETL

Додатки ETL завантажують інформацію з одного або декількох джерел, перетворюють її у формат, який підтримує система зберігання й обробки, яка є одержувачем даних, а потім завантажують у неї перетворену інформацію. Нині ETL-системи все більш широко застосовуються саме для консолідації даних з метою їхнього подальшого аналізу. Очевидно, що оскільки СД можуть будуватися на основі різних моделей даних, то і процес ETL повинен розроблятися з урахуванням усіх особливостей використовуваної у СД моделі. Крім того, бажано, щоб ETL-система була універсальною, тобто могла отримувати і переносити дані якомога більшої кількості типів і форматів.

Незалежно від особливостей побудови і функціонування ETL-система повинна забезпечувати виконання трьох основних етапів процесу перенесення даних:

1. Завантаження даних. Дані вилучаються з одного або декількох джерел і готуються до перетворення. Варто зазначити, що для коректного представлення даних після їхнього завантаження в СД з джерел повинні вилучатись не тільки самі дані, але й метадані.

2. Перетворення даних. Здійснюється перетворення форматів і кодування даних, а також їхнє узагальнення та очищення.

3. Завантаження даних – запис перетворених даних у відповідну систему зберігання.

Переміщення даних у процесі ETL можна розбити на послідовність процедур, представлених такою функціональною схемою (рис. 2.15):

1. Вилучення. Дані беруться з джерел і завантажуються в проміжну сферу.

2. Пошук помилок. Проводиться перевірка даних на відповідність специфікаціям і можливість подальшого завантаження в СД.

3. Перетворення. Дані групуються і перетворюються до вигляду, відповідного структурі СД.

4. Розподіл. Дані розподіляються на декілька потоків відповідно до особливостей організації процесу їхнього завантаження в СД.

5. Завантаження. Дані завантажуються у сховище-одержувач.

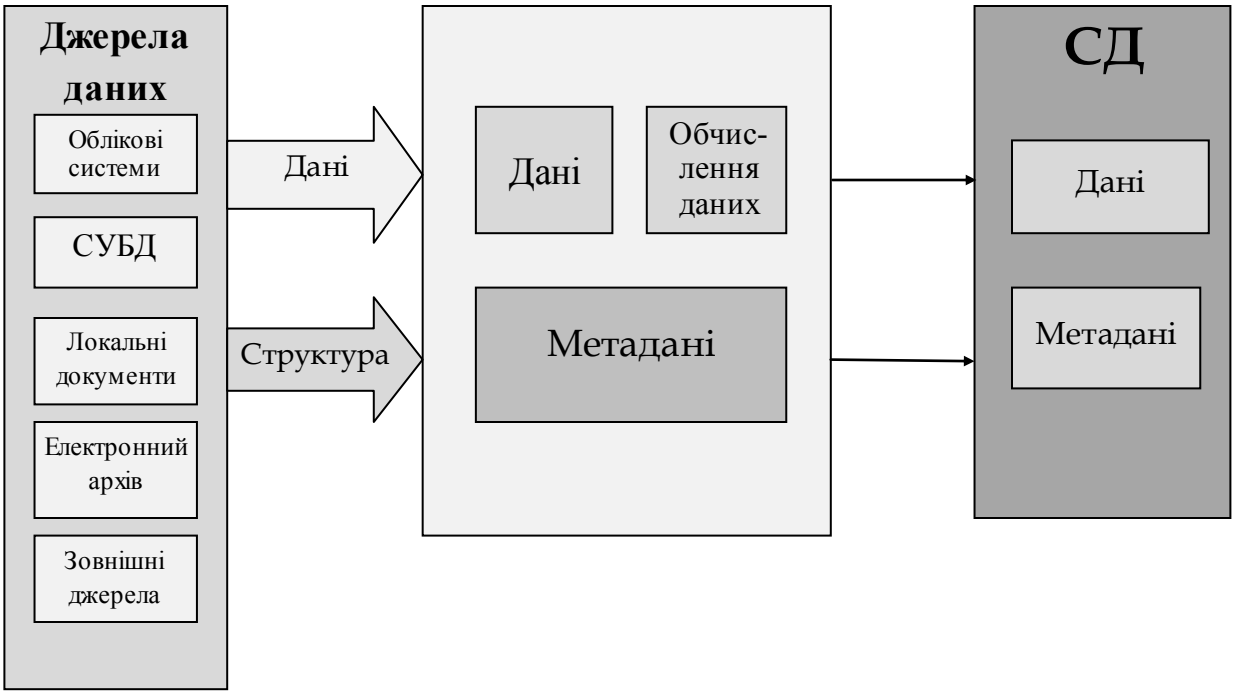


Рис. 2.15. Узагальнена структура процесу ETL

З точки зору процесу ETL архітектуру СД можна представити у вигляді трьох компонентів (рис. 2.16), таких як:

– джерело даних – містить структуровані дані у вигляді окремої таблиці, сукупності таблиць або просто файла, дані в якому, наприклад, впорядковані в типізовані стовпці, відокремлені один від одного деякими символами-роздільниками;

– проміжна сфера – містить допоміжні таблиці, створювані тимчасово і виключно для організації процесу вивантаження;

– одержувач даних – СД або просто БД, в яку повинні бути приміщені вилучені дані.

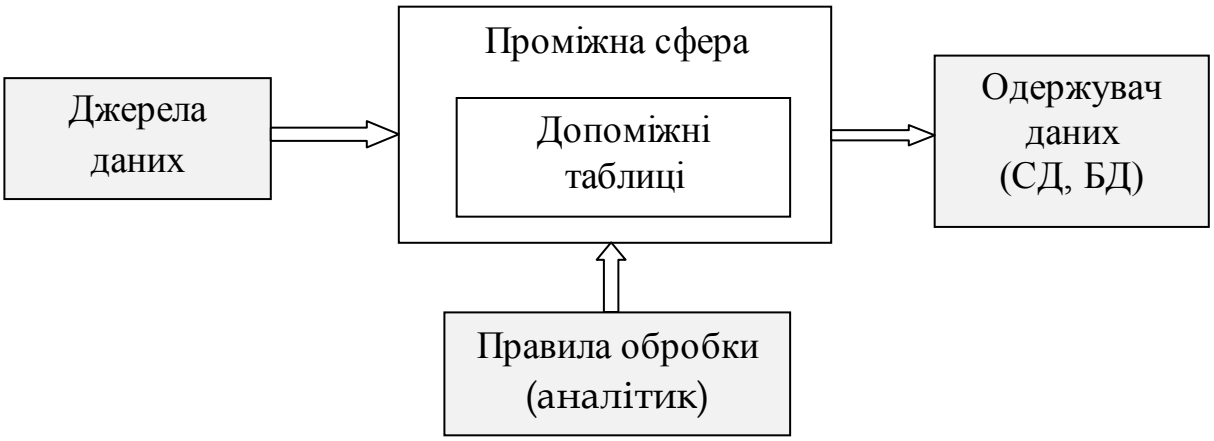


Рис. 2.16. Потоки даних в ETL

Переміщення даних від джерела до одержувача називається потоком даних. Вимоги до організації потоків даних описуються аналітиком.

ETL часто розглядають як просто підсистему перенесення даних з різних джерел у централізоване сховище. Завдання сховища – забезпечувати надійний і швидкий доступ до даних, підтримувати їхню хронологію, цілісність і несуперечливість. Тому, у першу чергу, ETL виявляється відокремленим від аналізу даних. ETL варто розглядати не тільки як процес перенесення даних з однієї програми до іншої, а й як інструмент їхньої підготовки до аналізу.

6.3. Вилучення даних в ETL

Під час розробки процедури вилучення даних, у першу чергу, необхідно визначити регламент завантаження СД і, відповідно, частоту вивантаження даних з OLTP-систем або окремих джерел. Наприклад, вивантаження може здійснюватися після закінчення заданого часового інтервалу (день, тиждень, місяць або квартал). У деяких випадках передбачається можливість позарегламентного вилучення даних після завершення певної бізнес-події (придбання нового бізнесу, відкриття філії, надходження великої партії товарів). Залежно від обсягу видобутих даних, складності доступу до них і швидкості роботи обладнання вивантаження даних займає певний час, яке так і називають – «вікно вивантаження». Очевидно, що упродовж «вікна вивантаження» різко збільшується навантаження на комп'ютерну з ETL організацію та її робота може виявитися частково або повністю паралізованою. Особливо це актуально для OLTP-систем, де в такі періоди може різко зрости час очікування відповіді. Тому «вікно вивантаження» намагаються вибрати так, щоб воно в мінімальному ступені впливало на робочий процес, наприклад в обідню перерву, відразу після завершення робочого дня або вночі.

Якщо вивантаження даних виробляти більш часто, то кількість «вікон завантаження» збільшується, але оскільки за менший період в OLTP-системі накопичується менше змін, то «вікна завантаження» стають коротшими і навантаження на систему – нижче.

Процедуру вилучення можна реалізувати двома основними способами:

1. Вилучення даних за допомогою спеціалізованих програмних засобів (рис. 2.17).

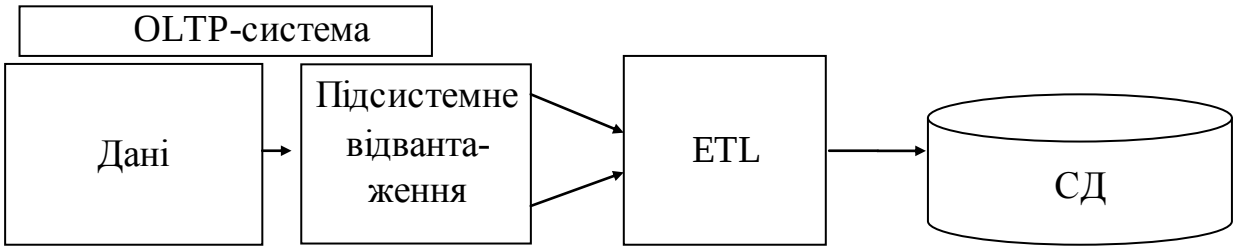


Рис. 2.17. Перший спосіб отримання даних в ETL

Переваги цього способу полягають у тому, що використання спеціалізованих програм дозволяє, по-перше, уникнути необхідності оснащувати OLTP-системи засобами вивантаження, по-друге, враховувати особливості всього ETL-процесу вже у процесі вивантаження.

2. Вилучення даних засобами тієї системи, в якій вони зберігаються.

Оскільки засоби «самовивантаження» розробляються з урахуванням особливостей структури даних OLTP-системи, це дозволяє адаптувати процедуру вилучення до структури видобутих даних, що в ряді випадків робить процес більш ефективним.

У деяких випадках, коли потрібно вивантажити дані з декількох джерел одного типу, для них створюється загальна таблиця; одне з її полів вказує на джерело, з якого були взяті дані. Приклад такої схеми організації вивантаження наведено на рис. 2.18.

Щоб розпочати процес видобування даних, у загальному випадку необхідно використовувати деяку службову інформацію:

– ім’я набору даних, з якого варто винести записи;

– номер запису, з якої починається вилучення;

– кількість видобутих записів;

– формат представлення даних;

– максимальна довжина запису, доступна для вилучення тощо.

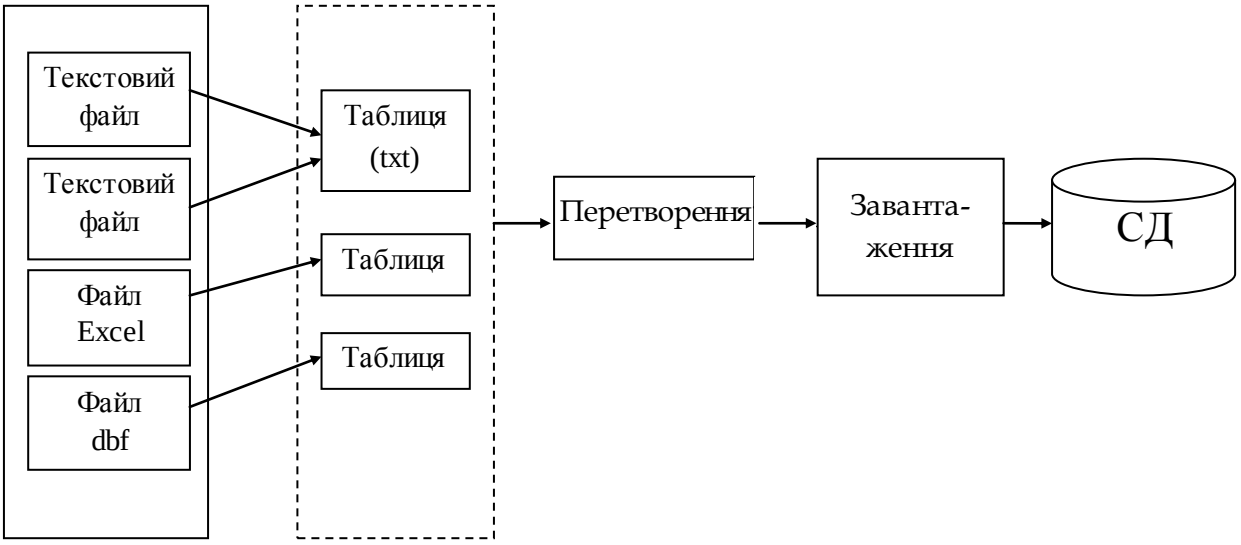


Рис. 2.18. Схема організації ETL

Вибір джерел даних. Усі джерела можна розділити на дві групи: розташовані в корпоративних інформаційних системах і на локальних комп'ютерах окремих користувачів. З точки зору дотримання регламенту і періодичності завантаження даних у СД переважні джерела з першої групи, оскільки вони також функціонують відповідно до визначеного порядку, який простіше узгодити з періодичністю завантаження даних у сховище. Що стосується джерел на локальних ПК, то зазвичай ніяких чітких правил роботи з ними не встановлюється: користувачі включають і вимикають комп'ютери, коли бажають; постійно виникають певні проблеми з ETL тощо. Проте саме на локальних комп'ютерах часто збирається досить цінна для аналізу інформація. Тому під час вибору джерел даних для завантаження у СД необхідно враховувати такі фактори:

– значимість даних з точки зору аналізу;
– складність отримання даних з джерел;
– можливе порушення цілісності та достовірності даних;
– обсяг даних у джерелі.

Як правило, доводиться знаходити компроміс між цими факторами. Наприклад, дані можуть являти собою безперечну цінність для аналізу, але складність їхнього вилучення чи некоректність структури може звести нанівець усі переваги від їхнього використання. В іншому випадку дані легкодоступні і не вимагають додаткової обробки під час завантаження в СД, але при цьому практично не являють собою інтерес з точки зору аналізу.

6.4. Особливості організації процесу вилучення даних

Після того як джерела, з яких будуть вилучатися дані, обрані, необхідно визначити, чи всі наявні в джерелах дані потрібні в СД.

Якщо вилучати потрібно не всі записи поспіль, а тільки певні, аналітик може описати набір умов, які дозволять відібрати тільки записи, що являють собою інтерес. Інший важливий момент – визначення глибини вивантаження даних за часом. Очевидно, що всі записи знадобляться тільки під час первинного заповнення сховища. У процесі його поповнення з джерел повинні вилучатись лише ті записи, які додавалися або змінювалися після минулого вилучення. Іноді сховища повністю очищають і перезавантажують. Але це можливо тільки в разі, якщо СД не досить велике і процес регенерації займає не значно багато часу, а також якщо регенерація відбувається рідко і за проміжок часу між ними велика частина даних змінюється або втрачає актуальність.

Питання визначення глибини вивантаження актуальне тільки у разі початкового завантаження сховища, коли потрібно визначити, інформація за який період часу є актуальною. У найпростішому випадку, коли ніяких міркувань з цього приводу немає, можна завантажити всі наявні записи. Однак цей підхід не завжди є оптимальним, оскільки в сховищі може виявитися багато інформації, що не являє собою цінності для аналізу. Таким чином, вибір глибини вивантаження історичних даних повинен забезпечити компроміс між обсягом вивантаження даних і їхньою цінністю з точки зору аналізу.

У разі повторних завантажень СД важливо не тільки визначити глибину вивантаження, а й організувати пошук змінених даних взагалі. Для цієї мети розроблені різні методики, наприклад використання міток часу. Для кожного зміненого запису створюється мітка, яка вказує на час зміни, і під час чергового вивантаження вилучаються тільки ті записи, мітки часу яких створені пізніше минулого вилучення.

Якщо джерелом даних є реляційна СУБД, то часто використовуються так звані тригери – спеціальні процедури, що запускаються автоматично у процесі виконання операцій вставки, оновлення або видалення. Тригери дозволяють зберігати змінені записи в спеціальній таблиці, з якої вони можуть бути вилучені під час наступного вивантаження. Однак використання тригерів істотно

збільшує навантаження на систему, тому якщо система вже працює з великим навантаженням, до застосування тригерів треба підходити обережно.

Процес вилучення даних у рамках ETL істотно залежить від типів і структури джерел даних. Можна виділити три різновиди джерел даних, з якими найчастіше стикаються організатори аналітичних проєктів.

Бази даних (SQL Server, Oracle, Firebird, Access тощо). У більшості випадків вилучення даних з баз даних не викликає проблем, оскільки структура даних у них чітко задана, відповідає певним стандартам і загальноприйнятим вимогам.

Структуровані файли різних форматів. Такі файли досить широко поширені, оскільки засоби їхнього створення загальнодоступні і не вимагають високої кваліфікації персоналу і продуктивності систем. До таких джерел належать текстові файли з роздільниками, файли електронних таблиць. Тут проблем більше, оскільки користувач може допускати помилки, пропуски, вводити суперечливі дані, втрачати фрагменти даних тощо. Єдиним плюсом є те, що для доступу до типових структурованих даних можна застосовувати такі стандартні засоби, як ODBC і ADO.

Неструктуровані джерела. Як правило, ця ситуація вимагає особливої уваги. Якщо уникнути використання неструктурованих джерел не виходить, потрібно застосувати спеціальні засоби їхнього перетворення в структорований вид. Коли джерело невелике, можливо, це вдасться зробити вручну. Але в більшості випадків доводиться розробляти спеціальний інструментарій, що враховує особливості організації даних у джерелі, і те, яку структуру з них варто створити. Існують також готові програмні системи для вирішення цього завдання. Остаточна мета структурування – так упорядкувати дані у файлі, щоб їх у певному вигляді можна було завантажити в реляційну таблицю.

Таким чином, вилучення даних з OLTP-систем, СУБД і окремих структурованих джерел у рамках ETL-процесу може виявитися завданням досить складним у технічному плані і важливим для ефективного аналізу даних. Тому розробкою та організацією процесу вилучення даних повинні займатися як технічні фахівці, так і аналітики.

6.5. Очищення даних в ETL

Наявність «брудних» даних – одна з найважливіших і важко формалізованих проблем аналітичних технологій взагалі і СД зокрема. Очищення даних обов'язкове під час їхнього перевантаження в сховище, і у процесі розробки стратегії ETL цьому приділяється велика увага. Варто зазначити, що, крім очищення даних перед їх завантаженням у сховище, користувач може виконати додаткове очищення засобами аналітичної системи вже після виконання запиту до СД. Таке дублювання цілком виправдане за рядом причин.

У даних, які з різних джерел, можуть міститися проблеми, через які виконати завантаження даних у СД буде неможливо. Остаточний користувач найчастіше не має уявлення про всі особливості даних у джерелах, з яких вони вилучаються, і тому не може розробляти стратегію очищення.

Вторинна очистка даних, передбачена в аналітичній системі, за своїми методами і цілями істотно відрізняється від очищення даних у процесі ETL. Доцільність застосування певного методу очищення даних, отриманих у результаті запиту із СД, визначається користувачем виходячи з особливостей конкретного завдання аналізу. При цьому часто використовується суб'єктивна думка аналітика, заснована на особистому досвіді.

Деякі види помилок можуть бути виявлені тільки після консолідації даних. Дійсно, не можна зробити висновок, що деяке значення є аномальним, доки не зроблено його порівняння із сусідніми значеннями.

Таким чином, первинна очистка даних у процесі ETL носить переважно технічний характер. Її основне завдання – підготувати дані до завантаження в сховище. Вторинна очистка в аналітичній системі спрямована на підготовку даних до вирішення конкретної аналітичної задачі. Тому обидва етапи очищення однаково важливі і необхідні.

Критерії оцінки якості даних. Щоб розробити методику очищення даних у процесі ETL, необхідно визначити критерії оцінки їхньої якості. Одним з таких критеріїв є критичність помилок. У зв'язку з цим дані можуть бути розділені на три категорії:

– дані високої якості, що не потребують очищення;

– дані, що містять критичні помилки, через які вони в принципі не можуть бути завантажені в СД (наприклад, буква або пропуск у числовому значенні, неправильний роздільник цілої і дробової частин числа тощо). Тобто критичними є помилки, які роблять неможливою подальшу роботу з даними;

– дані, що містять некритичні помилки, які не заважають їх завантаженню в СД, але при цьому дані є некоректними з точки зору їхнього аналізу.

Некритичні помилки можуть бути виправлені в процесі аналізу даних засобами аналітичної системи.

Існує декілька проблем, через які дані потребують очищення. Значно поширені проблеми, пов'язані з порушенням структури даних:

– коректність форматів і уявлень даних;

– унікальність первинних ключів у таблицях БД;

– повнота і цілісність даних;

– повнота зв'язків;

– відповідність деяким аналітичним обмеженням тощо.

Для кожного структурного рівня даних: окремого осередку, запису, таблиці, окремої або множини БД – характерні свої чинники, що знижують якість даних. Найбільш типові помилки, відповідні структурним одиницям БД, представлені в табл. 2.6.

Таблиця 2.6

Типові помилки, відповідні структурним одиницям БД

<table>
<tr><th>Комірка</th><th>Запис</th><th>Таблиця</th><th>Окрема БД</th><th>Множина БД</th></tr>
<tr><td>Орфографічні помилки. Пропуски в даних</td><td rowspan="4">Протиріччя між осередками</td><td rowspan="2">Дублювання записів</td><td rowspan="2">Цілісність даних</td><td>Невідповідності структури даних</td></tr>
<tr><td>Фіктивні значення</td><td>Однакові найменування різних атрибутів</td></tr>
<tr><td>Логічні невідповідності</td><td rowspan="2">Суперечливі записи</td><td rowspan="2">Протиріччя</td><td>Різне уявлення однотипних даних</td></tr>
<tr><td>Закодовані значення. Складові значення</td><td>Різна тимчасова шкала</td></tr>
</table>

На рівні окремої комірки таблиці найбільш характерні такі помилки.

Орфографічні або друкарські помилки можуть призвести до неправильного розуміння або спотворення даних.

Пропуски даних – відсутність даних у тих осередках, де вони повинні бути. Пропуски можуть бути викликані помилкою оператора або відсутністю відповідної інформації.

Фіктивні значення – дані, які не мають сенсу, ніяк не пов'язані з описуваним ними бізнес-процесом. Подібна ситуація можлива, якщо оператор системи OLTP, не маючи необхідної інформації, ввів в осередок довільну послідовність символів. Оператор буває змушений це зробити, коли система не дозволяє продовжувати роботу, доки відповідний осередок не буде заповнений. Отже, якщо клієнт забув вказати в анкеті вік, то оператор може ввести значення 0 або 999. Взагалі виявити і відфільтрувати фіктивні значення досить важко, оскільки поняття «сенс» є суб'єктивним і не піддається формалізації. Тому на етапі очищення даних засобами ETL можна виробити тільки загальні формальні методи виявлення фіктивних значень і боротьби з ними.

Логічні невідповідності виникають, якщо значення в клітинці не відповідає логічному змісту поля таблиці. Наприклад, замість найменування фірми вказано її банківський рахунок.

Закодовані значення – скорочення, абревіатури, заміна найменувань числовими кодами. Складові значення з'являються у файлах, де структура полів чітко не задана, наприклад у звичайних текстових файлах, електронних таблицях тощо.

На рівні запису виникають в основному проблеми суперечливості даних у різних осередках. Наприклад, сума, на яку був проданий товар, не відповідає добутку ціни за одиницю та кількості проданих одиниць.

На рівні таблиці основними проблемами є дублюючі і суперечливі записи.

Дублікати – це повністю ідентичні записи. Самі по собі вони не викликають порушень у структурі даних, і їхня наявність не є критичною. Однак дублікати також можуть викликати ряд проблем.

1. Збільшується обсяг даних і, відповідно, час, необхідний на їхнє завантаження й обробку.

2. Тимчасові витрати під час аналітичної обробки дублікатів збільшуються, при цьому якість аналізу не підвищується.

3. Дублікати можуть погіршити якість аналітичних моделей, заснованих на навчанні (нейронні мережі, дерева рішень). Якщо значний відсоток навчальної вибірки для таких моделей становитиме дублюючі записи, то результати навчання можуть погіршитися.

Аналітичні системи зазвичай містять засоби для виявлення і видалення дублюючих записів. Тому якщо в процесі ETL боротьба з дублікатами не була передбачена, це можна буде зробити у разі підготовки даних до аналізу безпосередньо в аналітичній системі.

Суперечливими є записи, які для двох різних об'єктів містять однакові значення унікальних атрибутів. Наприклад, у базі даних клієнтів для двох різних організацій вказані однакові банківські реквізити.

На рівні окремої БД основною проблемою є порушення цілісності даних, яке полягає в тому, що відбувається неузгодженість між різними об'єктами БД. Наприклад, якщо в БД є документ, який супроводжує угоду з будь-яким клієнтом, то, як правило, під час створення нового документа ім'я клієнта не вводиться вручну, а обирається зі списку клієнтів, що міститься в іншій таблиці БД і відкривається у разі заповнення відповідного поля. У результаті документ містить не власне ім'я клієнта, а посилання на відповідний об'єкт іншої таблиці. Якщо з будь-якої причини об'єкт (тобто запис про клієнта) буде видалений, то документ не зможе бути сформований через те, що об'єкт, на який є посилання, виявляється недоступний.

На рівні засобів БД виникають проблеми, пов'язані з відмінностями структури даних у різних базах:

– різні правила призначення імен полів (наприклад, різна кількість символів, допустимих в імені поля, наявність прогалин, символів національних алфавітів тощо);

– відмінності у використовуваних типах полів;

– однакові назви полів для різних атрибутів (наприклад, в одному джерелі даних поле Name містить назви фірм, а в іншому – найменування товарів);

– різна тимчасова шкала (наприклад, дані, що описують один і той самий бізнес-процес, зберігаються в двох різних файлах, але в одному з них представлена щоденна інформація, а в іншому – щотижнева).

Стратегія очищення даних повинна розроблятися з урахуванням особливостей предметної сфери, функціонування OLTP-систем та порядку збору даних. Крім того, необхідно пам'ятати, що повністю очистити дані вдається досить рідко. Існують проблеми, які не виходить вирішити незалежно від ступеня докладених зусиль. Бувають випадки, коли некоректне застосування методів очищення даних тільки погіршує ситуацію. Іноді використання досить складних алгоритмів очищення збільшує час перенесення даних у СД до неприйнятної величини. Тому не завжди варто прагнути до повного очищення даних. Краще забезпечити компроміс між складністю використовуваних алгоритмів, витратами на обчислення, часом, необхідним на очистку, і її результатами. Якщо достовірність якихось даних не впливає на результати аналізу, то від їхнього очищення, можливо, варто взагалі відмовитися.

6.6. Перетворення даних в ETL

Етап ETL-процесу, наступний за завантаженням, – перетворення даних. Його мета – підготовка даних до розміщення в СД і приведення їх до виду, найбільш зручному для подальшого аналізу. При цьому повинні враховуватися деякі висунуті аналітиком вимоги, зокрема, до рівня якості даних. Тому в процесі перетворення може бути задіяний найрізноманітніший інструментарій, починаючи від найпростіших засобів ручного редагування даних до систем, що реалізують досить складні методи обробки та очищення даних.

У процесі перетворення даних у рамках ETL найчастіше виконуються такі операції (рис. 2.19):

– перетворення структури даних;

– агрегування даних;

– переклад значень;

– створення нових даних;

– очищення даних.



Рис. 2.19. Процес перетворення даних в ETL

Перетворення структури даних. У багатьох випадках дані надходять у сховище, інтегруючись з багатьох джерел, які створювалися за допомогою різних програмних засобів, методологій, угод, стандартів тощо. Дані з таких джерел можуть відрізнятися своєю структурною організацією: угодами про призначення імен полів і таблиць порядком їх опису, форматами, типами та кодуванням даних, наприклад точністю представлення числових даних, що використовуються роздільниками цілої та дробової частин, роздільниками груп розрядів та ін.

При цьому якщо таблиці фактів найчастіше відповідають вимогам СД, то таблиці вимірювань потребують додаткової обробки і можуть бути об'єднанні.

Агрегування даних. Як правило, як джерела даних для сховищ, виступають системи оперативної обробки даних, облікові системи, файли різних СУБД, локальні файли окремих користувачів тощо. Загальною властивістю всіх цих джерел є те, що вони містять дані з максимальним ступенем деталізації: відомості про щоденні продажі або навіть про кожний факт продажу окремо, про обслуговування кожного клієнта та ін. Поширена думка, що таке детальне відтворення подій у досліджуваному бізнес-процесі тільки піде на користь, оскільки дані ніколи не бувають зайвими, і чим більше їх буде зібрано, тим точніше виявляться результати аналізу.

Це не зовсім так. Елементарні події, з яких складається бізнес-процес, які також називають атомарними (тобто неподільними), по суті, є випадковими величинами, схильними до впливу множини різних випадкових чинників, – від погоди до настрою клієнта. Отже, інформація про кожну окрему подію в бізнес-процесі практично не має цінності.

Для достовірного опису предметної сфери використання даних з максимальним рівнем деталізації не завжди доцільно, тому найбільший інтерес для аналізу являють собою дані, узагальнені за деяким інтервалом часу, групою клієнтів, товарів тощо. Такі великі цифри називаються агрегованими (іноді агрегатами), а сам процес їхнього обчислення – агрегуванням.

У результаті агрегування велику кількість записів про кожну подію в бізнес-процесі замінюють відносно невеликою кількістю записів, що містять агреговані значення. Фактично проводиться об'єднання декількох записів в одне з обчисленням агреговане значення на основі значень кожного запису. Під час обчислення агрегатів може бути використано декілька способів.

Середнє – для даних, розташованих у межах інтервалу, в якому вони узагальнюються, обчислюється середнє значення. Потім усі записи з цього інтервалу замінюються одним, що містить їхнє середнє значення (рис. 2.20).

Дата	Ціна	Кількість	Сума
07.04	150.00	20	3000.00
08.04	135.00	10	1350.00
09.04	220.00	15	3300.00
10.04	173.00	5	865.00
11.04	245.00	24	5880.00
12.04	96.00	12	1152.00
13.04	110.00	320	3520.00

Дата	Середня кількість	Середня сума	Максимальна сума	Мінімальна сума	Кількість сум	Медіана суми
07.04–13.04	16.85	2723.85	5860	865	7	3000

Рис. 2.20. Приклад агрегування

Сума – записи замінюються однією, в якій вказується сума значень.

Максимум – у результуючому записі залишається максимальне значення з усіх, що об'єднуються.

Мінімум – у результуючому записі залишається мінімальне значення.

Кількість унікальних значень – результатом агрегування буде кількість унікальних значень, що з'являються в осередках одного і того самого поля.

Кількість – результатом агрегування буде кількість записів, що містяться в полі.

Медіана – обчислюється медіана об'єднаних значень.

З усіх можливих варіантів агрегування варто вибрати найбільш значимі з точки зору запланованих напрямів аналізу, а від інших відмовитися. Очевидно, можна відмовитися від агрегатів, які мають малу кількість підлеглих агрегованих значень (наприклад, агрегування щомісячних продажів за квартал), оскільки їх легко обчислити в процесі аналізу. Або, навпаки, можна відмовитися від агрегатів з максимальним ступенем деталізації (наприклад, агрегування щоденних продажів). Другий варіант найкращий, оскільки, у першу чергу, обіцяє істотну економію, тому що кількість днів у році множиться на кількість товарів і продавців. Однак якщо дані про продажі є розрідженими, тобто не кожен товар продається щодня, то економія може виявитися досить незначною.

Вибір потрібних агрегатів завжди визначається особливостями бізнесу. При цьому варто пам'ятати, що агрегати, необхідні для аналізу, можуть бути обчислені безпосередньо під час виконання аналітичного запиту до СД, хоча тим самим час його виконання дещо збільшиться. Подібний підхід дозволяє, наприклад, відмовитися від агрегування рідко використовуваних даних.

Таким чином, вибір правильної стратегії агрегування даних в ETL – складне і суперечливе завдання. Збільшення кількості агрегатів у СД призводить до збільшення його розмірів і складності структури даних. Зниження кількості агрегатів у СД може призвести до необхідності їхнього обчислення в процесі виконання аналітичних запитів, що збільшить час очікування користувача. Отже, необхідно забезпечити розумний компроміс між цими факторами.

Часто дані в джерелах зберігаються з використанням спеціальних кодувань, які дозволяють скоротити надмірність даних і тим самим зменшити обсяг пам'яті, необхідної для їхнього зберігання. Так, найменування об'єктів, їх властивостей і ознак можуть

зберігатися в скороченому вигляді. У цьому випадку перед завантаженням даних у сховищі потрібно виконати переклад таких скорочених значень у більш повні і, відповідно, зрозумілі.

Створення нових даних. У процесі завантаження в СД може знадобитись обчислення деяких нових даних на основі існуючих, що зазвичай супроводжується створенням нових полів.

Створення нової інформації на основі наявних даних тісно пов'язане з таким важливим процесом, як збагачення даних, яке може здійснюватись (частково або повністю) на етапі перетворення даних в ETL. Агрегування також може розглядатися як створення нових даних.

Очищення даних. Збір даних у процесі ETL проводиться з великої кількості джерел, багато з яких не містять автоматичних засобів підтримки цілісності, несуперечності та коректного представлення даних. У зв'язку з цим під час перенесення інформації у СД доводиться мати справу з потоками «брудних» даних, які можуть стати причиною неправильних результатів аналізу і навіть зробити неможливим застосування деяких аналітичних алгоритмів і методів. З цієї причини в процесі ETL застосовується очищення – процедура коригування даних, які в будь-якому сенсі не задовольняють певним критеріям якості, тобто містять порушення структури даних, протиріччя, пропуски, дублікати, неправильні формати тощо.

Деякі проблеми в даних є критичними і навіть не дозволяють виконати завантаження даних у СД. Інші проблеми менш критичні, не заважають переносу даних у СД, але не дозволяють їх коректно аналізувати. Критичні проблеми в даних повинні усуватися безпосередньо в процесі ETL. Некритичні фактори, що знижують якість даних, можуть усуватися як у процесі ETL, так і під час підготовки їх до аналізу в аналітичній системі.

Очищення даних – одне з найбільш важливих і водночас найбільш складних завдань ETL-процесу, оскільки набір факторів, що знижують якість даних, досить різноманітний і може постійно змінюватися. Тому очищенню даних у разі розробки ETL-процесів приділяють велику увагу.

Вибір місця для виконання перетворень даних. У принципі перетворення даних може бути виконано на будь-якому етапі ETE-процесу. Але іноді потрібно вибрати оптимальне місце для

здійснення перетворення. Деякі види перетворень зручніше виконувати «на льоту», в процесі отримання даних з джерела, інші – в проміжній сфері та в процесі завантаження даних у СД.

1. Перетворення в процесі вилучення даних. На цьому етапі найкраще виконувати перетворення типів даних і виробляти фільтрацію записів, що являють собою інтерес для СД.

2. Перетворення в проміжній сфері перед завантаженням даних у сховищі – найкращий варіант для інтеграції даних з багатьох джерел, оскільки в процесі вилучення даних, очевидно, цього зробити не можна. У проміжній сфері доцільно виконувати такі види перетворень, як сортування, групування, обробка часових рядів тощо.

3. Перетворення в процесі завантаження даних у СД. Окремі прості перетворення, наприклад перетворення регістрів букв у текстових полях, можуть бути виконані тільки після завантаження даних у сховище.

Таким чином, усі операції перетворення, які можуть бути потрібні під час перенесення даних у СД, як правило, не зосереджуються на одному кроці ETL-процесу, а розподіляються за різними етапами залежно від того, де виконання перетворення більш ефективно.

Після того, як дані вилучені з різних джерел і виконані перетворення, агрегація й очищення даних, здійснюється останній етап ETL – завантаження даних у сховище. Процес завантаження полягає в перенесенні даних з проміжних таблиць у структури сховища даних. Від продуманості й оптимальності процесу завантаження даних багато в чому залежить час, необхідний для повного циклу поновлення даних у СД, а також повнота і коректність даних у сховищі.

Першими в процесі завантаження даних у СД зазвичай завантажуються таблиці вимірювань, які містять сурогатні ключі й іншу описову інформацію, необхідну для таблиць фактів.

Під час завантаження таблиць вимірів потрібно додавати нові записи і змінювати існуючі.

У процесі додавання нових даних у таблицю вимірювань потрібно визначити, чи не існує в ній відповідний запис. Якщо немає, то вона додається в таблицю. В іншому випадку можуть використовуватися різні способи обробки змін залежно від того, чи потрібно підтримувати стару інформацію в сховищі з метою її подальшого аналізу.

У разі завантаження таблиці фактів нова інформація зазвичай додається в кінець таблиці, щоб не змінювати існуючі дані. Однією з основних проблем цього етапу ETL є те, що далеко не завжди дані завантажуються повністю: в завантаженні деяких записів може бути відмовлено. Відхилення записів відбувається з таких причин.

На етапі перетворення даних не вдається виправити всі критичні помилки, які блокують завантаження записів у СД.

Некоректний порядок завантаження даних. Наприклад, здійснюється спроба завантажити факти для значення вимірювання, яке ще не було завантажено. Причиною цього є неправильна розробка ETL-процесів.

Переривання процесу завантаження або зупинка його користувачем.

Незалежно від причин результатом завжди виявляється наявність певної кількості записів, які не були в сховищі. Як правило, передбачена система сповіщення користувача про те, що сховище містить не повністю завантажені дані. Це необхідно тому, що якщо неповні дані будуть використані для аналізу, то наслідки можуть бути непередбачуваними.

У разі появи даних, спроба завантаження яких зазнала невдачі, необхідно передбачити ETL такі дії (рис. 2.21):

– зберегти дані, що не потрапили в СД, у вигляді таблиці або файла того ж формату, що і вихідна таблиця або файл (такі таблиці зазвичай називають таблицями виключень);

– провести аналіз відхилених даних для виявлення причин, через які вони не були завантажені;

– за необхідності провести додаткову обробку й очищення даних, після чого зробити повторну спробу їхнього завантаження.

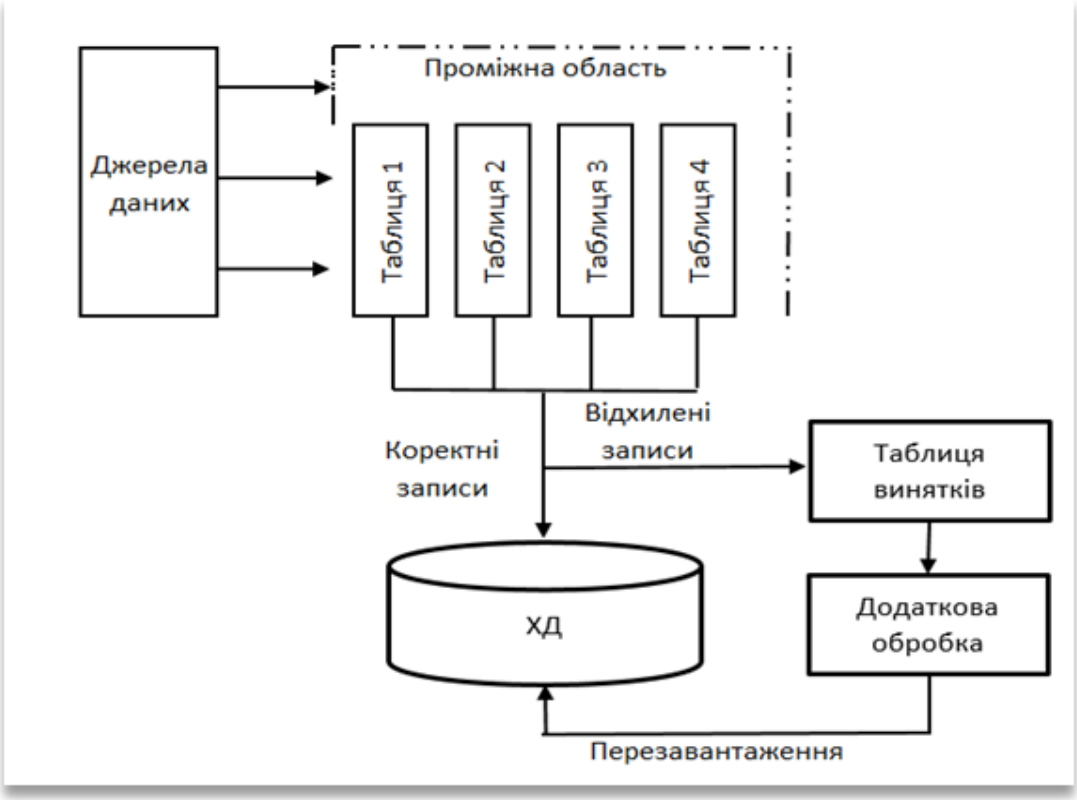


Рис. 2.21. Послідовність дій за наявності відхилених записів у процедурі завантаження в СД

Якщо і повторна спроба завантаження даних не досягла успіху, то в сховищі виявляться неповні дані, аналіз яких може призвести до неправильних висновків. Для вирішення цієї проблеми можна:

– відновити стан сховища, яким воно було до завантаження (наприклад, за допомогою резервної копії або операції відкоту, якщо вона передбачена);

– повністю очистити в СД таблиці з неповними даними, щоб виключити некоректні результати їхнього аналізу;

– залишити все як є і повідомити користувача про виниклі проблеми. У цьому випадку аналітик дізнається, що завантажена інформація недостовірна, і використовуватиме її з певними обмеженнями.

Багатопотокова організація процесу завантаження даних

Під час чергового завантаження в СД переноситься не вся інформація з OLTP-системи, а тільки та, яка була змінена упродовж проміжку часу, що пройшов з попереднього завантаження. При цьому можна виділити два види змін: додавання й оновлення (доповнення).

Додавання – в СД передається нова, що раніше не існувала, інформація, наприклад відомості про продажі, що відбулися з минулого завантаження, про появу нового клієнта, товару тощо.

Оновлення (доповнення) – в СД передається інформація, яка існувала раніше, але з якоїсь причини була змінена або доповнена (наприклад, змінено місто, в якому живе клієнт).

Для забезпечення цих функцій завантажуються дані, розподіляються за двома паралельними потоками (data flow): потік додавання й потік оновлення (рис. 2.22).

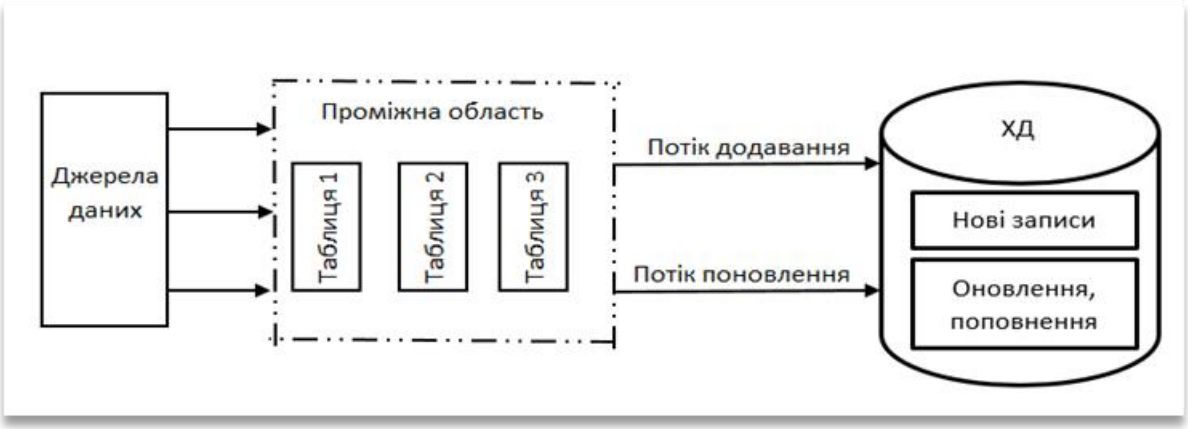


Рис. 2.22. Потік додавання і потік оновлення

Для розподілу даних під час завантаження на потоки використовуються засоби моніторингу змін у даних. Вони фіксують стан даних у деякі моменти часу і визначають, які дані були змінені або доповнені. Застосовуються такі методи:

– повне порівняння завантажуваних записів з усією інформацією, яка вже міститься в сховищі. Цей спосіб є найпростішим і водночас найбільш трудомістким у плані обчислювальних витрат;

– використання ознак модифікованих даних або полів Дата / Час для визначення останньої зміни запису (якщо такі поля передбачені в джерелі даних).

Розподіл даних у разі завантаження на потік додавання і потік оновлення дозволяє виконувати перенос даних у сховищі за допомогою звичайних запитів, не використовуючи будь-які фільтри.

Варто зазначити, що під час розробки методики завантаження даних у СД немає загального підходу до того, як модифікувати таблиці вимірювань. Після завершення завантаження

виконуються додаткові операції над даними, тільки що завантаженими в СД перед тим, як зробити їх доступними для користувача. Такі операції називаються постзавантажувальними. До них належать переіндексація, верифікація даних та ін.

З точки зору аналітика, найбільш важливим завданням є верифікація даних. Перш ніж використовувати нові дані для аналізу, корисно переконатися в їх надійності та достовірності. Для цих цілей можна передбачити комплекс верифікаційних тестів. Наприклад:

– під час підсумовування продажів щодо одного вимірювання результат повинен збігатися з відповідною сумою, отриманою щодо іншого пов'язаного з ним виміру, тобто сума продажів за всіма товарами за місяць повинна відповідати сумі угод, укладених з усіма клієнтами за той самий період;

– підсумковий показник за місяць повинен відповідати сумі щоденних або щотижневих показників у цьому місяці;

– сумарна виручка у всіх регіонах за поточний місяць повинна відповідати сумі продажів у всіх регіональних дилерських центрах.

Крім того, може виявитися корисним порівнювати дані не тільки в різних розрізах після їхнього завантаження в СД, але і з джерелами даних. Так, якщо значення будь-якого показника в джерелі і сховищі рівні, то все нормально, в іншому випадку дані, можливо, є некоректними.

Якщо тестування показало, що невідповідності, що дозволяють запідозрити втрату або недостовірність даних, відсутні, то можна вважати завантаження даних у СД успішним і приступати до аналізу нової інформації.

Досить часто виникають ситуації, коли дані для аналізу доводиться імпортувати безпосередньо з джерел, минаючи сховища даних та інші проміжні системи. Це може стати в нагоді в разі, коли організація відмовилася від використання СД або сховище є, але дані з деяких джерел не можуть бути в нього завантажені з технічних причин, через невідповідність моделям і форматам даних, що використовуються в СД.

Таким чином, частина даних надходить в аналітичну систему через СД з відповідним очищенням і підготовкою, а інша частина – безпосередньо з джерел (рис. 2.23).

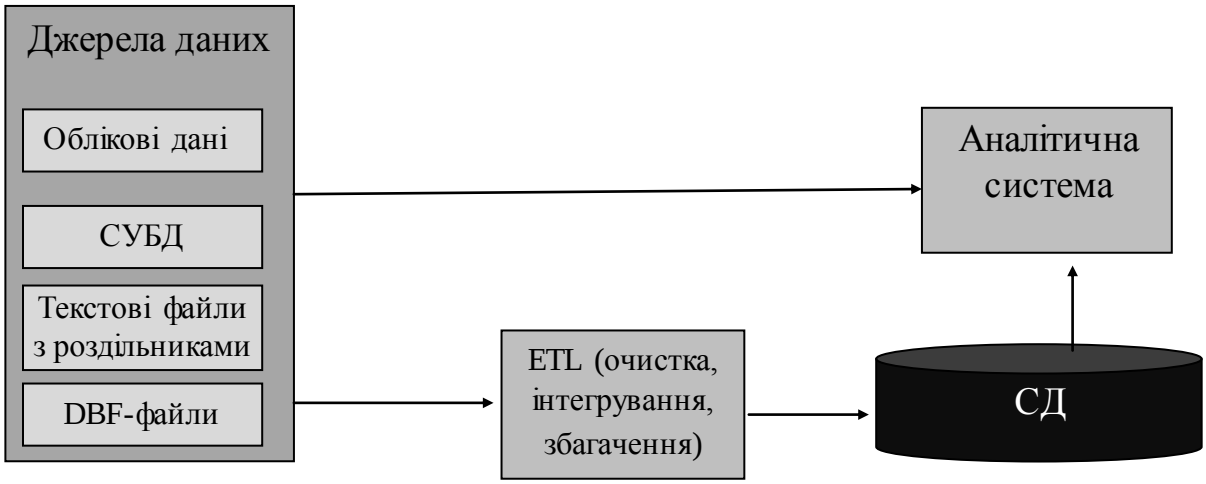


Рис. 2.23. Завантаження даних з локальних джерел

Вилучення даних безпосередньо з джерел має свої переваги і недоліки, а також породжує ряд проблем, пов'язаних як безпосередньо з процесом отримання, так і з якістю отриманих даних.

Основною перевагою, яке дає використання сховищ даних в аналітичних технологіях, є підвищення ефективності та достовірності аналізу даних, особливо для підтримки прийняття стратегічних рішень. Це досягається за рахунок оптимізації швидкості доступу до даних, можливості інтеграції даних різних форматів і типів, автоматичної підтримки цілісності і несуперечності, забезпечення хронології та історії даних.

Причини відмови від використання СД можуть бути такими:

1. Організація не має в своєму розпорядженні достатні ресурси (фінансові, технічні, кадрові) для розробки, придбання та підтримки СД.

2. Успадкована система аналітичної обробки ефективно функціонує з використанням звичайної реляційної СУБД.

3. Обсяг аналізованих даних невеликий, а застосовувані технології аналізу даних нескладні. Зокрема, не потрібна підтримка хронології даних, для аналізу використовуються тільки дані за актуальний період тощо.

4. Роль аналізу в діяльності організації невисока, адміністрація не усвідомила його значимість в отриманні конкурентних переваг.

Збагачення даних. У більшості випадків сховища даних створюються і підтримуються для забезпечення ефективного аналізу даних на підприємстві. Очевидно, що дані, зібрані для задач аналізу, повинні бути повними й достовірними, оскільки на основі неповних або недостовірних даних можна зробити неправильні висновки про стан бізнесу та шляхи його вдосконалення.

Неповні дані можуть з'явитися, наприклад, якщо частина відомостей про продажі філії фірми була загублена в процесі їхнього переносу в СД. Аналітик може прийти до висновку, що продажі в цій філії катастрофічно низькі, філія працює неефективно і її потрібно закрити, хоча насправді діяльність філії цілком успішна, а його співробітники добре справляються зі своїми завданнями. Недостовірні дані, які при цьому можуть бути повними, містять викривлену інформацію, що не дозволяє провести якісний аналіз. Тому в процесі завантаження в СД, а також підготовки до аналізу в аналітичному додатку дані перевіряються на повноту, цілісність, несуперечність, наявність помилок, пропусків, аномальних значень та інших чинників, які можуть призвести до некоректних результатів аналізу.

Крім достовірності та повноти даних, існує ще один фактор, що безпосередньо впливає на ефективність їхнього аналізу, – інформаційна насиченість. Взагалі зазначаючи, дані та інформація не зовсім одне й те саме. Інформація – це не будь-які дані. Дані – поняття об'єктивне. Вони або реально існують як зміни деякого фізичного процесу, або ні. А інформація в більшості випадків суб'єктивна. Якщо один експерт з певним рівнем компетентності, знань і досвіду побачить у деякому наборі даних корисну інформацію, то інший експерт з іншим рівнем досвіду і знань відшукає зовсім іншу інформацію або не знайде її зовсім.

Надійність і достовірність перевіряються практично на всіх етапах аналітичного процесу: спочатку на етапі завантаження даних у СД (у процесі ETL), потім у самому СД (автоматичний контроль) і, нарешті, в аналітичному додатку під час підготовки даних до аналізу.

Досить або недостатньо інформації для вирішення певної аналітичної задачі, кожен аналітик визначає сам на підставі значно

суб'єктивних критеріїв. Один аналітик навіть з мінімуму інформації дістане максимум корисних знань за допомогою особистого досвіду, навичок аналітичної роботи, вмілого застосування аналітичних методів та алгоритмів. Фахівцю з меншою кваліфікацією, можливо, не вдасться вирішити задачу з будь-якою кількістю даних. Крім того, самі аналітичні завдання розрізняються за рівнем складності і вимогам до інформативності вихідних даних.

Часто виникають ситуації, особливо під час вирішення нестандартних аналітичних задач, коли для аналізу потрібна інформація, якої чомусь не виявилося в наявності. Це може статися через непродуманий процес збору даних. Часом бази даних виявляються забиті будь-чим, тільки не даними, що мають пряме відношення до основних бізнес-процесів на підприємстві. Очевидно, що така частина інформації, що міститься в БД, не може зацікавити аналітика. Складається ситуація, яка проілюстрована лівою частиною рис. 2.24, коли у величезному масиві даних є тільки невелика їхня підмножина, що реально описує досліджуваний процес.

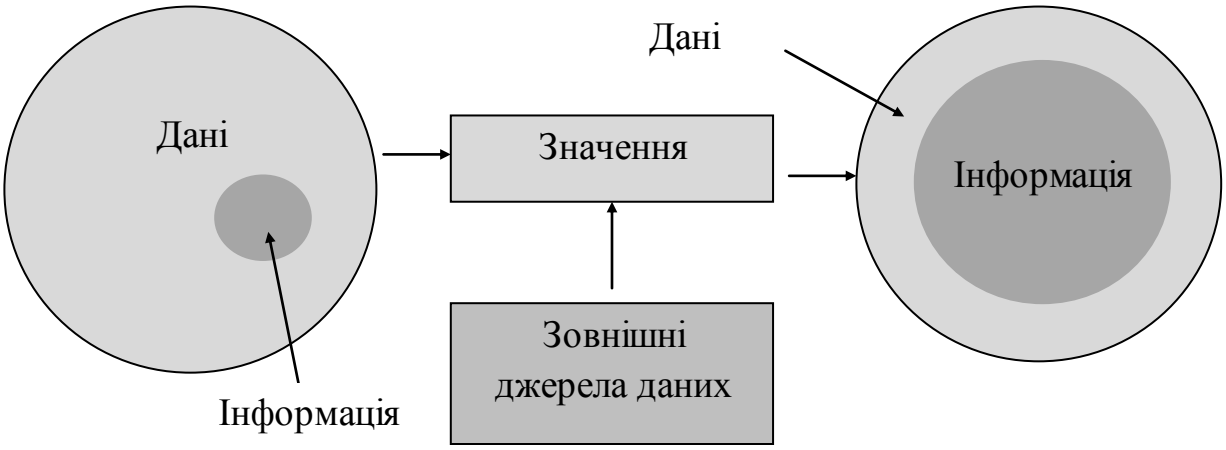


Рис. 2.24. Збагачення даних

Коли ж, нарешті, надходить час аналізувати дані, з'ясовується, що аналізувати нічого. У цей момент усвідомлюється необхідність збагачення даних. Воно може виконуватися за рахунок реорганізації самих даних: введення якихось кодувань, ознак станів об'єктів, підрозділи їх на категорії (наприклад, товари розподіляються за групами товарів) тощо. Може залучатися додаткова зовнішня інформація, наприклад історія курсів валют на день продажу, інформація про продажі конкурентів за той самий період та

ін. Поступово ситуація набуде вигляду, що представлена в правій частині схеми (рис. 2.24), коли корисної інформації стає більше.

Збагачення даних – процес насичення даних новою інформацією, яка дозволяє зробити їх більш цінними і значущими з точки зору вирішення певного аналітичного завдання. Можна виділити два основні методи збагачення даних: зовнішнє і внутрішнє.

Зовнішнє збагачення передбачає залучення додаткової інформації із зовнішніх джерел, що дозволить підвищити цінність і значимість даних з точки зору їх аналізу. Під підвищенням значущості даних мається на увазі, що на основі їхнього аналізу можна буде приймати управлінські рішення принципово нового рівня.

Зовнішніми джерелами можуть бути:

– інші підприємства та організації, що працюють у цій самій сфері діяльності або в суміжних сферах, причому як партнери, так і конкуренти;

– фінансово-кредитні установи, банки, страхові компанії;

– державні податкові і статистичні служби;

– органи державного і муніципального управління;

– різні служби соціальної сфери: міграційна служба, органи праці та зайнятості, система охорони здоров'я, пенсійні фонди тощо.

Джерелом інформації для збагачення даних можуть бути будь-які організації, які в процесі своєї діяльності збирають, структурують і зберігають інформацію, необхідну їм для здійснення своїх цілей.

Внутрішнє збагачення не передбачає залучення зовнішньої інформації. У цьому випадку підвищення інформативності та значущості даних може бути досягнуто за рахунок зміни їхньої організації. Не варто плутати внутрішнє збагачення зі звичайним перетворенням даних, що виконуються в процесі їхнього завантаження в СД або під час підготовки до аналізу в аналітичному додатку.

Внутрішнє збагачення зазвичай пов'язане з отриманням і включенням у набір даних корисної інформації, яка відсутня в явному вигляді, але може бути певним способом отримана за допомогою маніпуляцій з наявними даними. Потім ця інформація вбудовується у вигляді нових полів або навіть таблиць у СД і

може бути використана для подальшого аналізу. Для збагачення даних також може використовуватися інформація, отримана в процесі їхнього аналізу.

Застосування збагачення даних із зовнішніх джерел зазвичай пов'язане зі збором інформації про об'єкти предметної сфери, що беруть участь у досліджуваному бізнес-процесі. Для підприємств і організацій це можуть бути економічні показники (прибуток, чисельність працівників, обсяг продажів та ін.).

Збагачення – один з найважливіших етапів підготовки даних до аналізу. Використання цієї процедури в багатьох випадках дозволяє підняти якість аналізу на принципово новий рівень, особливо під час вирішення нестандартних завдань, навіть в умовах недостатньої інформативності даних, що надходять з OLTP та облікових систем. Крім того, збагачення даних якоюсь мірою дозволяє компенсувати прорахунки у стратегії збору і консолідації аналітичних даних.

Запитання до параграфа 6

1. Що таке ETL?
2. Які основні цілі ETL?
3. Які основні завдання процесу ETL?
4. Які є три основні етапи процесу перенесення даних?
5. Яка узагальнена структура процесу ETL?
6. Які є потоки даних в ETL?
7. У чому сутність процесу вилучення даних в ETL?
8. Яка схема організації ETL?
9. Які особливості організації процесу вилучення даних?
10. Яке завдання очистки даних в ETL?
11. Які значно поширені проблеми, пов'язані з порушенням структури даних?
12. Які є типові помилки, відповідні структурним одиницям БД?
13. Що таке дублікати?
14. Що таке протиріччя?
15. У чому сутність перетворення даних в ETL?

16. Яка загальна схема процесу перетворення даних в ETL?
17. Що таке агрегування даних?
18. Що таке очищення даних?
19. Коли здійснюється перетворення даних?
20. Яка сутність багатопотокової організації процесу завантаження даних?
21. Що таке верифікація даних?
22. Якими можуть бути причини відмови від використання СД?
23. У чому сутність збагачення даних?
24. Які є зовнішні джерела даних?

§ 7. Трансформація даних

7.1. Введення у трансформацію даних

Крім попередньої обробки даних, процес підготовки даних до аналізу зазвичай включає в себе ще один етап – трансформацію.

Трансформація даних – комплекс методів і алгоритмів, спрямованих на оптимізацію уявлення та форматів даних з точки зору завдань, які необхідно вирішити, і цілей аналізу. Трансформація не ставить за мету змінити інформаційний зміст даних, її завдання – представити інформацію в такому вигляді, щоб вона могла бути використана найбільш ефективно.

Взагалі трансформація даних – це досить широке поняття, яке не має чітко окреслених меж. У різних напрямах обробки даних цей термін іноді поширюють на будь-які маніпуляції з даними незалежно від їхніх цілей і методів. Проте в контексті аналітичних технологій трансформація даних має цілком конкретну мету і завдання, а також використовує відносно стабільний набір методів. До основних з них належать нормалізація, перетворення типів і форматів, сортування, групування, злиття та інші.

На кожному етапі аналітичного процесу є свої пріоритетні цілі трансформації. Наприклад, у системах оперативної обробки (OLTP) це – забезпечення підтримки коректності форматів і типів

даних, оптимізація процесів доступу і вивантаження даних. На етапі ETL-процесу трансформація проводиться з метою приведення даних у відповідність з моделлю, яка використовується в сховищі, а також забезпечення процесу консолідації даних і їхнього завантаження в сховище. В аналітичному додатку проводиться безпосередня підготовка даних до аналізу, об'єднання і виділення найбільш цінної інформації, забезпечення коректної роботи аналітичних алгоритмів, методів і моделей.

Трансформація даних є одним з найважливіших процесів в аналізі. Від того, наскільки грамотно вона буде проведена, залежить ефективність аналізу, достовірність і точність його результатів.

Кожна вибірка вихідних даних, що завантажується в аналітичний додаток, характеризується набором властивостей, які можуть вплинути на ефективність роботи моделі і знизити достовірність результатів аналізу. Навіть якщо дані очищені від таких факторів, що погіршують їхню якість, як дублікати, протиріччя, шуми, аномія, патерналізм значення, пропуски та ін., вони все ще можуть не відповідати методиці та цілям аналізу. Це пов'язано не з вмістом даних, а з їх поданням і внутрішньою організацією. Може виникнути парадокс: дані абсолютно коректні з точки зору якості, а їхнього інформаційного змісту цілком достатньо для вирішення аналітичного завдання, але уявлення й організація даних ускладнюють аналіз або позагально роблять його неможливим. Дані можуть бути роз'єднані, не впорядковані, представлені у форматах, з якими не працює певний алгоритм. Трансформація даних, тобто їхнє перетворення до певного подання, формату або виду, оптимальному з точки зору розв'язуваної задачі, покликана вирішити цю проблему.

Тоді як очищення даних займається тільки їх змістом, трансформація – це процес оптимізації їхнього представлення та організації з точки зору визнаного методу аналізу. Трансформація даних залежить від завдань, алгоритмів і цілей аналізу. Тобто для одного завдання можуть виникнути за необхідності одні види трансформації, а для іншого – інші.

Невідповідність уявлення й організації даних методикою і цілям аналізу може проявлятися як на рівні моделей, так і на рівні окремих завдань аналізу. Невідповідність на рівні моделей означає,

що навіть якщо дані в іншому представленні придатні для використання в нейронній мережі, то неможливо, що їхнє використання з іншими моделями Data Mining, наприклад з деревами рішень, виявиться ускладнене або взагалі неможливе. Що стосується рівня завдань, то для прогнозування досить мати одновимірний ряд спостережень за прогнозованим параметром досліджуваного процесу або об'єкта, тоді як для вирішення завдання класифікації потрібна вибірка, що містить декілька ознак класифікації об'єктів. Якщо для вирішення завдання регресії вихідні дані повинні бути числовими, то під час класифікації – категоріальними. Водночас усі три завдання Data Mining: класифікація, регресія і прогнозування – можуть бути вирішені в рамках однієї моделі (наприклад, нейросітьовий).

Трансформація даних – досить широке поняття, яке не має чітко окреслених кордонів. У різних напрямах обробки даних термін «трансформація» іноді поширюють на будь-які маніпуляції з даними незалежно від їх цілей і методів. Однак у контексті бізнес-аналітики трансформація даних має цілком конкретну мету і завдання, а також використовує досить стабільний набір методів.

Трансформація даних – комплекс методів і алгоритмів, спрямованих на оптимізацію уявлення та форматів даних з точки зору вирішуваних завдань і цілей аналізу. Трансформація даних не ставить за мету змінити інформаційний зміст даних. Її завдання – подати цю інформацію в такому вигляді, щоб вона могла бути використана найбільш ефективно.

Трансформація даних не обов'язково пов'язана тільки з аналітичним прикладанням, у певному вигляді вона виконується на всіх етапах аналітичного процесу: в реєструючих системах (OLTP-системах), у процесі перенесення і завантаження даних у СД (ETL-процес) і безпосередньо під час підготовки даних до аналізу в аналітичному додатку. Таке розподілення процесу трансформації обусловлене тим, що на кожному етапі він переслідує різні цілі. У системах первинної реєстрації даних (OLTP-системах) трансформація може здійснюватися для забезпечення технічної та логічної сумісності даних, їхньої підготовки до вилучення, перенесення в сховище даних тощо. Наприклад, адреси часто вводять одним рядком. Водночас для аналізу можуть являти собою інтерес відслушні

компоненти адреси, які мають як текстовий формат (вулиця, місто), так і числовий (номер будинку, офісу). За допомогою трансформації можна розподілити відповідні елементи за окремими полями і перетворити їх у потрібний формат. Наведемо ще один приклад. Різні системи реєстрації (сканер штрих-кодів, контрольно-касовий апарат тощо). Можуть видавати однотипні дані в різних форматах. Так, в одному випадку як роздільник цілої та дробової частин чисел може користуватися крапка, а в іншому – кома. Спроба вставити такі дані в одне поле бази даних може призвести до серйозних технічних проблем, а трансформація, тобто перетворення до єдиного формату, допоможе вирішити це питання.

Основними цілями трансформації даних на етапі процесу є приведення їх у відповідність з моделлю даних, використовуваної в сховище, здійснення коректної консолідації і, власне, завантаження в сховище.

Виникає питання: якщо трансформації даних так багато уваги приділяється на етапі збору та організації зберігання даних, то навіщо включати кошти трансформації в аналітичний додаток (що, безумовно, ускладнює і робить більш коштовним його розробку)? Відповідь проста. Не всі дані надходять в аналітичний додаток із систем, де вони пройшли попередню підготовку, таких як OLTP-системи і сховища даних. Але основна причина полягає в тому, що трансформація даних у цих системах здебільшого носить технічний характер і слабо пов’язана з можливими методами, алгоритмами і цілями аналізу.

Основні відмінності застосування трансформації даних на різних етапах аналітичного процесу представлені в табл. 2.7.

Хоча і OLTP-системи, і сховища даних у багатьох випадках розробляються з урахуванням перспективних напрямів аналізу, передбачити всі можливі ситуації досить важко. В умовах динаміки сучасного бізнесу не можна передбачити, який набір аналітичних рішень стане необхідним у найближчому майбутньому. Саме процес підготовки даних в аналітичному додатку безпосередньо перед виконанням аналізу, коли його цілі і методи чітки, дозволяє аналітику найкращим чином підготувати дані, виділити і підкреслити в них необхідну інформацію для найбільш ефективного її використання. Це завдання і вирішує трансформація даних.

Таблиця 2.7

Трансформація даних на різних етапах аналітичного процесу

OLTP-система	ETL-процес	Аналітичний додаток
Забезпечення підтримки коректності форматів і типів даних. Оптимізація процесів доступу і вивантаження даних	Забезпечення підтримки коректності форматів і типів даних. Перетворення даних з метою приведення у відповідність з моделлю даних, яка використовується в сховище. Забезпечення процесу консолідації даних і їхнього завантаження в сховище	Безпосередня підготовка даних до аналізу, об'єднання і виділення найбільш цінної інформації, забезпечення коректної роботи аналітичних алгоритмів, методів і моделей

Необхідність використання методів трансформації даних в аналітичному додатку обумовлена ще й їхньою доступністю аналітику. Дійсно, в процесі підготовки даних користувач аналітичного додатку має можливість на свій розсуд застосовувати певні методи трансформації даних, налаштовуючи їхні параметри, проводити експерименти, що дозволяють визначити вплив трансформації на результати аналізу. Що стосується сховищ даних і процесів, то, як правило, виконуються в них перетворення даних досить чітко, їхня адаптація, залежно від характеру даних і вирішуваних завдань аналізу, не передбачена, й аналітик не може втручатися в ці процеси.

Системи оперативної реєстрації даних (OLTP) зазвичай взагалі недоступні аналітику.

Основні методи трансформації даних

Перш ніж приступити до докладного розгляду методів, алгоритмів і цілей окремих напрямів трансформації даних, проведемо короткий огляд типових коштів трансформації, якими оснащується більшість аналітичних платформ.

1. Перетворення упорядкованих даних. Дозволяє оптимізувати представлення таких даних з метою забезпечення подальшого аналізу, наприклад рішення задачі прогнозування часового ряду або групування з часового періоду.

2. Квантування. Дозволяє розбити діапазон можливих значень числової ознаки на задану кількість інтервалів і привласнити номери інтервалів або інші мітки потрапили в їхнє значення.

3. Сортування. Дозволяє змінити порядок проходження записів вихідної вибірки даних, відповідно до алгоритму, певним користувачем. У деяких випадках сортування дає можливість спростити візуальний аналіз вибірки, оперативно визначити найбільші і найменші значення ознак тощо.

4. Злиття. Дозволяє об'єднати дві таблиці за однойменними полями або доповнити одну таблицю записами з іншої, які відсутні в додатковій. Злиття застосовується в тих випадках, коли інформацію в аналізованій вибірці даних необхідно доповнити інформацією з іншої вибірки. Під час об'єднання до записів вихідної вибірки додаються всі записи іншій. У разі доповнення до вихідної вибірки додаються тільки ті дані, які були відсутні у вихідній. Операція злиття є одним із способів збагачення даних: якщо вибірка містить недостатньо даних для аналізу, то її можна доповнити відсутньою інформацією з іншої вибірки.

5. Групування і розгрупування. Досить часто інформація, яка цікавить аналітиків, у таблиці виявляється «розбавлена» сторонніми даними, роз'єднана, розкидана окремими полями і записами. Використовуючи групування, можна узагальнити потрібну інформацію, об'єднати її в мінімально необхідну кількість полів і значень. Зазвичай передбачають можливість виконання і зворотної операції – розгрупування.

6. Налаштування набору даних. Дозволяє змінювати імена, типи, мітки і призначення полів вихідної вибірки даних. Наприклад, якщо поле, що містить числову інформацію, в джерелі даних за будь-якої причини має строковий тип, значення цього поля не можуть оброблятися як числа. Щоб робота з числовими даними цього поля стала можливою, їх варто перетворити до числового типу.

7. Таблична підстановка значень. Дозволяє робити заміну значень у вихідній вибірці даних на основі так званої таблиці підстановки. Таблиця підстановки містить пари «початкове значення – нове значення». Кожне значення вибірки даних перевіряється на відповідність вихідного значення таблиці підстановки, і якщо

таку відповідність знайдено, то значення вибору змінюється на відповідне нове значення з таблиці підстановки. Це досить зручний спосіб для автоматичного коректування значень.

8. Обчислювані значення. Іноді для аналізу потрібна інформація, яка відсутня в явному вигляді у вихідних даних, але може бути отримана на основі розрахунків наявних значень. Для цих цілей в аналітичний додаток включається своєрідний калькулятор, який дозволяє виконувати над даними вихідної вибірки різні обчислення. Оскільки аналізовані дані можуть бути різних типів (строковий, числовий, дата / час, логічний), то механізм розрахунків повинен підтримувати роботу не тільки з числовими даними, але і з даними інших типів, наприклад виділяти підстроку, виконувати логічні операції тощо.

9. Нормалізація. Нормалізація дозволяє перетворити діапазон зміни значень числової ознаки в інший діапазон, більш зручний для застосування до даних певних аналітичних алгоритмів, а також узгодити діапазони заради різних ознак. Часто використовується приведення до одиниці, коли весь наявний діапазон даних «стискається» в інтервал [0; 1] або [–1; 1]. Особливо важливо здійснити правильну нормалізацію даних в алгоритмах Data Mining, які засновані на вимірі відстані між векторами об’єктів у багатовимірному просторі ознак (наприклад, у кластеризації).

10. Трансформація упорядкованих даних. Багато аналітичних завдань, пов’язані з обробкою даних, що залежать від часу. Такі дані називають упорядкованими, або часовими рядами. У процесі обробки часових рядів потрібна спеціальна підготовка даних, щоб оптимізувати їх подання для вирішення певних аналітичних завдань: прогнозування, класифікації станів об’єктів, виявлення закономірностей, що пояснюють динаміку бізнес-процесів тощо – для всіх можливих інтервалів дати і часу.

Часовий ряд складається з послідовності спостережень за станом параметрів (ознак) досліджуваних об’єктів або процесів. Якщо спостереження містять одну ознаку, то ряд є одновимірним, а якщо дві або більше – багатовимірним. Оскільки значення часового ряду визначені тільки у фіксовані моменти часу, так звані відліки, послідовність його значень може бути представлена в такому вигляді: $X = \{x(1), x(2)...x\{n)\}$, де $x\ (n)$ – останнє значення

розглянутої часової послідовності. При цьому відлік часу покладається рівновіддаленим один від одного.

Часовий ряд завжди є впорядкованим у часі даних. Однак зворотне не завжди правильно: не всі впорядковані дані можуть утворювати тимчасової ряд. Наприклад, під час зняття показань зі свердловини вимірювання характеристик виробляються з деяким кроком глибини. Це теж впорядковані дані, але їх не можна назвати тимчасовими поруч, оскільки фіксується не час, а глибина. Проте багато міркування про тимчасові ряди поширюється на будь-які впорядковані дані.

Метою трансформації часових рядів є не зміна їхнього змісту, а подання інформації таким чином, щоб забезпечувалася максимальна ефективність вирішення певної задачі аналізу. Можна виділити два основних типи перетворення, які найбільш часто використовуються у разі підготовки часових рядів до аналізу.

Ковзне вікно застосовується під час вирішення задач прогнозування та класифікації станів бізнес-об'єктів, щоб перетворювати послідовність значень ряду в таблицю, яку можна використовувати для побудови моделей або будь-якої іншої обробки.

Перетворення дати і часу полягає у приведенні дати і часу до виду, найбільш зручного для візуального аналізу й обробки часового ряду. При цьому результати перетворення дати вже не є значеннями типу Дата / Час і можуть оброблятися як звичайні числа і рядки. Розглянемо ці перетворення більш докладно.

Ковзне вікно. Широко застосовується у процесі обробки часових рядів, наприклад, щоб побудувати модель прогнозу часового ряду. Метою прогнозування значень часового ряду є передбачення значення ознаки на основі попередніх її значень. Рішення завдання прогнозування можливо тільки в тому випадку, якщо значення часового ряду пов'язані між собою.

Для побудови прогностичної моделі необхідно знати три параметри. Інтервал прогнозу – часовий інтервал, на якому здійснюватиметься прогнозування: день, тиждень, місяць, квартал або рік. Горизонт прогнозу – на яку кількість інтервалів (днів, тижнів тощо) бажаємо отримати прогноз. Глибина занурення – кількість значень інтервалів прогнозу в минулому, яке будемо використовувати для передбачення значень інтервалів у майбутньому.

У разі вибору інтервалу прогнозу, можливо, стане необхідно агрегування даних всередині інтервалу. Наприклад, якщо в базі даних інформація про продажі представлена за днями, то для побудови прогнозу за тижнями доведеться об'єднати інформацію за окремі дні, обчисливши суму, середнє значення або використовуючи іншу функцію агрегації.

Якщо як інтервал прогнозу обрано тиждень, то глибина занурення – кількість тижнів у минулому, значення продажів за які будемо використовувати як вихідні дані для передбачення. Горизонт прогнозу в цьому випадку – кількість тижнів, значення продажів за які хочемо передбачити.

У процесі вибору глибини занурення і горизонту прогнозу варто керуватися такими міркуваннями. Модель прогнозу функціонує за принципом узагальнення, тобто висновок про можливе значення в майбутньому робиться на основі аналізу великої кількості значень з минулого. Отже, глибина занурення повинна в кілька разів перевищувати горизонт прогнозу.

Чим більша кількість значень з минулого використовується для прогнозування, тим вище ступінь узагальнення і вірогідніші результати передбачення. Проте при цьому є два обмеження. По-перше, глибина занурення повинна бути такою, щоб значення з минулого, які використовуються для прогнозування, залишалися актуальними, тобто правильно відображали поточну поведінку процесу. Використання для побудови моделі прогнозу занадто старих даних, які втратили актуальність, призведе до зниження вірогідності або до отримання некоректних результатів. По-друге, занадто велика глибина занурення збільшить розмірність вибірки даних, отриманої в результаті обробки ряду ковзаючим вікном, що спричинить за собою ускладнення моделі. Крім того, зростуть часові й обчислювальні витрати на аналіз даних. Чим далі простягається горизонт прогнозу, тим нижче вірогідність результатів.

Перетворення дати і часу. Зазвичай кожен елемент часового ряду являє собою відповідність «дата – значення», а в разі багатовимірного ряду – «дата – значення 1, значення 2 ...». Усі методи обробки, розглянуті раніше, були спрямовані на перетворення

значень, самі ж дати зазвичай залишалися без змін. Однак, як показує практика, під час вирішення різних завдань аналізу може бути корисним і перетворення дати.

Найчастіше дати у вихідних даних представлені в певному форматі, зазвичай «ЧЧ.ММ.РР», тобто число, місяць і рік, при цьому на відображення кожного елемента дати відводяться дві цифри. Можливі деякі варіації: вказується повністю рік (02.10.2006) або (і) місяць (12 вересня 2006) та ін. З точки зору аналізу найбільший інтерес являють собою не деталізовані дані за кожен день, а агреговані за тижнями, місяцями, кварталами тощо. У зв'язку з цим може виявитися корисним вилучення з дати додаткової інформації про часові інтервали, на яких проводиться аналіз.

Так, якщо базовий інтервал, за яким здійснюється аналіз діяльності підприємства, тиждень, то в деяких випадках може являти собою інтерес відображення дати не на кожен день, а на перший або останній день тижня.

Приведення дати до різних видів представлення дозволяє спростити візуальний аналіз вихідних даних, що залежать від часу, а також використати значення часу як вихідні дані для вирішення різноманітних аналітичних задач.

Можуть бути цікаві дані, отримані для різних інтервалів часу в межах однієї дати, тобто представлені в масштабі годин, хвилин і секунд. Для трансформації часу параметри години, хвилини і секунди вилучаються в окремі поля. Результати такого перетворення показані в табл. 2.8.

Таблиця 2.8

Час	**Час, год**	**Час, хв**	**Час, с**
12:10:39	12	10	39
13:15:20	13	15	20
16:20:44	16	20	44
18:09:56	18	9	56
19:15:30	19	15	30
21:22:50	21	22	50
23:12:40	23	12	40

7.2. Групування даних

Іноді високий ступінь деталізації даних може погіршити результати аналізу. Наслідком її є занадто велика кількість записів у результатній вибірці даних, а також низька значимість кожного окремого значення. Крім того, інформація виявляється «розмазаною» за всією вибіркою і її важко використовувати для аналізу. Щоденні показники деякого бізнес-процесу, наприклад продажів, не завжди стабільні під дією різних випадкових факторів, тому значення продажів, взяті окремо за кожен день, навряд чи дадуть надійну оцінку, яка характеризує розвиток процесу. Але якщо згрупувати щоденні показники за тиждень або місяць, то така оцінка буде більш обґрунтованою. Те ж саме можна зазначити і про ознаки об'єктів. Якщо об'єкти подібні між собою, то будуть близькі і їхні однойменні ознаки. Однак значення ознак також можуть бути схильні до випадкових варіацій. Тому якщо об'єднати ознаки за групами подібних об'єктів, то можна отримати показник, який більш надійно характеризує властивості сукупності в цілому.

Отже, групування даних – корисний інструмент, застосування якого в процесі підготовки даних до аналізу дозволяє більш ефективно використовувати інформацію.

Перед тим, як приступати до групування, необхідно визначити у вихідній вибірці поля вимірювань і фактів. Вимірювання – це дані, які характеризують досліджуваний процес якісно, а факти – кількісно. Наприклад, процес продажів описується кількісними і якісними даними. Якісні дані – найменування товарів і клієнтів. Очевидно, що просто вказати ці параметри недостатньо. З кожним проданим товаром або клієнтом пов'язаний набір числових показників – фактів: ціни, кількості, суми, знижки, націнки тощо.

Кожне найменування товару або клієнта, що міститься в полі вимірювання, називається значенням виміру. Сутність групування полягає в тому, що всі записи, що містять однойменні значення вимірювання, об'єднуються в одну, а відповідні факти агрегуються.

Розглянемо вибірку даних, що відображають продаж будматеріалів (табл. 2.9). Вона містить вимірювання Дата, Клієнт і Товар, а також факти Ціна, Кількість і Сума.

Таблица 2.9

Продажі будматеріалів

Дата	Клієнт	Товар	Ціна	Кіль-кість	Сума
01.03.2007	ООО «Полігон»	Цемент	150	20	3 000
01.03.2007	ЗАО «Монтажник»	Керамзит	100	60	6 000
0l.03.2007	ООО «Тандем»	Цегла	1 500	5	7 500
02.03.2007	ООО «Шплінт»	Плити	1 100	10	11 000
02.03.2007	ЗАО «Монтажник»	Блоки	900	20	18 000
02.03.2007	ООО «Полігон»	Цегла	1 500	10	15 000
03.03.2007	ООО «Агрострой»	Плити	1 100	8	8 800
03.03.2007	ЗАО «Монтажник»	Керамзит	100	30	3 000
03.03.2007	ООО «Тандем»	Блоки	900	10	9 000
03.03.2007	ООО «Полігон»	Плити	1 100	30	33 000
04.03.2007	ООО «Шплінт»	Керамзит	100	100	10 000
04.03.2007	ООО «Тандем»	Цегла	1 500	10	15 000
04.03.2007	ЗАО «Монтажник»	Цемент	150	20	3 000
04.03.2007	ООО «Полігон»	Блоки	900	15	13 500
05.03.2007	ООО «Агрострой»	Плити	1 100	20	22 000
05.03.2007	ООО «Шплінт»	Керамзит	100	50	5 000
05.03.2007	ЗАО «Монтажник»	Цемент	150	40	6 000
05.03.2007	ООО «Тандем»	Блоки	900	15	13 500
06.03.2007	ООО «Полігон»	Цегла	1 500	6	9 000
06.03.2007	ООО «Агрострой»	Керамзит	100	70	7 000
06.03.2007	ЗАО «Монтажник»	Цемент	150	30	4 500

Групування може проводитися за одним з трьох фактів. Якщо як факт для групування вибирається Дата, то всі записи, що містять однакові дати, будуть об’єднані в одну (табл. 2.10).

Таблиця 2.10

Варіант групування 1

Дата	Ціна	Кількість	Сума
01.03.2007	583,33	85	16 500
02.03.2007	1 166,67	40	44 000
03.03.2007	800,00	78	53 800
04.03.2007	662,50	145	41 500
05.03.2007	562,50	125	46 500
06.03.2007	583,33	106	20 500

При цьому для вимірювання Ціна була обрана функція агрегації з обчисленням середнього значення, для вимірювань Кількість і Сума – з обчисленням суми. Таким чином, цей варіант групування дозволяє отримати інформацію про те, яка кількість одиниць товару, за якою середньою ціною і на яку суму була продана в межах кожної дати.

Групування щодо вимірювання Клієнт дозволяє отримати ті ж факти, що і в попередньому прикладі, але вже не для кожної дати, а для кожного клієнта (табл. 2.11).

Таблиця 2.11

Варіант групування 2

Дата	Ціна	Кількість	Сума
ЗАО «Монтажник»	258,33	200	40 500
ООО «Агрострой»	766,67	98	37 800
ООО «Полігон»	1 030,00	81	73 500

Ціна	Кількість	Сума
1 200,00	40	45 000
433,33	160	26 000

Отримана в результаті групування інформація може використовуватися для різних завдань. Так, розподіл сум продажів за датами дозволить проаналізувати динаміку продажів, виявити тенденції тощо. Групування за полем Клієнт – оптимізувати роботу з клієнтами, наприклад надати найбільш активним з них знижки. Групування за полем Товар – визначити найбільш і найменш продажні товари, оцінити внесок конкретного товару в загальний обсяг продажів.

До найбільш типових функцій агрегування належать: сума, середнє, кількість, максимум, мінімум, медіана.

Для строкових значень як функції агрегації можуть використовуватися тільки максимум, мінімум, кількість, перший, останній.

7.3. Розгруповування даних

У процесі групування даних відбувається об'єднання фактів за значенням деякого вимірювання із застосуванням певної функції агрегації, але при цьому втрачається інформація, пов'язана з іншими вимірами, тобто зв'язок фактів зі значеннями інших вимірів. І якщо у нас є прогноз, побудований на основі згрупованих даних, то за ним нічого не можемо зазначити про значення, на основі яких він був отриманий. За необхідності цю інформацію можна відновити за допомогою розгрупування даних. Це не є операцією, яка зворотня групуванню. Факти відновлюються не точно, а пропорційно їх відомому вкладу під час агрегування. Загалом, розгруповування не може бути ніяк пов'язане з групуванням.

Розгруповування може виконуватися для всіх видів обробки даних, коли має місце перетворення деякого набору значень в одне значення за допомогою будь-якого алгоритму, наприклад, коли на основі декількох членів ряду формується прогнозоване значення. При цьому отримані значення відображатимуть частку кожного вихідного значення в результаті.

7.4. Злиття даних

У практиці аналізу досить часто зустрічається ситуація, коли необхідні дані доводиться збирати з декількох таблиць. Потреба у цьому зазвичай виникає в таких випадках.

Дані, які потрібні для аналізу, «розкидані» за декількома таблицями.

Дані в початковій таблиці несуть недостатньо інформації для якісного аналізу, і тому потрібна процедура їхнього збагачення, яка зазвичай пов'язана з додаванням у таблицю даних зі сторонніх джерел.

Ситуація, коли аналізовані дані виявляються в декількох таблицях або беруться з окремих джерел, а не з централізованого сховища даних, може бути наслідком непродуманого процесу консолідації та інтеграції даних. Недостатньо інформативна вибірка також зустрічається досить часто. Отже, якщо для вирішення однієї аналітичної задачі інформативність множини даних може

бути достатня або навіть надлишкова, то для іншої вона виявиться недостатньою і вимагатиме збагачення.

За необхідності об'єднання даних виконується процедура злиття. Таблиця, до якої у процесі злиття додаються дані з іншої, називається вихідною. Вихідна і яку пов'язують таблиці повинні мати одне або декілька однакових полів, на основі яких буде проводитися пов'язування двох таблиць – це поля зв'язку. Решта полів, які унікальні для кожної з таблиць, можуть бути приєднані до результуючого набору після злиття.

Існує кілька способів злиття, які застосовуються залежно від того, які дані і в якому вигляді повинні бути об'єднані в результуючій таблиці.

Об'єднання. Об'єднання застосовується в тих випадках, коли до рядків вихідної таблиці потрібно додати всі рядки таблиці, яку пов'язують, при цьому рядки додаються знизу.

Внутрішнє з'єднання. Внутрішнє з'єднання дозволяє отримати в результуючому наборі тільки ті записи, для яких значення в одному з полів зв'язку збігаються. Тобто в таблиці, отриманої в результаті внутрішнього сполучення, залишаються тільки ті записи, які містять однакові значення в заданому полі (або заданих полях).

Зовнішнє з'єднання. Під час зовнішнього з'єднання всі записи однієї таблиці доповнюються значеннями з іншої, якщо значення цих записів за ключовими полями збігаються. Тобто таблиці пов'язуються за полем, і якщо існують записи, де значення цього поля в обох таблицях ідентичні, то записи будуть доповненні значеннями, які відсутні в одній таблиці і присутні в іншій. Фактично цей механізм дозволяє додавати поля з однієї таблиці в іншу, але не за всіма записами, а тільки за тими, значення яких у полі зв'язку збігаються для обох таблиць. Розрізняють ліве і праве зовнішнє з'єднання. При лівому записи вихідної таблиці доповнюються значеннями з пов'язаної таблиці, а при правому – навпаки.

Повне зовнішнє з'єднання. У результуючий набір включаються всі рядки і поля як вихідної, так і яку пов'язують таблиць. При цьому якщо до деяких записів значення поля зв'язку для обох таблиць збігаються, то всі поля цього запису заповнюються відповідними значеннями.

Якщо збіг за полем зв'язку відсутній, то інші поля такого запису будуть заповнені порожніми значеннями. Фактично це означає, що для значень вихідної таблиці відсутні відповідні значення в таблиці, яку пов'язують.

Квантування. Одним з перетворень, які часто використовуються під час підготовки даних до аналізу, є квантування. В основі операції квантування полягає процедура, що складається з двох кроків:

1. Діапазон значень, у межах якого змінюється деяка числова величина (ознака, показник тощо), розбивається на задану кількість інтервалів, кожному з яких присвоюється певний номер. Ці інтервали називаються інтервалами квантування, а присвоєні їм номери – рівнями квантування.

2. Кожне значення замінюється номером інтервалу квантування, в який потрапило це значення.

За необхідності замість номера інтервалу квантування використовуються інші значення, пов'язані з цим інтервалом. Це можуть бути верхня або нижня межа інтервалу, його середнє значення або довільна мітка, обрана користувачем.

У разі квантування необхідно визначити, яку з меж інтервалу варто включити в цей інтервал. Оскільки нижня межа діапазону завжди належить нижньому інтервалу, то і для інших інтервалів є сенс включити нижню межу. Єдиним винятком є верхній інтервал, який включає в себе як верхню, так і нижню межу.

У бізнес-аналітиці квантування сприяє досягненню таких цілей.

Змінюється вигляд даних: з безперервних вони можуть бути перетворені в дискретні.

Квантування може використовуватися для скорочення розмірності даних, а саме для зменшення кількості різноманітних значень ознаки. Наприклад, якщо для аналізу клієнтів банку, які отримують кредит, інтерес являють собою не окремі клієнти і суми кредитів, а групи, які об'єднують клієнтів. У деяких випадках становить інтерес використання не самих номерів інтервалів, а інших значень, пов'язаних з ними. Наприклад, нижня межа інтервалу, верхня межа інтервалу, середина інтервалу тощо. Мітку інтервалу користувач може задати довільно.

Використання міток інтервалів дає можливість зробити результати квантування більш наочними і відразу визначити мітки класів, якщо метою квантування є розбиття ознаки за категоріями.

Вибір кількості інтервалів квантування. У квантуванні важливо правильно вибрати кількість інтервалів. У результаті квантування здійснюється перехід від точних даних до деяких інтервальних оцінок, і при цьому неминуча втрата інформації. Фактично ряд значень, отриманих у результаті квантування, просто висловлює відносини між вихідними значеннями ознаки. Те, що два значення розташовані в двох сусідніх інтервалах квантування, не дозволяє точно визначити, наскільки одне з них більше або менше іншого. Потрібно зазначити тільки, що вони не відрізняються більше ніж на дві ширини інтервалу. Отже, чим більше інтервалів використовується під час квантування, тим точніше уявлення про вихідні значення даних. У разі зменшення ширини інтервалу в межі отримаємо вихідний набір значень.

Іншими словами, змінюючи кількість інтервалів квантування, можна перейти від точного відтворення вихідних значень даних до повної втрати інформації про мінливість значень ознаки. На практиці вибрати кількість інтервалів квантування можна виходячи з таких міркувань.

Якщо квантування виконується для перетворення безперервних даних у дискретні, то кількість інтервалів визначатиметься кількістю унікальних значень (міток, категорій), які використовуються під час вирішення завдання аналізу.

Необхідно враховувати потрібну точність опису даних. Наприклад, може бути поставлена умова, що кількість інтервалів квантування має бути такою, щоб ширина інтервалу не перевищувала 10 % від повного діапазону зміни вихідних значень.

Методи квантування. Крім вибору кількості інтервалів, під час виконання операції квантування потрібно вибрати метод. Вибір методу квантування залежить від характеру даних. Розрізняють два основні методи квантування:

– рівномірне (однорідне) квантування;

– нерівномірне (неоднорідне) квантування.

У процесі рівномірного квантування діапазон зміни значень ознаки поділяється на інтервали однакової ширини, а за нерівномірним – ширина інтервалів може бути різною.

Перший метод використовується, якщо дані рівномірно розподілені по всьому діапазону їхньої зміни, тобто в результаті квантування не буде інтервалів, в яких значення майже відсутні або заповнені досить щільно. В іншому випадку кращі результати дасть другий метод.

Нормалізація і кодування даних

Алгоритми Data Mining працюють тільки з числами, оскільки їхня робота базується на математичних операціях. Тим часом не будь-яка вхідна або вихідна змінна спочатку представлена у вигляді числа. Перш ніж подавати такі змінні на вхід алгоритму, їх потрібно перетворити. Цей процес називається кодуванням.

Як входами, так і виходами алгоритмів можуть бути абсолютно різні величини. Очевидно, що результати моделювання не повинні залежати від одиниць виміру цих величин. Щоб аналітичний метод трактував їхнє значення однаково, всі вхідні і вихідні величини повинні бути приведені до єдиного масштабу. Такий процес називається нормалізацією (нормування). Вона полягає в перетворенні вихідного діапазону, в межах якого розподілені значення певної ознаки, до іншого, зазвичай меншого діапазону, наприклад [–1; 1] або [0; 1].

Далеко не всі алгоритми вимагають нормалізації. Наприклад, дерево рішень або лінійна регресія не потребують приведення входів і виходів до єдиного масштабу, а ось у нейронній мережі без цієї процедури не обійтися. У більшості випадків нормалізація і кодування «зашиті» всередині алгоритмів, проте аналітик повинен розбиратися в їх різновидах, оскільки вибір схеми кодування і нормалізації істотно впливає на результат.

Кодування категоріальних даних. Під час використання таких моделей Data Mining, як нейронні мережі, дерева рішень, регресія, дані повинні подаватися на вхід у числовому вигляді. Однак часто як вихідні дані для аналізу використовуються категорії. У цьому випадку може виникнути потреба застосування до них перетворення, аналогічного нормалізації для числових ознак, тобто приведення категоріальних значень до числових.

Запитання до параграфа 7

1. Що таке трансформація даних?
2. Яка відмінність трансформації від очищення?
3. Які є основні методи трансформації даних?
4. Що є метою трансформації часових рядів?
5. У чому сутність перетворення дати і часу?
6. Що таке групування даних?
7. У чому сутність розгрупування даних?
8. Що таке злиття даних?
9. Які є способи злиття даних?
10. Що таке квантування?
11. Яка мета квантування?
12. Які є методи квантування?
13. Що таке нормалізація і кодування даних?

§ 8. Візуалізація даних

8.1. Вступ у візуалізацію даних

Візуалізація даних – це наочне представлення масивів різної інформації. Існує декілька типів візуалізації[22]:

1. Звичайне візуальне представлення кількісної інформації у схематичній формі. До цієї групи можна віднести: кругові та лінійні діаграми, гістограми і спектрограми, таблиці і різні точкові графіки тощо.

2. Дані під час візуалізації можуть бути перетворені у форму, що покращує сприйняття й аналіз цієї інформації. Наприклад, карта і полярний графік, тимчасова лінія і графік з паралельними осями, діаграма Ейлера тощо.

3. Концептуальна візуалізація дозволяє розробляти складні концепції, ідеї і плани за допомогою концептуальних карт, діаграм Ганта, графів з мінімальним шляхом та інших подібних видів діаграм.

---

[22] Як і для чого використовувати візуалізацію даних? Центр політичних студій та аналітики «Ейдос». URL : http://eidos.org.ua/novyny/yak-i-dlya-choho-vykorystovuvaty-vizualizatsiyu-danyh/

4. Стратегічна візуалізація переводить у візуальну форму різні дані про аспекти роботи організацій. Це будь-які діаграми продуктивності, життєвого циклу і графіки структур організацій.

5. Графічно організувати структурну інформацію за допомогою пірамід, дерев і мап даних допоможе метафорична візуалізація, яскравим прикладом якої є мапа метро.

6. Комбінована візуалізація дозволяє об'єднати декілька складних графіків в одну схему, як у мапі з прогнозом погоди.

Візуальна інформація краще сприймається. Фізіологічно сприйняття візуальної інформації є основною для людини. Є численні дослідження, які підтверджують, що:

– 90 % інформації людина сприймає через зір;

– 70 % сенсорних рецепторів містяться в очах;

– близько половини нейронів головного мозку людини задіяні в обробці візуальної інформації;

– на 19 % менше під час роботи з візуальними даними використовується когнітивна функція мозку, що відповідає за обробку та аналіз інформації;

– на 17 % вище продуктивність людини, що працює з візуальною інформацією;

– на 4,5 % краще згадуються докладні деталі візуальної інформації;

– в 60 000 разів швидше сприймається візуальна інформація порівняно з текстовою.

Крім обробки нашим мозком, візуалізація даних має декілька переваг:

– акцентування уваги на різних аспектах даних;

– аналіз великого набору даних зі складною структурою;

– зменшення інформаційного перевантаження людини й утримування її уваги;

– однозначність і чіткість виведених даних;

– виділення взаємозв'язків і відносин, що містяться в інформації.

Візуалізація має критичне значення для аналізу даних. Вона як «перша лінія атаки», яка дозволяє розкривати складну внутрішню структуру даних, яку неможливо усвідомити в інший спосіб.

Дані самі по собі – це здебільшого послідовність бітів та байтів, що зберігається у файлі на жорсткому диску комп'ютера чи в інший спосіб, і вони є невидимими. Щоб побачити їх та зрозуміти їхнє значення, потрібно візуалізувати ці дані. Тому однією з важливих складових бізнес-аналітики є візуалізація даних – подання їх у вигляді, який забезпечує ефективну роботу аналітика, полегшує їхнє сприйняття, дозволяє отримувати нову інформацію. Спосіб візуалізації має максимально відображати поведінку і специфіку даних, інформацію, яка в них міститься, тенденції, закономірності тощо. Вибір способу візуалізації залежить від характеру досліджуваних даних, а також від мети і завдання аналізу.

Візуалізація пов'язана не тільки з інтерпретацією інформації, оцінкою якості та достовірності результатів аналізу. Її рекомендується застосовувати на всіх етапах аналітичного процесу.

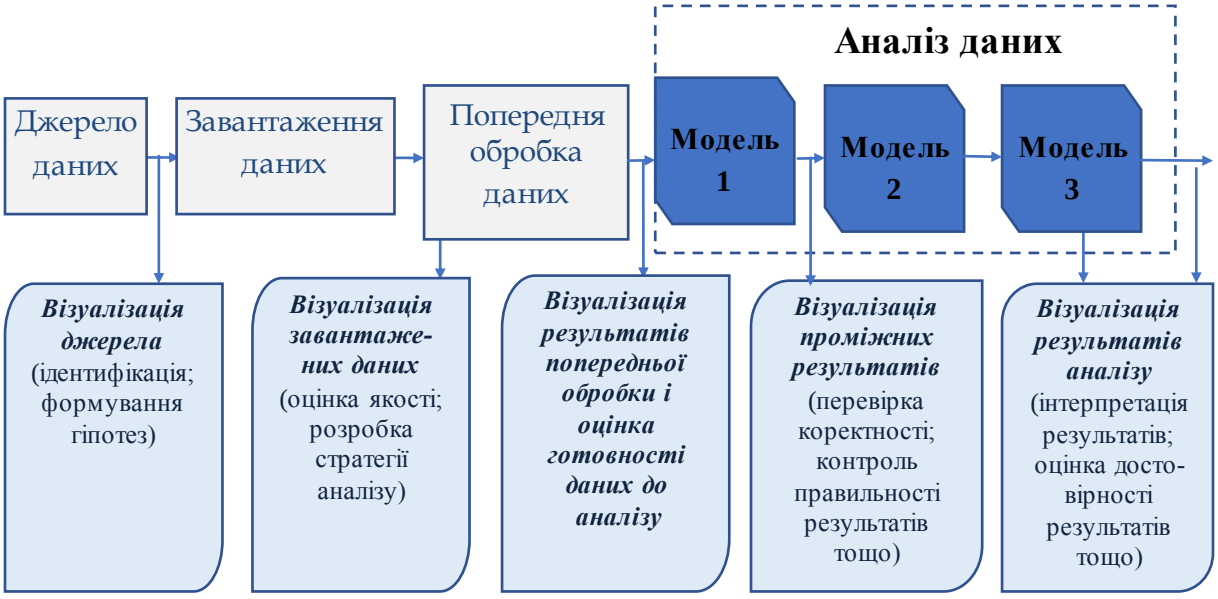


Рис. 2.25. Цілі і завдання візуалізації на різних етапах аналітичного процесу

Кожна нова візуалізація імовірно дасть будь-які ідеї щодо даних, методів їх обробки тощо. Деякі з цих ідей можуть бути вже відомими (але, можливо, наразі не доведеними), тоді як інші можуть бути цілком новими або навіть несподіваними.

Візуалізація використовується на різних етапах аналітичного процесу для досягнення різних цілей (відповідно до етапу, на якому проводиться візуалізація) і вирішення відповідних завдань (рис. 2.25).

1. Візуалізація джерел даних. У джерелі даних перед їх завантаженням в аналітичну систему потрібно візуально оцінити:

– характер, тип і поведінку даних;

– динамічний діапазон значень;

– ступінь гладкості;

– наявність факторів, що знижують якість даних, таких як шуми, аномальні і пропущені значення.

Візуальний аналіз джерела даних дозволяє:

– побачити, чи відповідають дані очікуванням;

– оцінити ступінь придатності даних до аналізу;

– висунути гіпотези про закономірності процесів, що описуються даними;

– визначити, які види очищення і попередньої обробки необхідно застосувати до даних.

Візуалізація джерел даних дозволяє визначити метод завантаження даних в аналітичний додаток і параметри, які при цьому повинні бути використані.

Важливо правильно налаштовувати параметри завантаження під час імпорту великих масивів даних. Якщо масив даних має великий обсяг, то процес завантаження в аналітичний додаток може виявитися досить тривалим, а після довгого очікування може з'ясуватися, що дані були завантажені некоректно, оскільки було неправильно вказано деякі параметри.

2. Візуалізація завантажених даних. Після завантаження даних з джерела в аналітичний додаток робота з вибіркою починається з візуального аналізу. Проте цілі, завдання і методи візуального аналізу тепер будуть іншими. Потрібно переконатися, що дані завантажилися правильно (не з'явилися пропуски, зберіглася структура рядків і стовпців тощо). Спотворення даних під час завантаження може статися через невідповідність їхніх типів налаштуванням параметрів завантаження. Якщо дані завантажені коректно, то, як правило, далі рекомендується оцінити ступінь їхньої гладкості, наявність шумів і аномальних викидів. Особливу увагу варто звернути на пошук фрагментів даних з певними особливостями.

3. Корисно отримати попередні статистичні характеристики даних, а саме: мінімальне і максимальне значення, дисперсію і середньоквадратичне відхилення тощо. Після цього здійснюється візуалізація результатів попередньої обробки даних. За результатами візуального аналізу здійснюються висновки про доцільність (необхідність і можливість) застосування певних видів очищення та трансформації даних, виробляється методика і стратегія їхнього аналізу.

4. Візуалізація даних у процесі їхньої аналітичної обробки. Як правило, аналітичні процедури є багатокроковими, тобто в процесі аналізу до даних послідовно застосовується кілька алгоритмів або моделей. Вибірка, яка формується на виході кожного алгоритму або моделі, може подаватися на вхід наступного етапу обробки. Очевидно, що якщо дані, які надійшли з попереднього етапу, виявляться некоректними, то подальша обробка втрачає сенс. Тому значно важливо передбачити візуалізацію проміжних результатів аналізу з метою перевірки коректності використовуваних моделей і алгоритмів.

5. Візуалізація результатів аналізу. Після отримання остаточних результатів аналітичної обробки на перший план виходить задача їхньої інтерпретації та оцінки достовірності отриманих результатів. Проте варто зауважити, що навіть за умови, що в процесі аналізу були отримані достовірні і цінні результати, невдалий вибір візуалізації не дозволить їх інтерпретувати, побачити в них залежності і закономірності.

8.2. Групи методів візуалізації

У бізнес-аналітиці використовується кілька десятків методів візуалізації. Вибір методу визначається особливостями і характером даних, специфікою розв'язуваної задачі, власними уподобаннями.

Табличні та графічні. Як правило, таблиці застосовуються в тому випадку, коли користувачеві необхідно працювати з окремими значеннями даних, вносити зміни, контролювати формати даних, пропуски, протиріччя тощо. Стандартне табличне подання дає можливість сортувати, експортувати та фільтрувати дані. Належним чином побудована таблиця може досить ефективно унаочнювати певні властивості даних.

Графічні методи дозволяють краще побачити загальний характер даних: закономірності, тенденції, періодичні зміни. Графічні методи більш ефективно зіставляють дані. Наприклад, досить побудувати графіки двох досліджуваних процесів в одній системі координат, щоб оцінити ступінь їхньої подібності та відмінності.

Одновимірні і багатовимірні. Одновимірні візуалізатори являють собою інформацію лише про один вимір даних, тоді як багатовимірні – про два або більше. Якщо графік показує залежність суми продажів від дати, то він буде одновимірним, оскільки на ньому відображатиметься тільки один вимір – дата, значенням якого відповідатиме факт – ціна. Необхідно звернути увагу на те, що з однією відміткою часу може бути пов'язане одне або декілька значень даних.

Двовимірні дані мають два окремих виміра. Типовим прикладом можуть бути географічні дані, які характеризуються двома атрибутами: довгота і широта. Зазвичай такі дані відображаються у вигляді точок у двовимірній системі координат. Ця простота візуалізації одновимірних і двовимірних даних може бути оманлива. За великим обсягом даних їхня візуалізація може не дати бажаного ефекту. Наприклад, під час візуалізації великої кількості даних на довгому часовому інтервалі відображення їх на екран призведе до необхідності значного зменшення масштабу і, як наслідок, втрати наочності.

Якщо ж інформація про продажі наводиться за датами і найменуваннями товарів, то з'являється ще один вимір – товар, і тоді для коректного представлення даних використовується багатовимірний візуалізатор. Популярними багатовимірними візуалізаторами є: OLAP-куб, багатовимірна діаграма, карта Кохонена та ін. Прикладами багатовимірних даних є реляційні таблиці баз даних, де атрибутами даних є колонки. Для відображення таких даних на екрані монітора потрібні спеціальні методи візуалізації, наприклад паралельні координати.

Методи загального призначення і спеціалізовані. Методи візуалізації загального призначення не пов'язані з будь-яким певним видом завдань аналізу або типом даних і можуть використовуватися на будь-якому етапі аналітичного процесу. Це своєрідні

типові візуалізатори: графіки та діаграми, графи, гістограми і їхні різновиди, статистичні характеристики тощо. Водночас існує ряд завдань, специфіка яких вимагає застосування спеціалізованих візуалізаторів. Наприклад, карти Кохонена розроблені для візуалізації результатів кластеризації, матриці класифікації використовуються в основному для перевірки спроможності класифікаційних моделей, а за допомогою діаграм розсіювання оцінюється коректність регресійних моделей.

Не всі дані можуть бути описані в термінах вимірювань. До таких даних належать тексти, гіпертексти тощо. Вони характеризуються тим, що не можуть бути описані кількісними характеристиками, і, як наслідок, не всі методи візуалізації можуть бути для них використані. Тому перед їхньою візуалізацією для них потрібні додаткові перетворення, що готують ці тексти до вигляду, який придатний для використання методів візуалізації.

Дані можуть перебувати в деяких взаємозв'язках з іншими частинами інформації. Для подання таких взаємозв'язків широко використовуються графи.

Граф складається з набору вузлів і з'єднують їх лінії, які називаються ребрами (дугами). Прикладами таких даних можуть бути: електронна пошта, що пересилається між людьми, гіперпосилання між Web-документами, файлова структура диска тощо. Для відображення таких зв'язків використовуються спеціальні методи візуалізації.

Інший клас даних – алгоритми і програми. Копіювання великих програмних проєктів є проблемою. Метою їхньої візуалізації є підтримка в розробці програмних засобів. Наприклад, відображення потоку інформації в програмі для кращого розуміння алгоритмів і написаного коду, відображення структури тисячі рядків вихідного коду у вигляді графа і підтримки процесу налагодження, візуалізація помилок. Існує велика кількість інструментів і систем, які реалізують ці завдання. Для візуалізації перерахованих типів даних використовуються різні візуальні образи і методи їхнього створення (наприклад, UML – універсальна мова моделювання).

Під час вивчення різних видів візуалізації рекомендується розглядати їх у контексті завдань, для яких вони найбільш часто застосовуються. Виділяються такі групи методів візуалізації:

– загального призначення – застосовуються для вирішення типових задач аналізу даних: візуальної оцінки якості і характеру даних, розподілу значень ознак, статистичних характеристик тощо;

– OLAP-аналіз – комплекс методів для візуалізації багатовимірних даних;

– оцінка якості моделей – дозволяє оцінювати різні характеристики моделей, такі як точність, ефективність, достовірність результатів, стійкість тощо;

– інтерпретація результатів аналізу – слугує для представлення остаточних результатів аналізу у вигляді, найбільш зручному з точки зору їхньої інтерпретації аналітиком та сприйняття аудиторії.

Підсистеми візуалізації даних містяться не тільки в спеціалізованих аналітичних платформах, а й практично у всіх програмних засобах, які пов'язані з обробкою даних. Однак в аналітичних платформах візуалізації даних приділяється особлива увага, оскільки вона є однією зі складових аналітичного процесу, без якої неможливо ефективно вирішити поставлені завдання.

Кращих результатів можна досягти, якщо вважати візуалізацію не окремою підсистемою, а частиною аналітичного процесу як, наприклад, моделювання, очищення і трансформація. Це дозволить отримати максимум корисної інформації у випадках, коли застосування інших методів є неефективним.

У випадку коли для побудови якісної моделі даних з різних причин (наприклад, є пропуски даних) недостатньо, візуалізація дозволяє висувати гіпотези, робити висновки на основі експертних оцінок (або інших методів), розробляти способи підвищення інформативності даних.

Можна виділити набір засобів візуалізації, які включаються до складу будь-якого аналітичного додатка і застосовуються на всіх вищезазначених етапах. Такі засоби візуалізації називаються візуалізаторами загального призначення. До них належать:

– графіки;

– діаграми;

– гістограми;

– статистика.

Графіки. Графіки являють собою лінії, що відображають залежність між декількома змінними в деякій системі координат. Лінія на графіку складається з множини точок, положення кожної з яких визначається значеннями залежної і незалежної змінної (змінних). Найчастіше використовується декартова система координат (X, У, Z). Також може застосовуватися полярна система координат (г, θ), де положення точки на координатній площині залежить від відстані до початку координат г і кута θ.

Незважаючи на свою відносну простоту, графіки є ефективним засобом візуального аналізу даних, адже часто саме з побудови графіка і починається робота з даними. За допомогою графіків оцінюється ступінь гладкості даних, наявність у них шумової складової, аномальних викидів і пропусків. Особливо корисні вони під час аналізу часових рядів. Іноді одного погляду на графік достатньо, щоб виявити наявність тренда, сезонної компоненти, оцінити ступінь впливу випадкової складової на досліджуваний процес.

На рис. 2.26 представлені елементи звичайного (зверху) і згладженого (знизу) графіків.

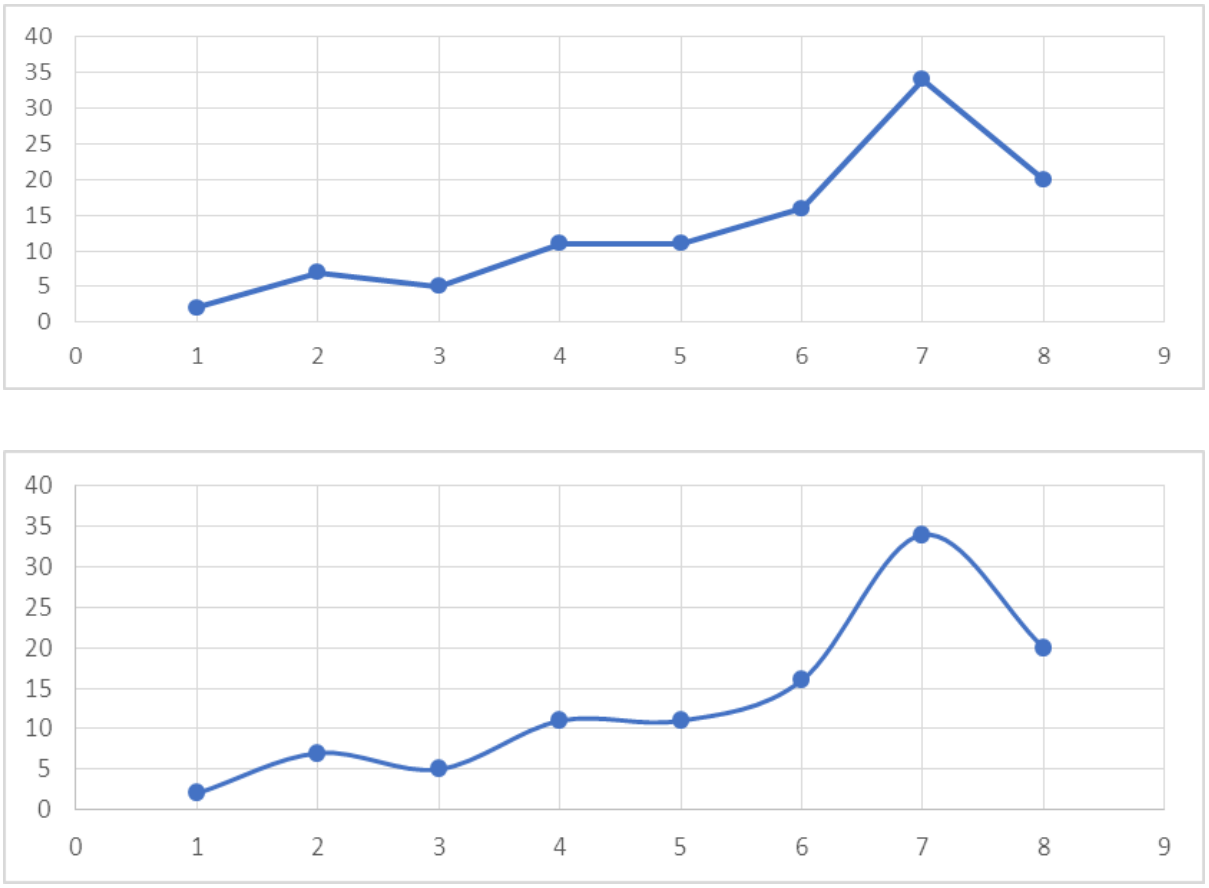


Рис. 2.26. Звичайний і згладжений графіки

У багатьох випадках гладкі графіки зручніші, ніж ламані, для візуального сприйняття. Часом вони коректніше відображають реальні бізнес-процеси, які також найчастіше змінюються плавно. Іноді точки, за якими будується графік, взагалі не з'єднують, у цьому випадку графік називається точковим.

Якщо на графіку у вас можуть запитати кілька рядів даних, то в одній системі координат будується декілька ліній. Але для цього необхідно, щоб усі ряди відображаються на графіку, мали однакові одиниці виміру і могли бути представлені в одному і тому самому масштабі. Наприклад, якщо потрібно порівняти щомісячні продажі за три роки, то можна скористатися таким графіком, як на рис. 2.27.

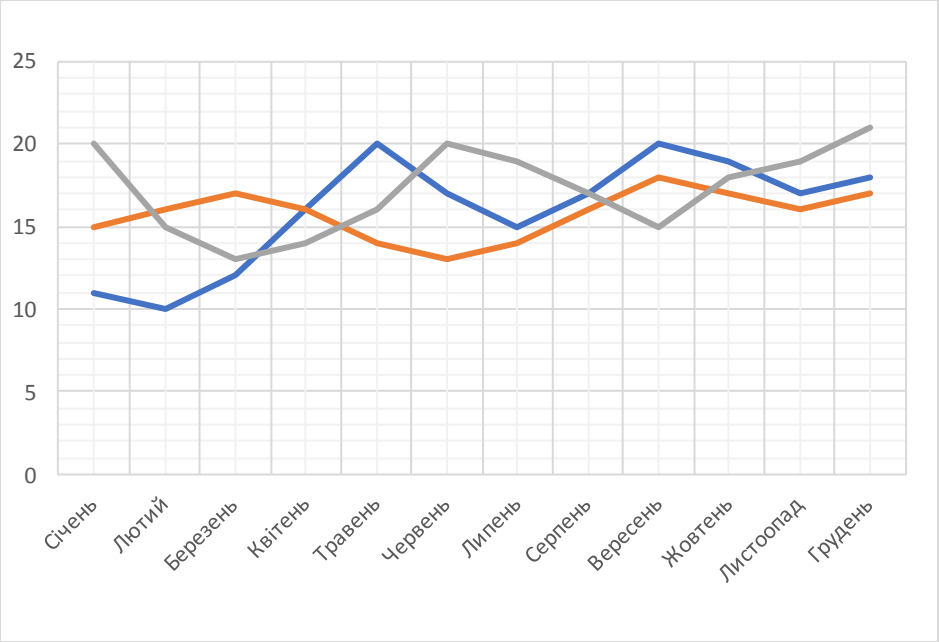
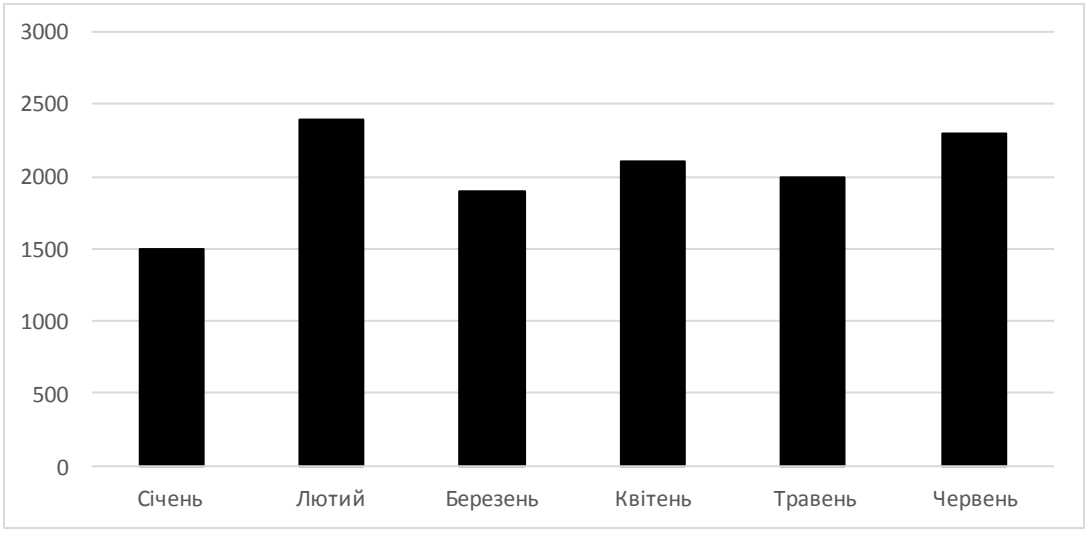


Рис. 2.27. Графік щомісячного продажу за три роки

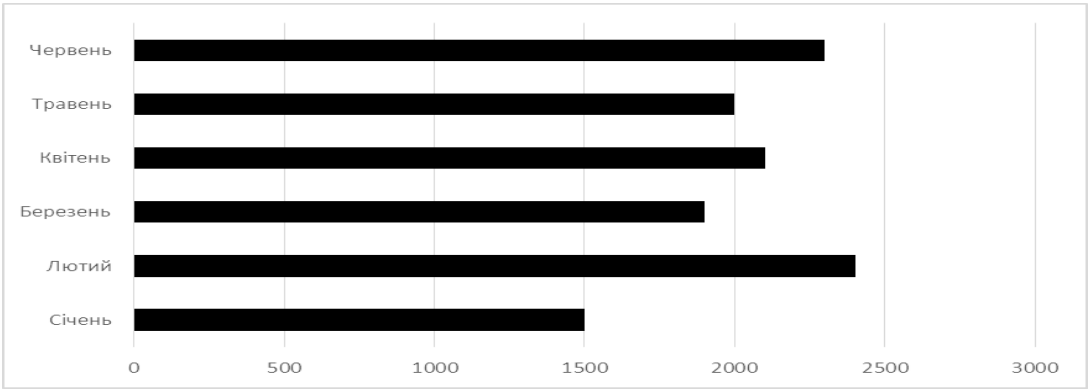
Якщо значення в рядах даних полягають у різних діапазонах і відрізняються на декілька порядків, то відображення цих рядів на одному графіку може викликати певні труднощі. Припустимо, потрібно порівняти динаміку продажів декількох товарів. Але ціни на різні товари навіть в одній групі можуть відрізнятися в десятки і сотні разів. Так, коштовні агрегати для автомобіля (коробка передач, двигун та ін.) продаються одинично, а дрібні деталі за ціною в декілька гривень (гайки, болти та ін.) – сотнями. У цьому випадку вдаються до такого прийому, як стиснення діапазону значень, що відображаються на графіку. Наприклад, це можна зробити за допомогою логарифмування або нормалізації.

Діаграми. За допомогою графіка найзручніше відображати безперервні (числові) величини, оскільки можна отримати достатню кількість точок, щоб його побудувати. Якщо ж йдеться про категоріальні (дискретні) значення, то більш придатним засобом візуалізації є діаграма. Принципової різниці між поняттями «графік» і «діаграма» немає. За графіком традиційно розуміють уявлення залежностей у вигляді ліній, тоді як на діаграмі значення відображаються за допомогою найрізноманітніших об'єктів і фігур. Як правило, на діаграмах по горизонтальній осі X відкладаються категорії, а по вертикальній осі – У-значення.

Найпростіші і часто використовувані діаграми – стовпчикові. На них значення кожної категорії представляється у вигляді стовпчика, висота якого пропорційна відповідному значенню (рис. 2.28, а). Різновидом стовпчикової діаграми є лінійчата діаграма, яка відрізняється від стовпчикової тим, що вісь категорій відкладається вертикально, а вісь значень – горизонтально (рис. 2.28, б).



а)



б)

Рис. 2.28. Стовпчикова (а) і лінійчата (б) діаграми

Ще одним поширеним видом діаграм є кругова діаграма. Її досить зручно використовувати, якщо потрібно показати частку, яку вносить певне значення в загальну суму. Ця частка може бути виражена як в абсолютних одиницях, так і у відсотках (наприклад, відсоток від виручки, який забезпечив товар А). Приклад кругової діаграми розподілу місячних продажів побутової техніки й електроніки за категоріями товарів представлений на рис. 2.29.

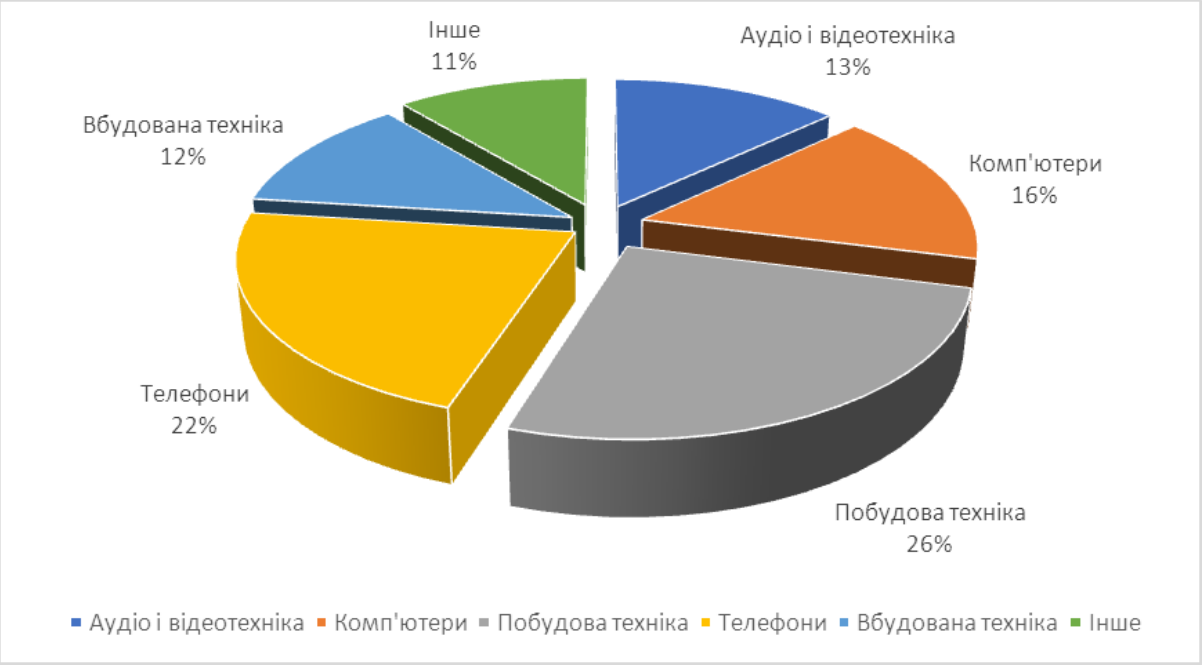


Рис. 2.29. Кругова діаграма

Хорошу наочність забезпечують пелюсткові діаграми. На них кожна категорія даних представлена у вигляді окремої осі (пелюстки), на якій відображається відповідне значення. Потім значення на всіх осях з'єднуються лініями. Якщо в ряду даних усі крапки мають однакові значення, то пелюсткова діаграма набуває вигляду кола. Наявність на діаграмі декількох рядів дозволяє порівнювати, наприклад, динаміку зміни структури продажів у групах товарів за місяцями (рис. 2.30).

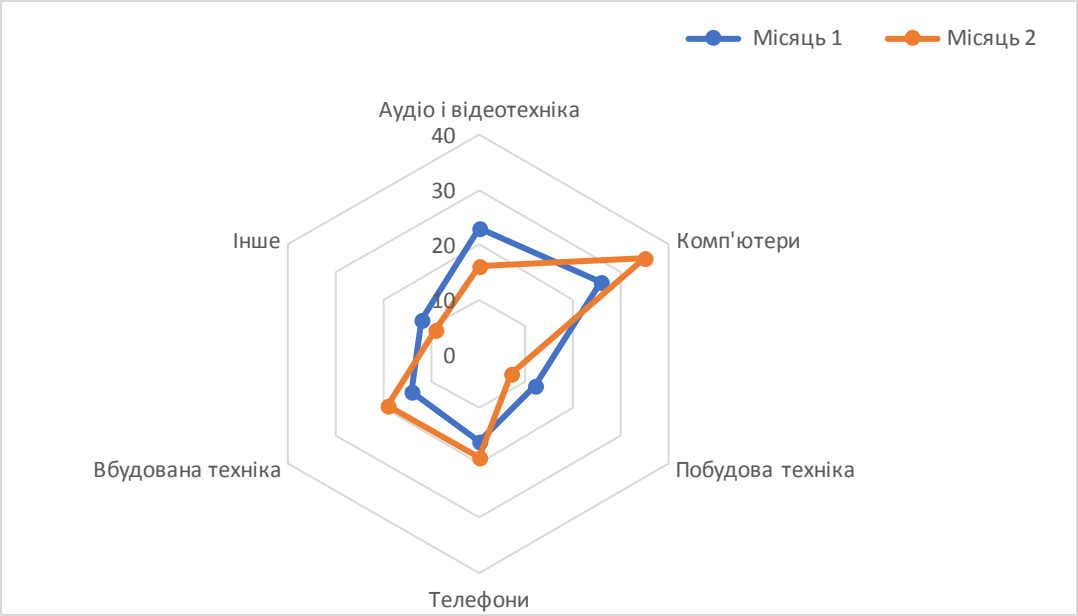


Рис. 2.30. Динаміка зміни структури продажів у групах товарів за місяцями

Існує безліч інших видів діаграм: діаграми з накопиченням, зі сферами, бульбашкові, кільцеві та ін. При цьому різновид діаграми підбирається виходячи з тієї інформації, яку бажає отримати від неї користувач.

Гістограми. Гістограма показує розподіл набору даних всередині вибірки (наприклад, кількість позичальників банку в декількох вікових групах) у вигляді стовпчиків.

Гістограми широко використовуються в статистиці для визначення найбільш імовірних значень, які може набувати деяка величина, а також для визначення закону розподілу, якому підпорядковується випадкова величина.

Гістограма будується в такий спосіб. Нехай досліджуваною величиною є щоденні продажі торгової точки упродовж місяця. При цьому мінімальне спостережуване значення становить 10 тис. грн, а максимальне – 100 тис. грн.

Розіб’ємо діапазон зміни величини на 9 піддіапазонів по 10 тис. грн і підрахуємо, скільки разів значення продажів потрапляє в певний піддіапазон. Результати зведемо в табл. 2.12.

Таблиця 2.12

Вихідні дані для гістограми

Діапазон у тис. грн	**10–20**	**20–30**	**30–40**	**40–50**	**50–60**	**60–70**	**70–80**	**80–90**	**90–100**
Частота	1	1	2	6	10	6	3	1	1

Як бачимо, продажі на суму від 10 до 20 тис. грн спостерігалися тільки один раз, від 30 до 40 тис. грн – два рази тощо. На основі отриманої таблиці будуємо гістограму (рис. 2.31).

По горизонтальній осі гістограми відкладаються значення продажів, а по вертикальній – кількість або частота спостережень, значення яких потрапили в заданий діапазон (тому іноді гістограму називають частотним полігоном).

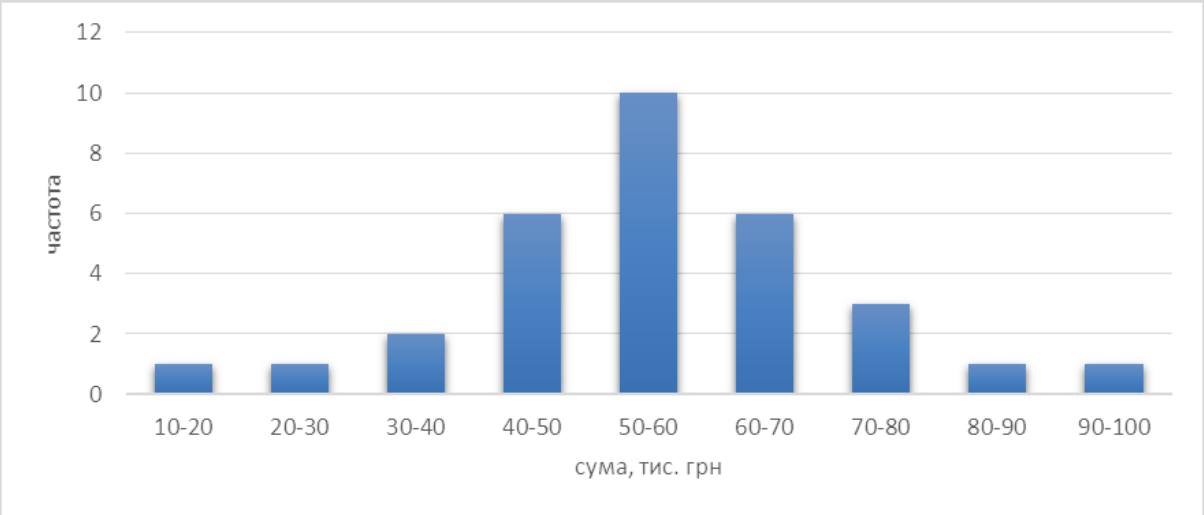


Рис. 2.31. Гістограма на основі отриманої таблиці

Гістограма на рис. 2.31 показує, що найбільша кількість спостережень потрапила в діапазон 50–60 тис. грн. Таким чином, значення з цього діапазону можна розглядати як найбільш імовірні. Цю інформацію використовують для відновлення пропущених значень під час очищення даних, для планування грошових надходжень, закупівель тощо.

Що стосується крайніх елементів гістограми, то вони являють собою рідкісні події – екстремально високі або екстремально низькі значення продажів. Видно, що в діапазон 10–20 тис. грн

потрапило лише одне значення, отже, ймовірність такої події мала і його не варто включати в розгляд. Іноді використовують нормовану гістограму, що дозволяє оперувати не значеннями спостережень, а їх можливостями. Для цього кожен елемент гістограми поділяється на кількість спостережень, тобто в нашому випадку на 31 (кількість днів у місяці). Розглянемо рис. 2.32.

Тепер висота стовпчика визначається не кількістю спостережень, що потрапили у відповідний діапазон, а ймовірністю попадання в нього. На рисунку видно, що ймовірність попадання значення в діапазон 50–60 тис. грн становить приблизно 0,32, або 32 %. Відповідно, ймовірність появи значень у діапазоні 10–20 тис. грн не перевищує 0,03, або 3 %.

У нормованої гістограмі сума значень усіх її елементів повинна дорівнювати 1, оскільки сума ймовірностей усіх можливих подій (попадання значення в будь-якій діапазон) є 1.

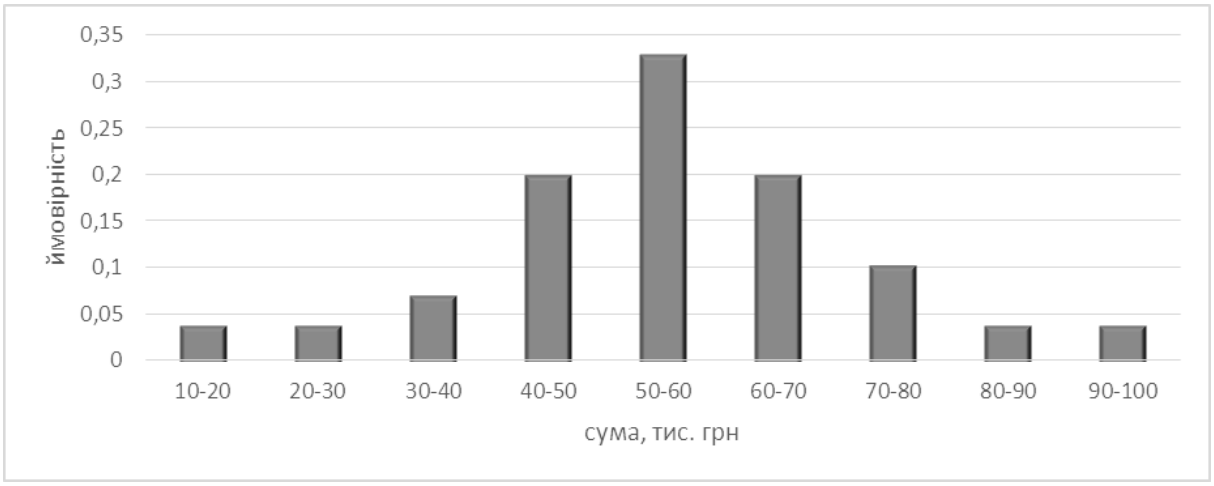
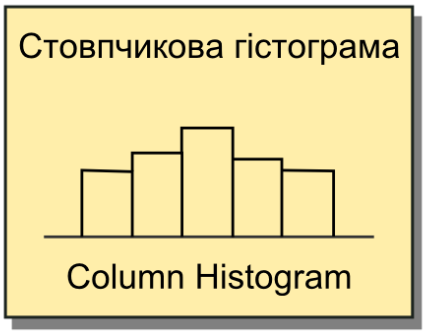
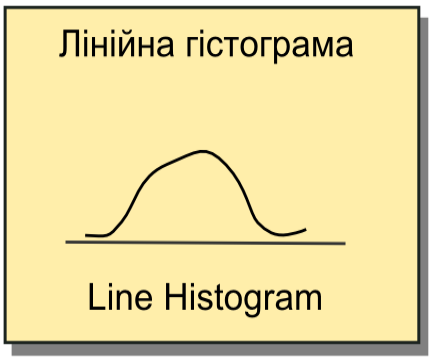
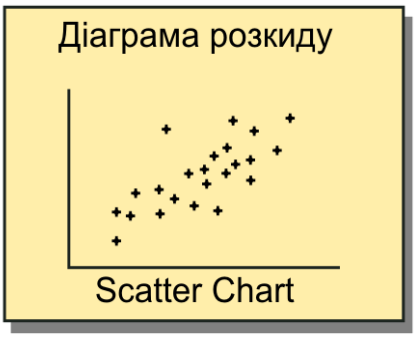
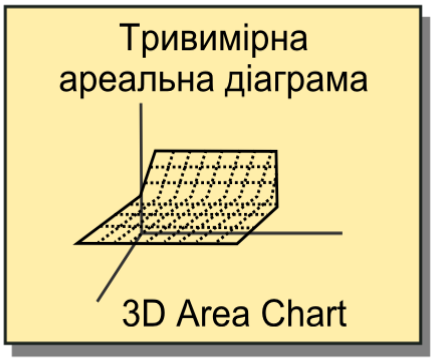


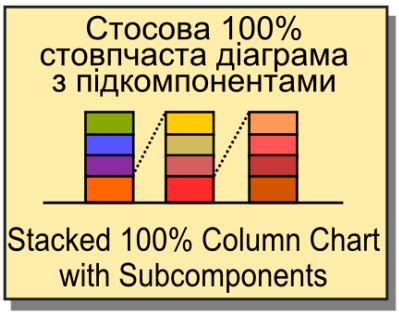
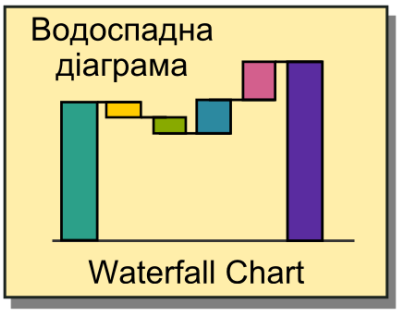
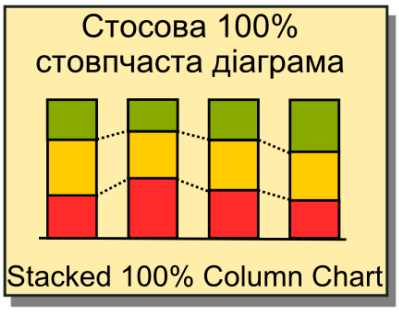
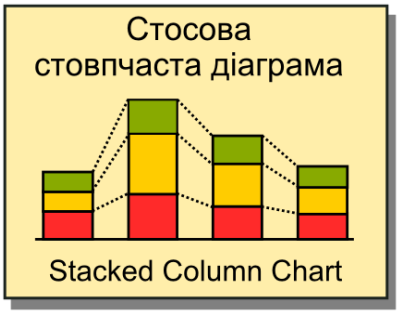
Рис. 2.32. Нормована гістограма

Зазвичай під час побудови гістограми аналітик має можливість задати кількість піддіапазонів, на яку розбиватиметься вихідний діапазон зміни величини (фактично це кількість стовпців гістограми). Тут існують різні рекомендації. На практиці можна керуватися таким емпіричним правилом. Кількість стовпців у гістограмі має бути такою, щоб у ній не утворювалися провали, різкі викиди або множинні піки. Вона повинна бути досить гладкою, щоб по ній можна було визначити характер розподілу спостережуваної величини. Часто хороших результатів вдається домогтися під час використання 10–15 стовпців.

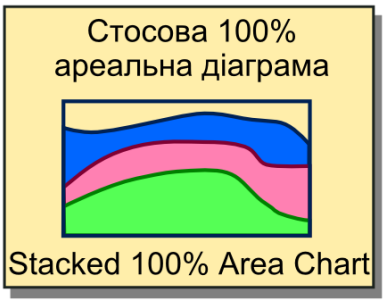
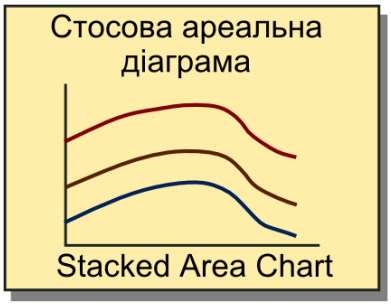
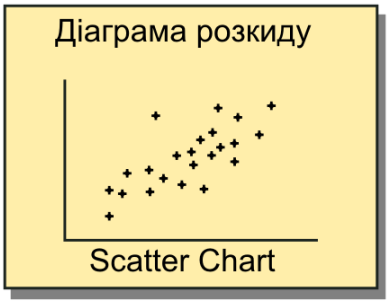
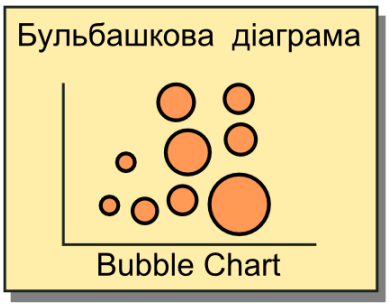
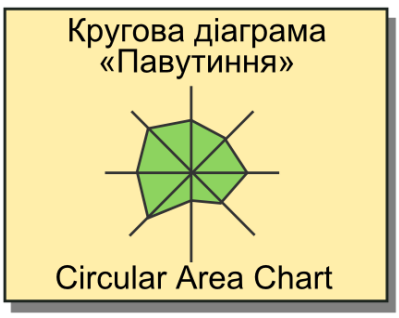
Таблиця 2.13

Ціль візуалізації	Особливості даних	Особливості подання	Візуалізація
1	2	3	4
Розподіл	Одна змінна	Небагато точок даних / категорій / інтервалів	
Розподіл	Одна змінна	Багато точок даних	
Розподіл	Дві змінні		
Розподіл	Три змінні		

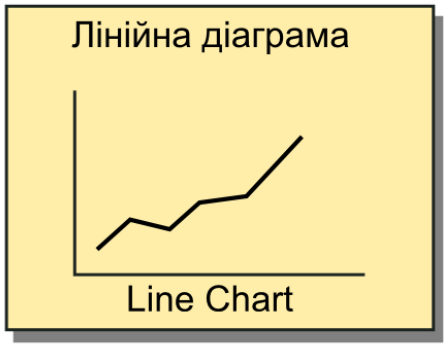
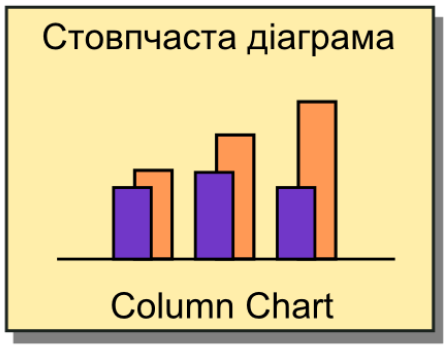
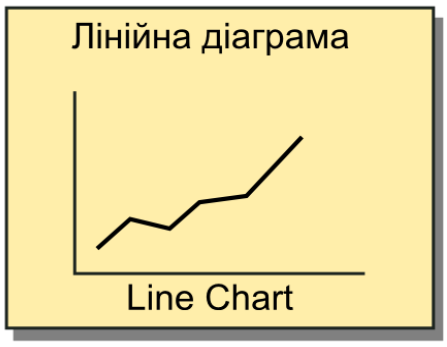
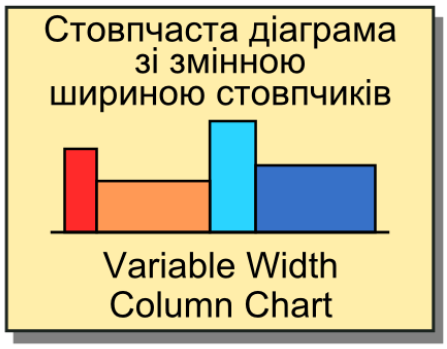
Продовження таблиці 2.13

1	2	3	4
Компо-зиція (будова)	Статична	Проста частина В цілому	
Компо-зиція (будова)	Статична	Компоненти компонентів	
Компо-зиція (будова)	Статична	Накопичення додавань і віднімань у цілому	
Зміна впродовж часу	Небагато періодів	Важливі тільки відносні відмінності	
Зміна впродовж часу	Небагато періодів	Важливі й абсолютні, і відносні відмінності	

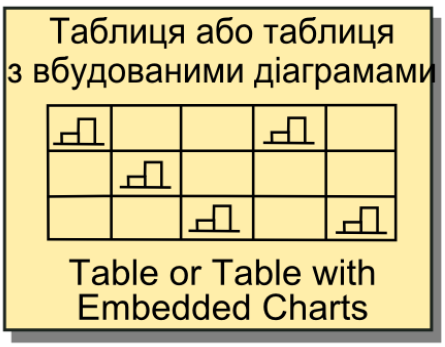
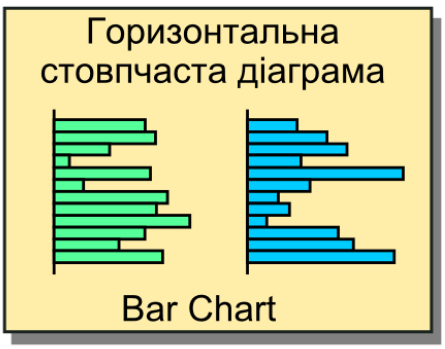
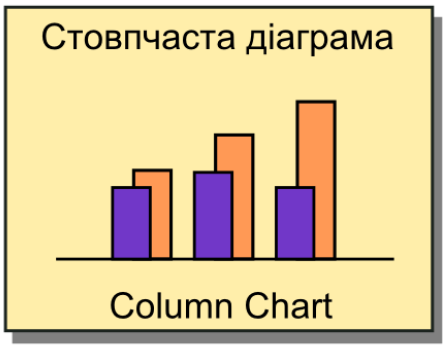
Продовження таблиці 2.13

1	2	3	4
Зміна впродовж часу	Багато періодів	Важливі тільки відносні відмінності	
Зміна впродовж часу	Багато періодів	Важливі й абсолютні, і відносні відмінності	
Зв’язок чи залеж-ність	Дві змінні		
Зв’язок чи залеж-ність	Три змінні		
Порів-няння	У часі	Багато періодів – циклічні дані	

Продовження таблиці 2.13

1	2	3	4
Порів-няння	У часі	Багато періодів – ациклічні дані	
Порів-няння	У часі	Мало періодів, одна чи декілька категорій	
Порів-няння	У часі	Мало періодів, багато категорій	
Порів-няння	Між об’єктами	Дві змінні на об’єкт	

Продовження таблиці 2.13

1	2	3	4
Порів-няння	Між об’єктами	Одна змінна на об’єкт, багато категорій	
Порів-няння	Між об’єктами	Одна змінна на об’єкт, мало категорій, багато об’єктів	
Порів-няння	Між об’єктами	Одна змінна на об’єкт, мало категорій, мало об’єктів	

Під час вибору відповідного графіка можна керуватися також діаграмою з книги Джина Желязни «Говори мовою діаграм»[23].

[23] Желязны Д. Говори на языке диаграмм : пособие по визуальным коммуникациям для руководителей ; пер. с англ. М. : Институт комплексных стратегических исследований, 2004. 220 с.

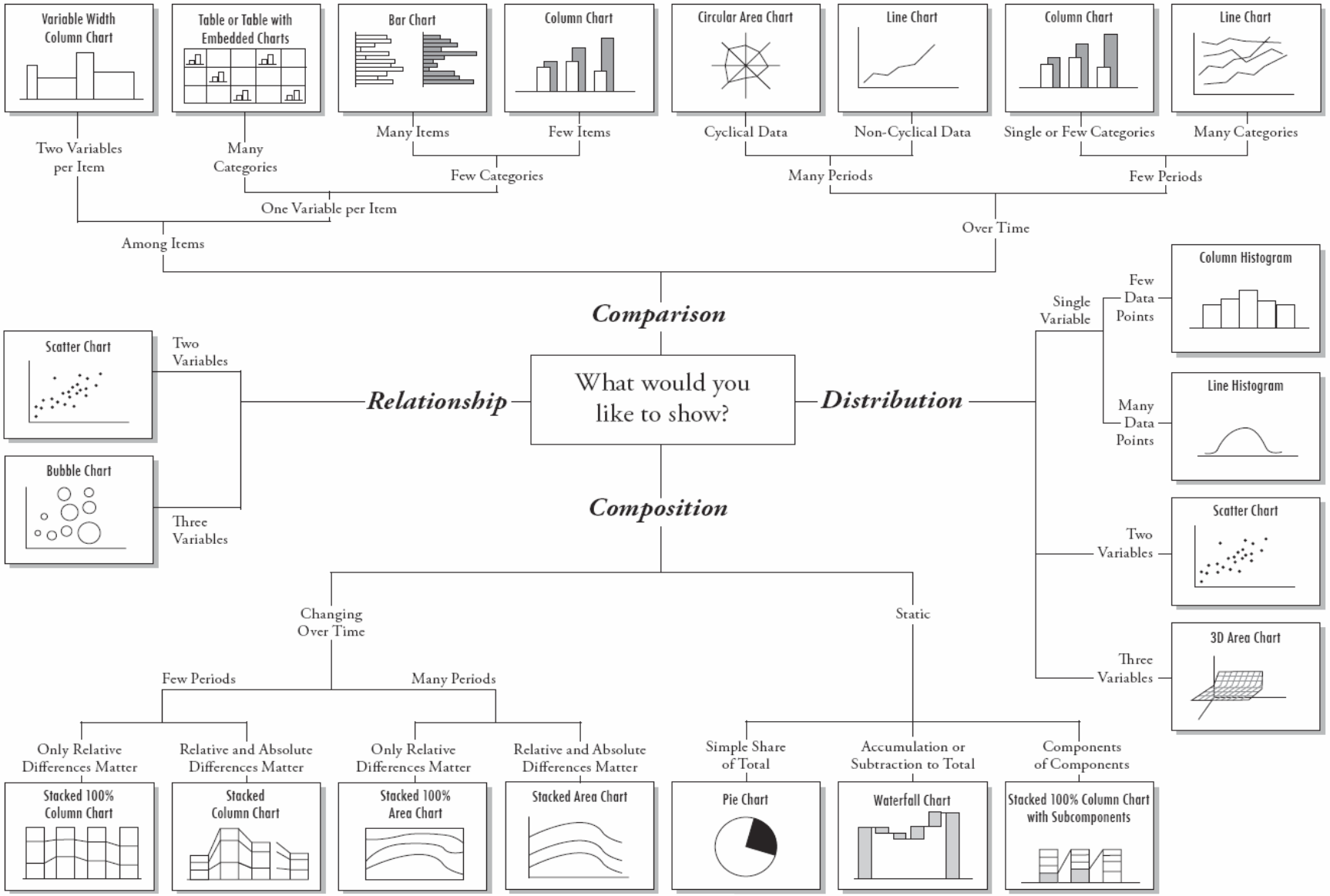


Рис. 2.33. Діаграма Джина Желязни

Статистика. Ще одним поширеним засобом візуалізації, яке входить до складу більшості аналітичних програм, є інформація про статистичні характеристики досліджуваної вибірки. Вони зазвичай подаються в табличному вигляді і вказуються для кожного поля (атрибута) вибірки. Статистичні характеристики дозволяють висувати гіпотези про поведінку даних і властиві їм закономірності, контролювати результати обробки даних на різних етапах аналітичного процесу. Типовий вид візуалізатора статистики представлений у табл. 2.14.

Таблиця 2.14

Статистичні характеристики

№ з/п	Мітка поля	Статистика: кількість значень = 150					
		мінімум	максимум	середнє	стандартне відхилення	кількість унікальних значень	кількість пропусків
1.	Довжина чашелі стіки	43,00	79,00	58,43	9,25		0
2.	Ширина чашелі стіки	20,00	44,00	30,57	40,34		0
3.	Довжина пелюстки	10,00	69,00	37,58	17,59		0
4.	Ширина пелюстки	1,00	25,00	11,99	7,60		0
5.	Клас квітки	6	9	8	1,14	3	0

Статистика, як правило, включає такі характеристики. Мінімум і максимум дозволяють визначити діапазон зміни значень величини. Знання мінімального і максимального значень дає можливість побачити, чи полягає величина в діапазоні, допустимому для застосування в певній аналітичній моделі, вибрати коректну методику порівняння величин тощо.

Середнє значення і математичне сподівання дозволяють висувати гіпотези про найбільш імовірні значення, які може приймати досліджувана величина.

Стандартне (середньоквадратичне) відхилення і дисперсія показують ступінь розсіювання значень величини щодо середнього.

Знання цих характеристик дозволяє оцінювати гладкість рядів даних, наявність у них шумів, контролювати ступінь згладжування даних у процесі їхньої попередньої обробки тощо.

Розподіл – вказується відповідність досліджуваної вибірки деякому статистичному розподілу (нормальному, рівномірному, експоненціального та ін.). Знання розподілу досліджуваної величини дозволяє пояснити особливості її поведінки, наприклад визначити, які значення є найбільш імовірними.

У набір визначених статистичних характеристик можуть включатися медіана, коефіцієнти асиметрії й ексцесу та інші статистичні величини. Також візуалізатор статистики часто містить кількість унікальних значень для дискретних величин і кількість пропущених значень, виявлених у вибірці.

OLAP-аналіз. Більшість реальних бізнес-процесів є складними, оскільки в них бере участь багато об'єктів, які перебувають у найрізноманітніших відносинах і з кожним з яких може бути пов'язано декілька числових характеристик. Тому під час візуалізації даних часто постає необхідність представити складні дані в такому вигляді, щоб людина могла їх осмислити й інтерпретувати.

Можливим виходом з цієї ситуації є декомпозиція складної таблиці на більш прості, де міститься інформація про продажі окремих товарів і груп товарів за містами, інтервалами дат, постачальниками тощо. Але такий підхід дозволяє лише частково вирішити проблему. Справа в тому, що процес аналізу даних вимагає порівняння і зіставлення даних, обчислення різних показників і характеристик тощо. Якщо дані про продажі розподілені за декількома таблицями, кожна з яких складена для одного міста, то як знайти, наприклад, товар, продажі якого були максимальними в кожному місті. Для цього доведеться сформувати зведений звіт і виявити відповідні максимуми. З урахуванням величезних обсягів даних кожна така операція може зайняти досить багато часу і вимагати більших обчислювальних витрат.

Усі ці складності, що виникають під час обробки великих і складних масивів даних, вирішуються методом візуалізації багатовимірних табличних даних – OLAP-аналізом.

В основі OLAP полягає багатовимірне представлення даних, які можуть бути розділені на кількісні та якісні. Якісні дані являють собою значення, виражені в категоріальній формі. Зазвичай це найменування товарів, груп товарів, організацій, назви міст, ПІП співробітників тощо. З кожним об'єктом пов'язані ознаки, які кількісно описують його. Для товару це може бути ціна, кількість або сума; для міста, в якому розташоване торгове представництво, – відстань до нього і кількість жителів; для співробітника – заробітна плата і стаж роботи.

З кожним якісним значенням аналізованого бізнес-процесу пов'язані один або декілька кількісних показників. У рамках багатовимірної моделі дані, що якісно описують досліджуваний бізнес-процес, називаються вимірами. Вимірами можуть бути Товар, Місто, Клієнт, Організація та ін. До них відносять і дату. Дані, які кількісно описують процес або об'єкт, називаються фактами. Елементи фактів: Кількість, Сума, Вік, Дохід, Торгова націнка тощо. Іншими словами, з кожним виміром пов'язані один або декілька фактів.

Вимірювання несуть смислове навантаження, а факти – кількісне. Щоб достовірно відокремити вимірювання від фактів, досить зіставити значення з питанням. Кожен вимір містить обмежений набір значень (категорій). Для вимірювання Товар значеннями будуть найменування товарів: Холодильник, Праска, Пилосос та ін. Кожному значенню вимірювання в багатовимірній моделі відповідають одне або декілька значень фактів. Так, для значень вимірювання Товар можуть зазначатися ціна і кількість на складі.

Абревіатура OLAP також зустрічається в теорії сховищ даних. Але якщо там вона вживається в контексті моделей зберігання даних (багатовимірної, реляційної), то тут під OLAP мається на увазі метод візуалізації, при цьому модель зберігання даних може бути будь-якою.

Щоб побудувати OLAP-куб, користувач повинен вказати системі такі параметри:

– які вимірювання і факти включати в куб;

– методи агрегації значень фактів.

Візуалізація OLAP-куба проводиться за допомогою спеціального виду таблиць, які будуються на основі зрізів OLAP-куба, що містять необхідну користувачу інформацію. Зрізи, в свою чергу, є результатом виконання відповідного запиту до бази даних.

Як правило, в процесі побудови зрізів користувач за допомогою миші і клавіатури маніпулює заголовками вимірювань, домагаючись найбільш інформативного представлення даних у кубі. Залежно від положення заголовків вимірювань у таблиці автоматично формується запит до бази чи сховища даних.

На рис. 2.34 показано приклад візуалізації даних за допомогою OLAP-куба, де представлена інформація про кількість і суми продажів за аптеками в розрізі року і місяця.

Виміри

Факти

		Відділ. Назви							
		Аптека 1		Аптека 2		Аптека 3		Всього	
		Кількість	Сума	Кількість	Сума	Кількість	Сума	Кількість	Сума
2004	Січень	573	33284,1					573	33284,1
	Лютий	623	33809,2					623	33809,2
	Березень	534	32241,2					534	32241,2
	Квітень	527	33488,0	353	19370,6			880	52858,6
	Травень	449	22377,6	224	10759,3			673	33136,9
	Червень	425	21364,1	186	8160,5			611	29524,6
	Липень	373	13536,4	164	8158,1			537	21694,5
	Серпень	312	14324,6	227	10764,9			539	25089,5
	Вересень	453	22436,4	278	15008,1			731	37444,5
	Жовтень	536	31328,3	361	21777,8	566	35965,4	1463	89071,5
	Листопад	588	33413,9	281	16416,8	603	46399,3	1472	96229,9
	Грудень	591	32596,5	350	21365,7	655	43121,3	1596	97083,5
	Всього	5984	325200,1	2424,0	131781,8	1824	125486,0	10232	582468,0
Всього		5984	325200,1	2424,0	131781,8	1824	125486,0	10232	582468,0

Рис. 2.34. Приклад OLAP-куба

У кубі на рис. 2.34 для кожного місяця і року перераховані всі продажі, які були зроблені в межах цього періоду, із зазначенням кількості проданих одиниць товару і їхньої загальної суми. Візуальний аналіз дозволяє помітити, що найбільше продажів припало на «Аптеку 1».

Для більш детального аналізу представляють інтерес продажу за днями. Щоб перетворити куб у відповідний вид, досить додати ще один вимір – День місяця, як показано на рис. 2.35.

Таким чином, OLAP-куб можна використовувати не тільки як метод візуалізації, але і як засіб оперативного формування звітів та подання інформації в потрібному розрізі (так звана аналітична звітність).

			Відділ. Назви					
			Аптека 1		Аптека 2		Всього	
Рік	Місяць	День міс.	Кількість	Сума	Кількість	Сума	Кількість	Сума
2004	Січень		202	13544,3			202	13544,3
	Лютий		221	13544,3			221	13544,3
	Березень		183	10951,5			183	10951,5
	Квітень		206	13363,5	122	9308,2	329	22671,7
	Травень		141	5670,5	73	3762,4	214	9433,0
	Червень		183	9965,7	94	3822,1	277	13787,8
	Липень		136	5528,0	55	3314,7	191	8842,7
	Серпень		130	4928,6	57	2576,4	187	7505,0
	Вересень		137	7072,7	70	4300,3	207	11373,0
	Жовтень	1	15	885,3	25	647,0	40	1532,3
		2	10	971,2	21	937,7	31	1908,9
		3	25	827,5	12	453,0	37	1280,5
		4	16	950,2	14	1161,7	30	2111,9
		5	17	1309,6	18	1305,0	35	2614,6
		6	31	1289,6	16	1330,4	47	2620,0
		7	20	1580,7	12	932,1	32	2512,8
		8	14	719,3	9	656,2	23	1375,5
		9	21	2099,9	5	389,5	26	2489,4
		10	18	1080,9	9	921,4	27	2002,3
		Всього	187	11714,1	141	8734,1	328	20448,2
	Листопад		186	10308,2	81	4984,4	267	15292,6
	Грудень		192	9856,4	116	7360,3	308	17216,7
	Всього		2104	112845,7	799	448162,9	2903	561008,6
Всього			2104	112845,7	799	448162,9	2903	561008,6

Рис. 2.35. OLAP-куб з доданим виміром «День місяця»

Найбільший інтерес OLAP-куб представляє з точки зору візуального аналізу даних, пошуку особливостей і закономірностей у даних. При цьому існує два підходи.

1. Аналітик може задатися кількома запитаннями: «Який товар найпопулярніший?», «Де продажі були найкращими і з чим

це пов'язано?» та ін. Потім за допомогою маніпуляцій вибирається таке представлення куба, яке дозволяє відповісти на поставлені запитання.

2. Можна підійти до проблеми і з іншого боку – послідовно перебирати можливі варіанти представлення даних, і за їхньою допомогою зіставляти дані, виявляти закономірності і зв'язки між елементами даних, висувати гіпотези тощо.

Вміле використання OLAP-аналізу і робота з кубом часом дозволяють отримати результати навіть у тих випадках, коли методи Data Mining виявляються малоефективними.

Для того щоб зробити процес більш гнучким і ефективним, в OLAP-системах зазвичай передбачається набір операцій з вимірами, що дозволяють отримати максимум можливих представлень даними.

Заміна і додавання вимірювань. Будь-які виміри куба можна замінити іншими, присутніми у вихідній множині даних.

Видалення вимірювань. Якщо в процесі роботи з кубом якийсь вимір втратив інформативність, то його можна виключити з куба. Залишати в кубі зайві вимірювання небажано: чим вище розмірність куба, тим складніше для розуміння й осмислення представлена в ньому інформація.

Приховування вимірювань. Користувач може вибрати в кубі кілька найбільш інформативних параметрів і працювати з ними.

Зміна порядку вимірювань. Під час роботи з декількома вимірами в кубі закладається можливість обрати порядок їхнього відображення.

Відбір значень вимірів. Часто потреби у відображенні всіх можливих значень вимірювання немає. Тому щоб не захаращувати куб непотрібними даними, зайві значення вимірювань можуть бути тимчасово приховані, а якщо стане необхідно – відображені знову.

Транспонування. Якщо у процесі роботи з кубом з'ясується, що результати вимірювання, які відображаються у стовпцях, зручніше відображати в рядках або навпаки, то відповідне перетворення можна легко зробити за допомогою операції транспонування. Приклад транспонування представлений на рис. 2.36.

	Місяць					
	Жовтень		Листопад		Грудень	
Відділ. Назви	Кількість	Сума	Кількість	Сума	Кількість	Сума
Аптека 1	187	11714,1	186	10308,2	192	9856,4
Аптека 2	141	8734,1	81	4984,4	116	7360,3
Аптека 3	13	10490,6	204	14347,7	216	12580,3
Всього	465	30938,9	471	29640,2	524	297973,0

	Відділ. Назви					
	Аптека 1		Аптека 2		Аптека 3	
Місяць	Кількість	Сума	Кількість	Сума	Кількість	Сума
Жовтень	187	11714,1	141	8734,1	328	20448,3
Листопад	186	10308,2	81	4984,4	267	15292,5
Грудень	192	9856,4	116	7360,3	308	17216,7
Всього	556	31878,7	338	21078,8	557	52957,5

Рис. 2.36. Транспонування OLAP-куба

У разі роботи з OLAP-кубами широко застосовується ще одна операція, яка називається деталізацією. Її необхідність викликана тим, що в більшості випадків значення в кубі є агрегованими, наприклад у межах деякої дати або інтервалу дат. Водночас користувача можуть цікавити й атомарні (тобто деталізовані) значення, на основі яких були отримані агреговані.

Операція деталізації полягає у відображенні набору записів вибірки даних, у результаті агрегування яких було отримано відповідне значення OLAP-куба.

У табл. 2.15 представлений фрагмент вибірки даних щодо кредитування.

Таблиця 2.15

Код клієнта	**Дата**	**Сума кредиту**	**Мета кредитування**
1	**2**	**3**	**4**
120	08.01.2007	50 000	Інше
121	08.01.2007	9 000	Покупка комп’ютера
122	09.01.2007	43 000	Оплата освіти
123	09.01.2007	2 500	Оплата послуг
124	09.01.2007	48 500	Покупка вікон
125	09.01.2007	45 500	Оплата освіти

Продовження таблиці 2.15

1	2	3	4
120	09.01.2007	32 000	Оплата послуг
121	09.01.2007	2 500	Оплата послуг
122	09.01.2007	6 500	Оплата послуг
123	09.01.2007	13 500	Оплата послуг
124	09.01.2007	31 500	Оплата освіти
125	09.01.2007	6 000	Інше
120	09.01.2007	28 000	Покупка комп'ютера
121	09.01.2007	13 500	Покупка вікон
122	09.01.2007	19 500	Оплата освіти
123	10.01.2007	18 000	Покупка вікон
124	10.01.2007	45 000	Турпоїздка

У табл. 2.15 представлений фрагмент куба, побудованого на основі цієї вибірки. Дані про кредити агреговані за датами: в кубі бачимо не суми окремих кредитів, а значення, отримані їх сумовуванням за окремими цілями кредитування в межах однієї дати. Іншими словами, з куба можна почерпнути тільки інформацію про те, на яку загальну суму було видано кредитів на певну мету в межах певної дати, а суми, видані окремим клієнтам, залишаються невідомими.

	Ціль: кредитування			
Дата	Інше	Оплата освіти	Оплата послуги	Всього
04.01.2007	135000,0	23500,0	45500,0	204000,0
05.01.2007	34500,0	38000,0	40500,0	113000,0
06.01.2007		48000,0		48000,0
07.01.2007		51000,0	31000,0	82000,0
08.01.2007	100000,0			100000,0
09.01.2007	42000,0	**139500,0**	114000,0	295500,0
Всього	311500,0	300000,0	231000,0	942500,0

Дата	Сума кредиту	Ціль кредитування
09.01.2007	43000	Оплата освіти
09.01.2007	45500	Оплата освіти
09.01.2007	31500	Оплата освіти
09.01.2007	19500	Оплата освіти

Рис. 2.37. Деталізація значень фактів у кубі

Результатом застосування деталізації до відповідного осередку куба буде Таблиця, в якій представлені записи, що містять дані, пов'язані з цією клітинкою (шляхом агрегування яких вона була отримана) (рис. 2.37).

Існує й інший варіант деталізації – у вигляді послідовного розгортання і згортання вимірювань, як це показано на рис. 2.37.

8.3. Візуалізатори для оцінки якості моделей

У процесі вирішення завдань у бізнес-аналітиці доводиться створювати одну або декілька моделей. Вони можуть бути засновані на машинному навчанні як нейронні мережі, дерева рішень тощо або чітко задані шляхом безпосереднього програмування перетворень, які ця модель повинна виконувати. Незалежно від виду побудованої моделі перш ніж застосовувати її на практиці, необхідно оцінити її якість, тобто визначити, наскільки правильно і точно вона вирішує поставлене завдання.

Можна виділити дві складові якості моделі:

– адекватність – наскільки точно модель описує досліджуваний об'єкт або процес;

– коректність – наскільки правильно модель може працювати з усіма можливими вхідними даними.

Досягти абсолютно точного опису моделлю реального бізнес-процесу або об'єкта практично неможливо, оскільки в такому випадку модель виявиться досить складною і важкореалізовуваною. Найчастіше шукають розумний компроміс між точністю і складністю. Наприклад, задається допустима точність 10 %, яку повинна дотримуватися побудована модель. Тобто потрібно, щоб класифікаційна модель зі 100 прикладів неправильно класифікувала не більше 10.

Але навіть якщо сама по собі модель володіє достатнім ступенем точності й адекватності, у разі надходження на її вхід некоректних даних вона може повести себе непередбачувано.

Щоб звести до мінімуму ймовірність виявлення неіснуючих закономірностей під час аналізу, необхідно досліджувати якість роботи моделей. Такі дослідження повинні проводитися як у процесі побудови моделі, так і після його завершення. Так, у процесі

навчання нейронної мережі з метою створення класифікатора зазвичай потрібно декілька етапів навчання. Після кожного етапу контролюється точність моделі, після чого коригуються конфігурація мережі, навчальна вибірка або параметри навчання. І це відбувається доти, доки модель не стане задовольняти заданим вимогам (рис. 2.38).

Після того як модель доведена до необхідного рівня точності, перевіряється її стійкість до некоректних даних. Якщо під час появи некоректних даних модель видає неправильні результати, то, можливо, змінювати її не буде потрібно, оскільки виявиться достатнім застосувати певні методи трансформації та очищення даних, що дозволить забезпечити коректність вихідних даних.

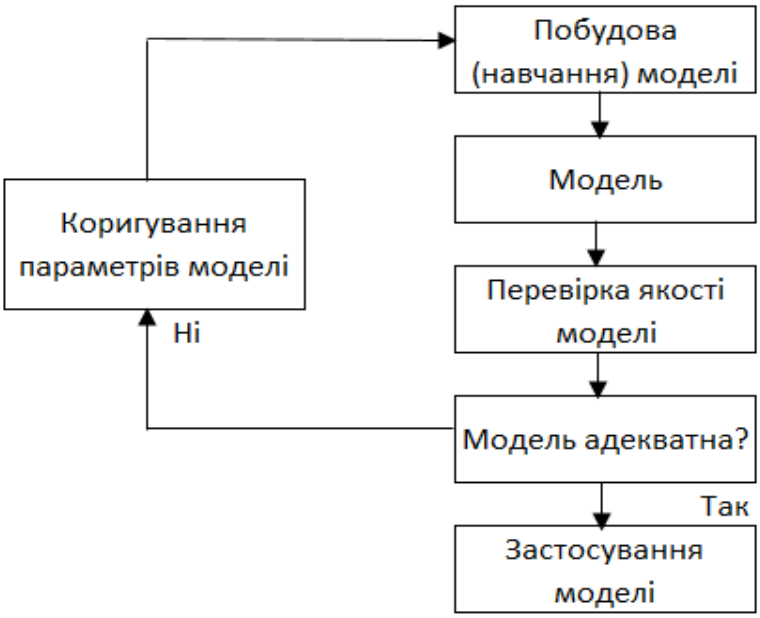


Рис. 2.38. Ітераційний характер моделювання

Для дослідження адекватності та коректності моделей до складу аналітичних додатків включаються спеціальні візуалізатори, які дозволяють оцінити точність моделі і її реакцію на певні вхідні дані.

Типовий набір візуалізаторів для оцінки якості моделей такий:

– матриця класифікації;

– діаграма розсіювання;

– ретропрогноз;

– графіки контролю процесу навчання.

Матриця класифікації, або таблиця спряженості, застосовується для оцінки якості класифікаційних моделей. Як відомо, завдання класифікації полягає у віднесенні об'єкта до одного із заздалегідь визначених класів на основі набору ознак цього об'єкта. Таким чином, вхідні дані для класифікаційної моделі являють собою ознаки (атрибути об'єктів), а на виході класифікаційної моделі повинна формуватися мітка класу, до якого належить об'єкт. При цьому вона повинна бути категоріального виду.

Для кожного прикладу класифікаційна модель формує на виході мітку класу, до якої належить об'єкт з набором ознак, зазначених у прикладі. Якщо мітка класу, сформована моделлю, збігається з цільовою міткою класу з прикладу, то такий приклад (об'єкт) є правильно розпізнаним, в іншому випадку – неправильно розпізнаним.

Співвідношення кількості правильно і неправильно розпізнаних об'єктів слугує критерієм якості моделі. Чим більше об'єктів вибірки було розпізнано правильно, тим краще працює модель.

Однак з точки зору оцінки якості моделі інтерес представляє не тільки загальна кількість правильно і неправильно розпізнаних об'єктів, але і їхня кількість у кожному класі. Якщо в одному з класів кількість неправильно розпізнаних прикладів істотно більша, ніж в інших, є привід замислитися про те, чи правильно підібрані елементи в навчальній вибірці, і розробити відповідні заходи для оптимізації моделі. Для оперативної візуалізації кількості правильно і неправильно розпізнаних прикладів у кожному класі початкової вибірки слугує матриця класифікації.

Матриця класифікації показує, яка частка від загальної кількості прикладів розпізнана неправильно. Залежно від особливостей розв'язуваної задачі і необхідної точності допускається певний відсоток неправильно розпізнаних прикладів.

Матриця класифікації дозволяє визначити класи, для яких допущено найбільшу кількість помилок класифікації.

Інша найважливіша задача Data Mining – регресія. Формальна відмінність регресії від класифікації полягає в тому, що вихідна змінна є безперервною, тому використання матриць класифікації для оцінки якості регресійних моделей неприйнятне. Регресія

просто встановлює взаємний зв’язок між значеннями вхідних і вихідних змінних моделі.

Візуалізатором, який застосовується для оцінки якості моделей у разі безперервної вихідної змінної, є діаграма розсіювання. Діаграма розсіювання є графіком, по одній осі якого відкладаються цільові значення вихідної змінної, а по інший – реальні значення, отримані на виході моделі. На такому графіку можна побудувати лінію ідеальних значень. На цій лінії полягатиме будь-яка точка, для якої реальне вихідне значення, що сформоване моделлю, дорівнюватиме цільовому значенню. Така лінія, очевидно, виходитиме з початку координат і перетинатиме координатну площину по діагоналі.

Водночас будь-яка точка, для якої реальне вихідне значення буде відмінне від еталонного цільового значення, відхилиться від лінії ідеальних значень. При цьому величина відхилення дорівнюватиме помилці, яка допущена моделлю на цьому прикладі (рис. 2.39).

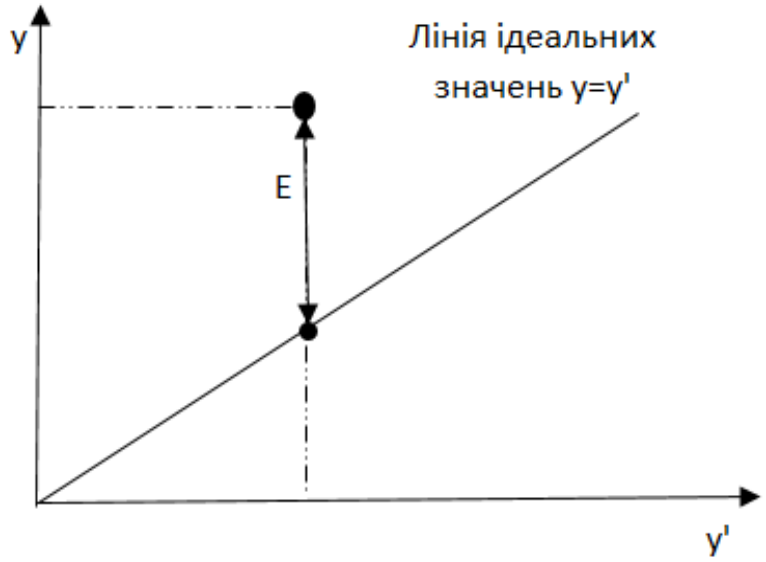


Рис. 2.39. Величина відхилення дорівнюватиме помилці, яка допущена моделлю на цьому прикладі

Приклад діаграми розсіювання представлений на рис. 2.40. Діагональна лінія на рисунку – це лінія ідеальних значень. Точками, розсіяними вздовж лінії ідеальних значень, позначені реальні вихідні значення моделі.

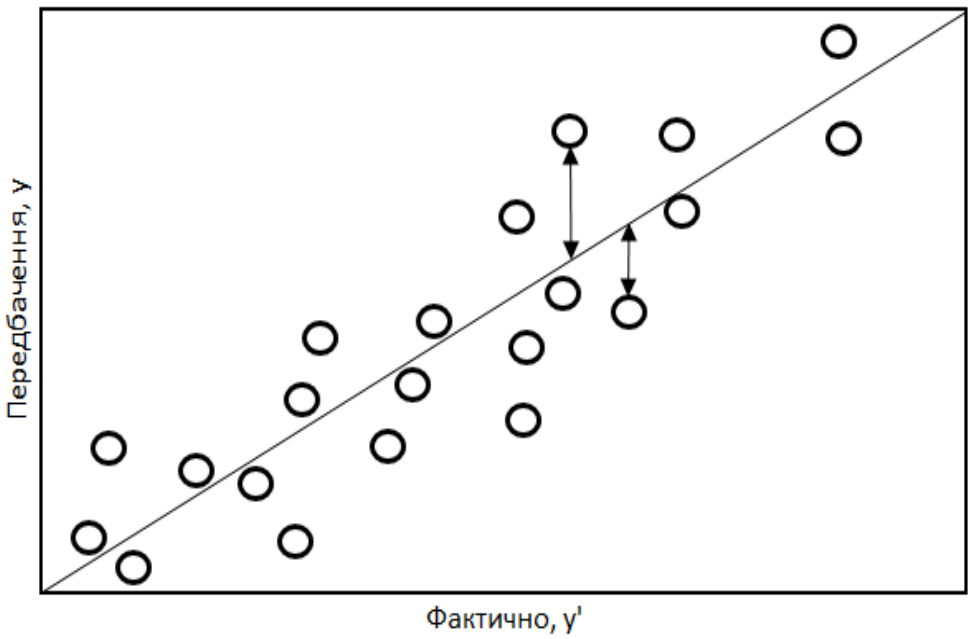


Рис. 2.40. Приклад діаграми розсіювання

Сенс діаграми розсіювання досить простий. Якщо всі точки (або хоча б основна їхня маса), що являють собою реальні вихідні значення моделі, зосереджені поблизу лінії ідеальних значень, то модель працює добре.

Якщо хмари, утворені точками виходів моделі, мають значний розкид, то більшість вихідних значень має велику помилку і в цьому випадку якість моделі невдала. Приклад такої діаграми розсіювання представлений на рис. 2.41.

Особливий інтерес можуть представляти окремі точки, найбільш далеко віддалені від лінії ідеальних значень. Їхня поява обумовлена тим, що на відповідних прикладах модель дає досить високу помилку, оскільки в цих прикладах містяться аномальні значення, нехарактерні для поведінки вибірки в цілому. Аналіз таких відхилень дозволяє виявляти рідкісні, але важливі події, що впливають на досліджуваний процес. Якщо ж аномальне значення – просто наслідок помилки, то його варто виключити з розгляду.

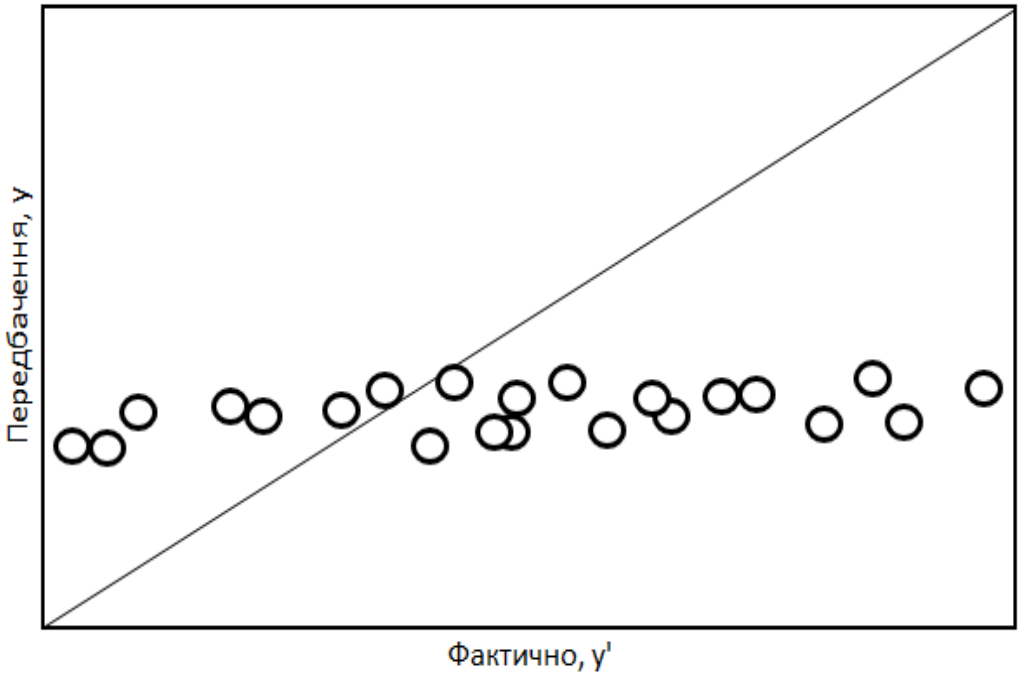


Рис. 2.41. Діаграма розсіювання для невдалої моделі

Ретропрогноз

Окремий випадок задачі регресії – прогнозування часових рядів. Щоб скласти прогноз, будується регресійна модель, яка на основі заданих значень ряду розраховує прогнозовані значення. Вона має деякий набір параметрів, що дозволяють отримати прогноз з урахуванням поведінки ряду в минулому (рис. 2.42).

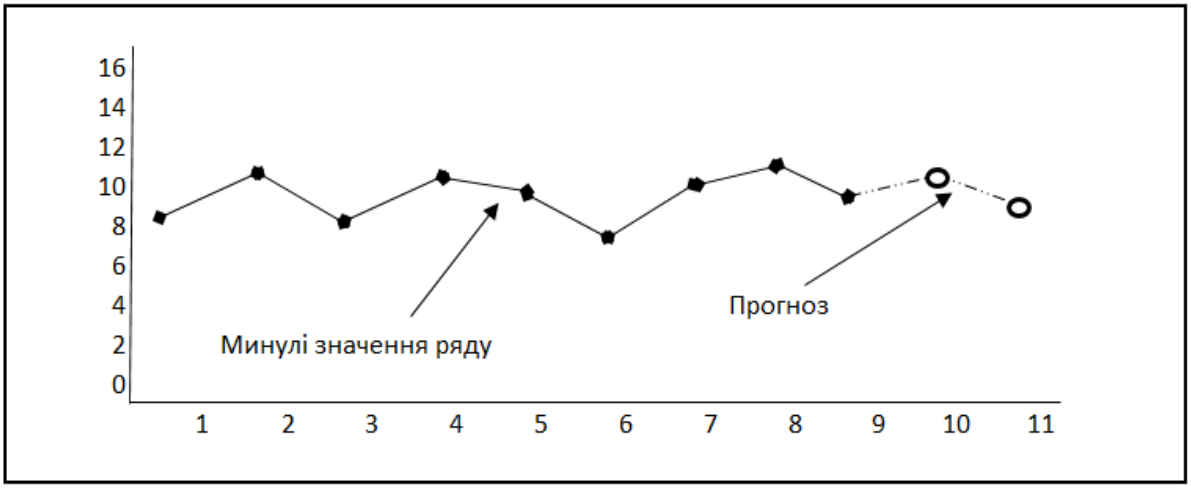


Рис. 2.42. Діаграма прогнозу

Для перевірки якості прогностичної моделі використовується спеціальний тип візуалізатора – ретропрогноз. Слово «ретропрогноз» означає «прогноз у минулому», «прогноз минулих значень». Його сутність полягає в тому, щоб застосувати побудовану модель до підмножини даних з минулого й оцінити, наскільки отриманий прогноз відповідає реальним значенням ряду, які мали місце в минулому.

Для побудови ретропрогнозу потрібно вибрати деяку підмножину даних з минулого, щоб використовувати їх як вихідні. Модель формує набір прогнозних значень, які потім порівнюються з даними, які реально мали місце в минулому. Якщо в результаті такого порівняння виявиться, що між значеннями ретропрогнозу і реальними даними існує велика розбіжність, то це дасть привід засумніватися в коректності моделі. Сам характер розбіжності (його величина, знак тощо) дозволяє виробити методику корекції моделі.

Принцип роботи ретропрогнозу пояснюється за допомогою діаграми на рис. 2.43.

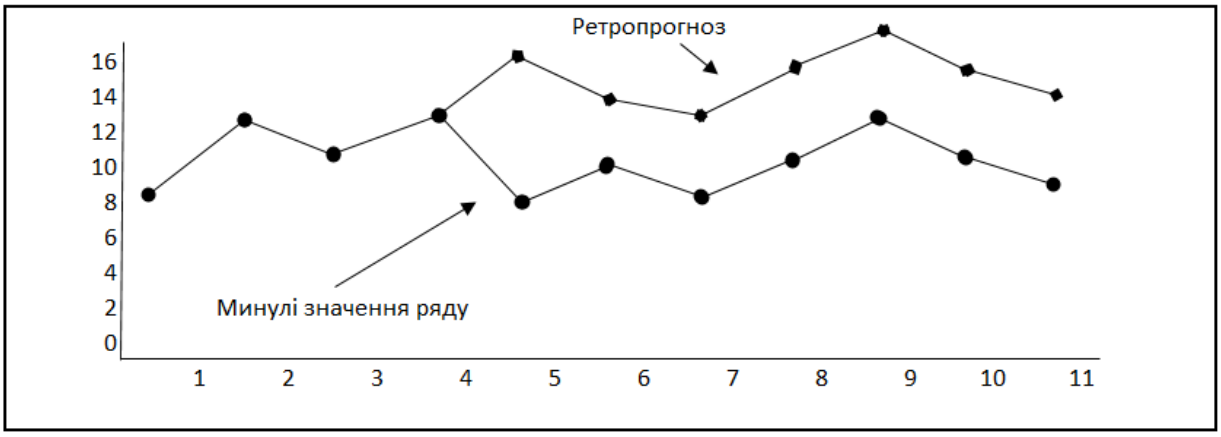


Рис. 2.43. Діаграма ретропрогнозу

На рис. 2.43 можна спостерігати добре узгодження значень ретропрогнозу і реальних значень ряду, що дозволяє припустити високу якість прогностичної моделі, за допомогою якої був побудований ретропрогноз. На рис. 2.44 представлений протилежний випадок: ретропрогноз виявляється істотно завищеним або заниженим, що свідчить про недостатню точність моделі.

Під час вибору множини значень часового ряду, які будуть використовуватися як вихідні дані для ретропрогнозу, варто

враховувати його актуальність. У часовому ряді можна виділити два види даних – актуальні і застарілі. Перші сформовані під дією поточних умов, які впливають на процес, який описується часовим рядом. Згодом умови можуть змінюватися, і тоді дані, сформовані під дією, що втратила чинність умов, виявляються застарілими, а дані, сформовані під дією нових умов, – актуальними. Зазвичай актуальними є найсвіжіші дані.

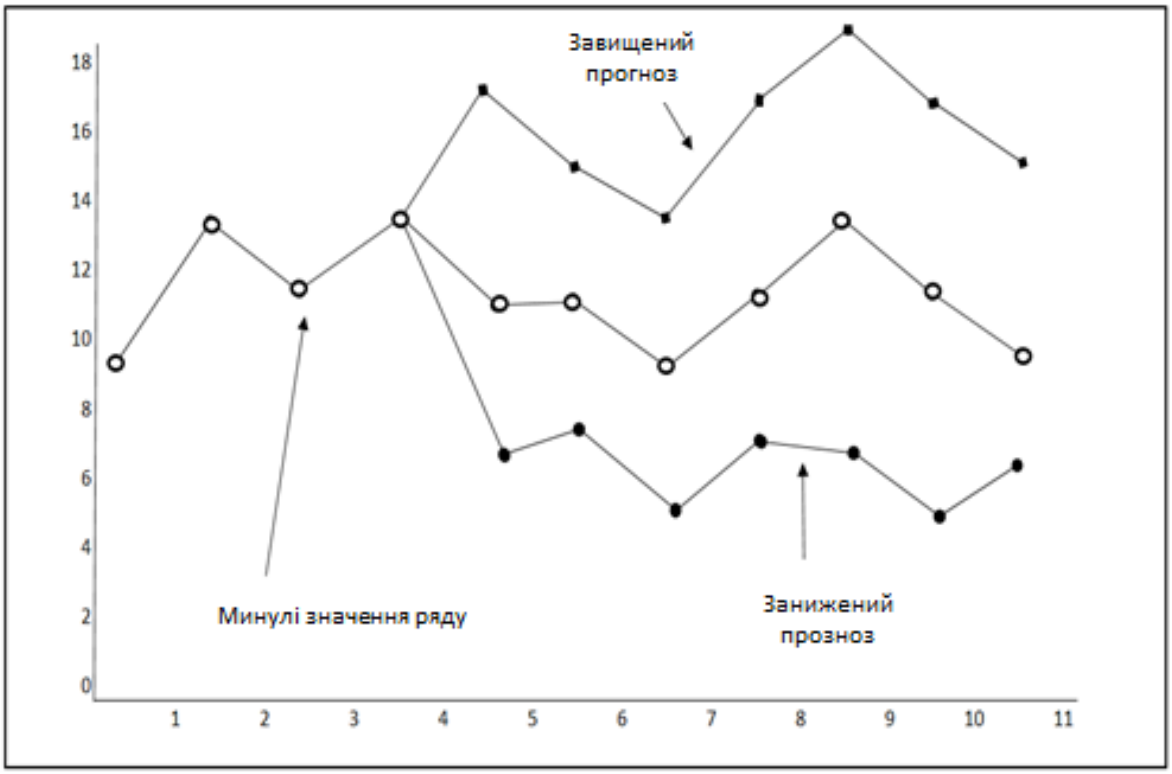


Рис. 2.44. Ретропрогноз з великим відхиленням прогнозованих значень від реальних

Якщо модель прогнозу була побудована на основі актуальних даних, то і ретропрогноз повинен будуватися на основі свіжих даних. Якщо ж ретропрогноз буде побудований на основі застарілих даних, то швидше за все його значення значно розходитимуться з реальністю. Але це буде пов’язано не з низькою якістю моделі, а з тим, що вона використовується для даних, що підкоряються іншим закономірностям, які в цій моделі не враховуються.

У бізнес-аналітиці широко використовуються моделі, засновані на машинному навчанні, наприклад нейронні мережі, дерева рішень тощо. У процесі настройки своїх параметрів модель набуває властивостей, які забезпечують виконання нею необхідного перетворення даних.

Керуючим параметром, на основі якого розраховується зміна стану моделі на кожній ітерації, є помилка неузгодженості між реальним вихідним значенням моделі, яка була визначена на цьому навчальному прикладі, і розрахованим значенням з цього прикладу (тобто між тим, що є, і тим, що повинно бути).

Помилка між реальним і розрахованим виходом, отримана в результаті навчання, – основний критерій оцінки якості навчання і точності моделі. Очевидно, що в міру того, як модель навчається, помилка зменшується, і як тільки вона досягає деякого заданого значення, процес навчання зупиняється.

Під час навчання моделі цікавим є спостереження за зміною вихідної помилки на кожній ітерації. За зміною помилки можна зробити висновок про те, як йде процес навчання, чи є сенс його продовжувати і які його передбачувані результати. Наприклад, якщо помилка швидко зменшується, це означає, що навчання просувається швидко й ефективно. Якщо помилка перестала зменшуватися і залишається однаковою на великій кількості ітерацій, можна припустити, що процес навчання «пробуксовує». У цьому випадку навряд чи має сенс продовжувати навчання, оскільки модель вичерпала свої можливості і налаштувати її точніше неможливо. Тому до складу аналітичних додатків часто включають спеціальний візуалізатор, який дає можливість спостерігати за динамікою зміни помилок. Можлива реалізація такого візуалізатора представлена на рис. 2.45.

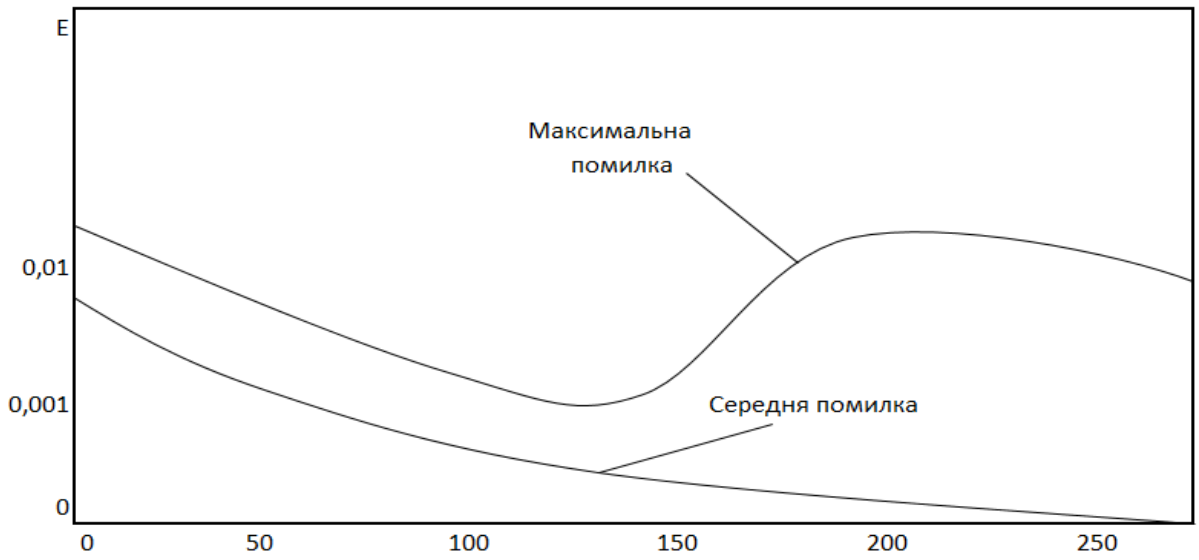


Рис. 2.45. Візуалізатор для спостереження за помилками у процесі навчання (*i*-номер ітерації)

8.4. Візуалізатори, що застосовуються для інтерпретації результатів аналізу

Як на вході, так і на виході аналітичного процесу присутні деякі дані. Дані на вході зазвичай називаються вихідними. Передбачається, що вони мають властивості, містять закономірності, структури й особливості, виявлення і розуміння яких дозволяє отримати нові знання про досліджуваний об'єкт. На виході також є дані, які називаються результатами. На їхній основі робляться висновки і приймаються рішення.

Між результатами аналізу та отриманими на їхній основі знаннями і рішеннями полягає процедура інтерпретації результатів, яка може виявитися не менш складною і проблемною, ніж сам аналіз. Неправильна інтерпретація може не тільки знецінити результати аналізу, але і призвести до помилкових висновків, що спричинить неправильні управлінські рішення. Саме тому в бізнес-аналітиці так багато уваги приділяється візуалізації результатів аналізу, що дозволяє зробити їхню інтерпретацію якомога ефективнішою і звести до мінімуму можливість помилкових висновків.

Крім візуалізаторів загального призначення, до складу аналітичних додатків включається набір спеціалізованих засобів візуалізації, які пов'язані з представленням результатів. Основна складність в інтерпретації результатів аналізу полягає в тому, що набори даних, як правило, мають велику розмірність.

Існує два основних способи опису даних. Перший полягає в побудові ієрархічної структури даних, тобто у разі необхідності поділу даних на класи, підкласи та ін. Другий – у скороченні розмірності вихідного простору даних. Розглянемо декілька найбільш поширених типів візуалізаторів, що застосовуються для інтерпретації результатів аналізу:

– деревовидні візуалізатори;

– візуалізатори зв'язків;

– карти.

Деревовидні візуалізатори можуть використовуватися скрізь, де результати аналітичної обробки даних утворюють ієрархічні рівні і можуть бути розділені на предків і нащадків.

Дерева. Однією з найбільш популярних класифікаційних моделей є дерева рішень. Вони являють собою ієрархічну послідовність правил виду «якщо ... то ...». Наприклад, класифікаційне правило може бути сформульоване таким чином: «Якщо значення ознаки > 10, то об'єкт належить до класу А». У процесі побудови моделі генерується ієрархічна послідовність правил, які застосовуються до досліджуваної вибірки і розподіляють усі вхідні в неї об'єкти за заздалегідь визначеними класами.

Для інтерпретації результатів класифікації за допомогою дерева рішень використовуються візуалізатори, що показують структуру дерева, а також сформульовані в ньому правила.

Цінність методу не тільки в тому, що він дозволяє класифікувати об'єкти, але і в тому, що він описує їхні властивості. Так, за деревом рішень можна оцінити ступінь подібності класів.

Застосування дерев для візуалізації асоціативних правил. Дерева рішень часто називають індуктивними, оскільки правила, з яких вони складаються, розбивають вихідну множину даних на підмножини, відповідні класам. Поряд з індуктивними правилами в Data Mining широко використовуються так звані асоціативні правила. При цьому на відміну від індуктивних правил, які слугують для розпізнавання об'єктів за належністю до певного класу, асоціативні правила встановлюють взаємний зв'язок між двома подіями.

Наприклад, якщо подією А є продаж товару X, а подією В – продаж товару Y, то асоціація між цими подіями читається так: «З А варто В». При цьому подія А називається умовою, а подія В – наслідком. Якщо асоціація між А і В є правилом, це означає, що якщо відбулася подія А, то з великою ймовірністю настане подія В. Чи є певна асоціація правилом, визначається шляхом аналізу їхньої досить великої кількості. Так, якщо було встановлено, що зі 100 транзакцій, в яких з'являлася подія А, 90 транзакцій містить подія В (тобто ймовірність спільної появи А і В буде 0,9), то таку асоціацію з високим ступенем вірогідності можна вважати правилом.

Основне застосування асоціативні правила знаходять у роздрібній торгівлі. Якщо відомо, що продаж товару X з великою ймовірністю тягне за собою продаж товару Y, то логічно клієнту, який купив товар X, запропонувати і товар Y.

Пошук асоціативних правил дозволяє виявляти групи товарів, які часто продаються разом, а це, в свою чергу, допомагає оптимізувати асортимент товарів, закупівлі, розміщення товарів у торговому залі, на вітрині тощо.

Для інтерпретації результатів пошуку асоціативних правил (як і індуктивних) часто використовують представлення у вигляді дерев. Але на відміну від індуктивних правил у деревах рішень асоціативні правила не використовують ієрархію, оскільки правила шукаються незалежно для всіх можливих асоціацій.

Один з можливих варіантів візуалізації асоціативних правил за допомогою дерева представлений на рис. 2.46. У кожному вузлі може бути передбачено виведення значень підтримки або достовірності пов'язаної з ним асоціації.

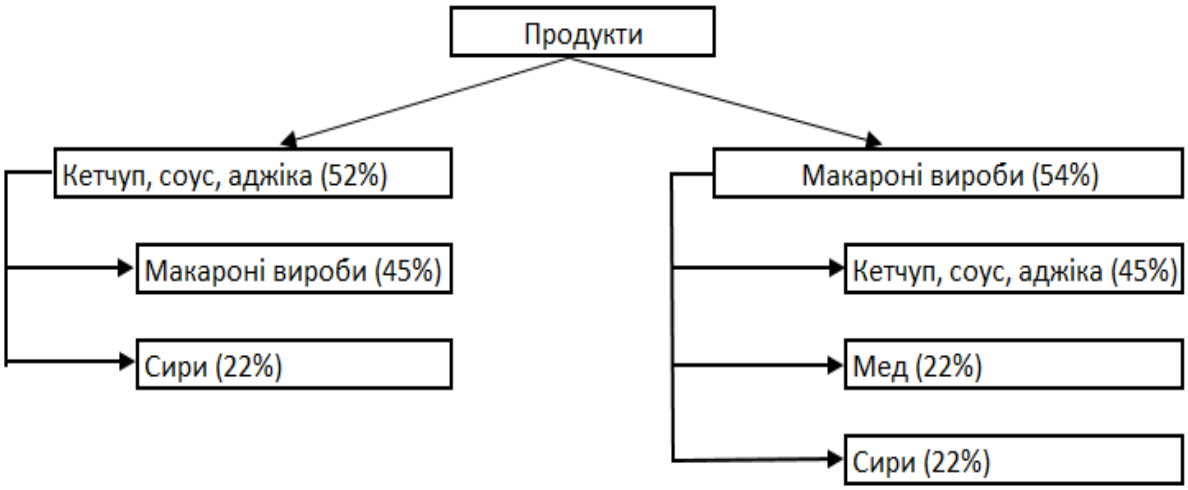


Рис. 2.46. Дерево асоціативних правил

Кластеризація. Однією з найважливіших задач Data Mining є кластеризація – об'єднання об'єктів у групи на основі подібності їхніх ознак. Такі групи називаються кластерами. Попадання двох об'єктів в один кластер дозволяє припустити високу ступінь подібності їх властивостей, і навпаки, якщо об'єкти в результаті кластеризації потрапили в різні кластери, то вони істотно відрізняються один від одного за своїми ознаками.

У результаті кластеризації деякої множини даних формується певна кількість кластерів. Одним з популярних видів кластеризації є ієрархічна кластеризація, яка, в свою чергу, містить два методи:

– агломеративну кластеризацію – кластери формуються шляхом злиття кластерів меншого розміру в кластери більшого розміру;

– розділову кластеризацію – кластери більшого розміру поділяються на декілька кластерів меншого розміру.

Для візуалізації результатів ієрархічної кластеризації використовують спеціальний різновид діаграм – так звана дендрограма. Дендрограма показує ступінь близькості окремих об’єктів і кластерів, а також наочно представляє в графічному вигляді послідовність їхнього об’єднання або поділу. Кількість рівнів дендрограми відповідає кількості кроків злиття або поділу кластерів. Приклад дендрограми представлений на рис. 2.47. У нижній частині рисунка розташована шкала, на якій відкладається відстань між об’єктами в просторі ознак.

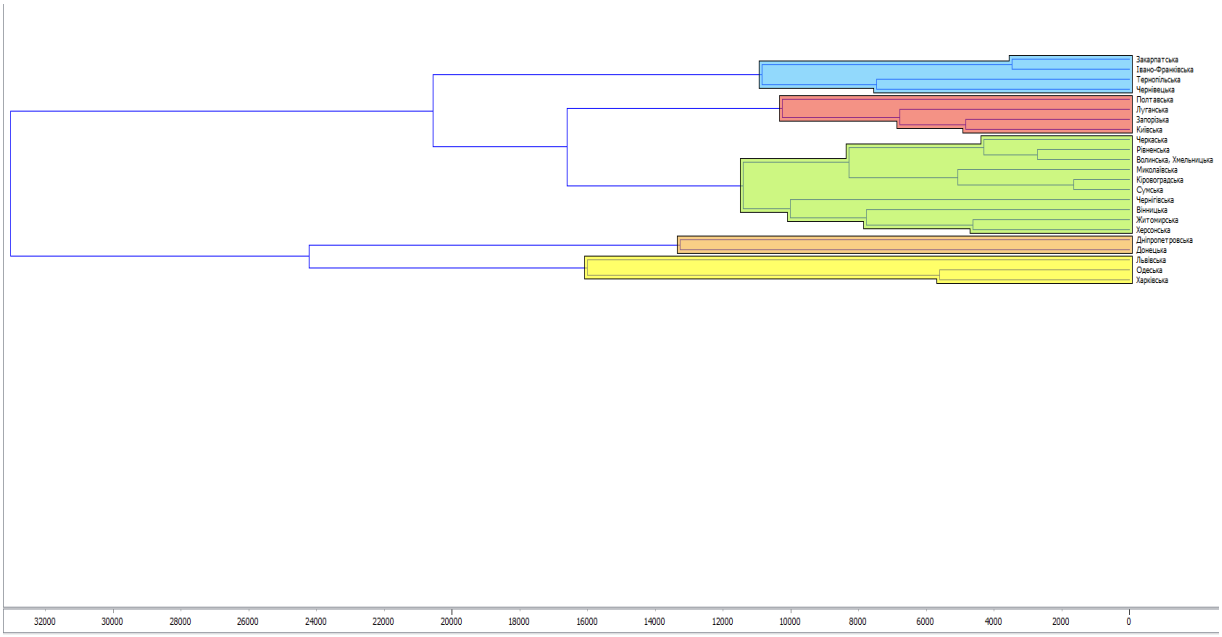


Рис. 2.47. Приклад дендрограми

Візуалізація зв’язків. Нерідко потрібно досліджувати характер і ступінь взаємної залежності між різними об’єктами. Для аналізу можна використовувати візуалізацію зв’язків, коли

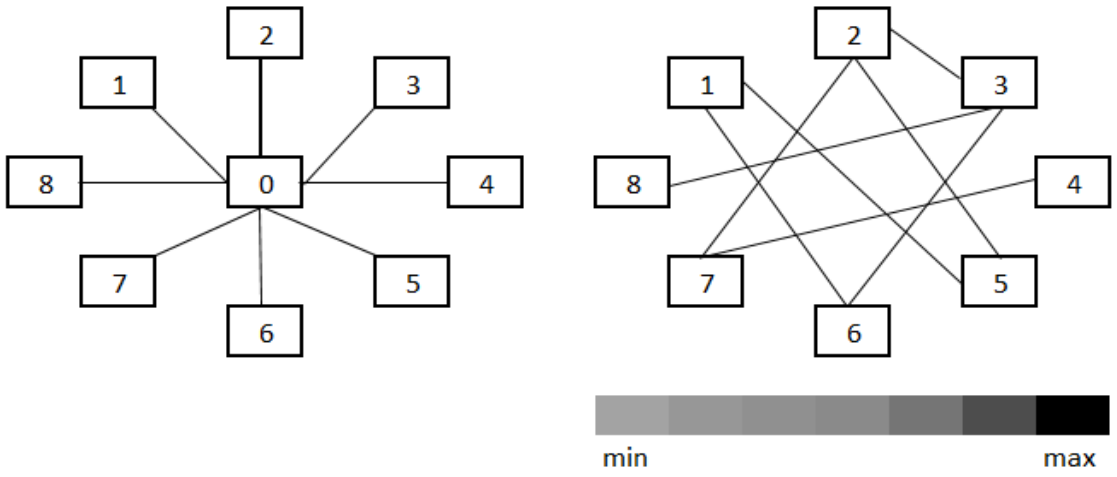
об’єкти представляються у вигляді деяких значків, а зв’язки між ними – у вигляді ліній, що з’єднують відповідні значки. При цьому сила зв’язку, тобто ступінь взаємної залежності об’єктів, може показуватися різними способами. Найчастіше для цього використовують:

– товщину лінії: чим сильніше зв’язок між об’єктами, тим товстіша з’єднує їх лінія (рис. 2.48, а);

– колір лінії, при цьому обираються відтінки певного спектра (рис. 2.48, б).

У принципі допускається одночасно використовувати і товщину, і колір, хоча такі діаграми більш складні для сприйняття.

Крім того, можна по-різному розташовувати об’єкти на площині. У першому випадку показуються зв’язки між одним обраним об’єктом, який міститься в центрі, і всіма іншими (рис. 2.48, а). У другому випадку візуалізуються всі попарні зв’язки між об’єктами (рис. 2.48, б). Якщо кількість об’єктів значна, зв’язків стає ще більше, тому аналізують тільки найбільш сильні або найбільш слабкі, накладаючи фільтр на значення сили зв’язку.



а) б)

Рис. 2.48. Варіанти діаграм зв’язків між об’єктами: а) «зірка»; б) «мережа»

Об’єкти порівнюються між собою на основі будь-якої ознаки. Якщо ознак декілька, вибирається спосіб розрахунку критерію зв’язку.

Карти. Карти дозволяють наочно представити дані, пов'язані з географічним розташуванням досліджуваних об'єктів і процесів. Це можуть бути демографічні дані (наприклад, розподіл показників смертності або народжуваності за різними регіонами), а також дані, що відображають міграцію населення і робочої сили, забезпечення регіонів енергоносіями, динаміку поширення епідемій та ін. У бізнес-аналітиці це – інформація, пов'язана з рівнем споживання в регіонах, характером попиту і пропозиції за різними видами товарів, відомості про продажі, що здійснюються регіональними дилерами, про логістичні і транспортні потоки тощо.

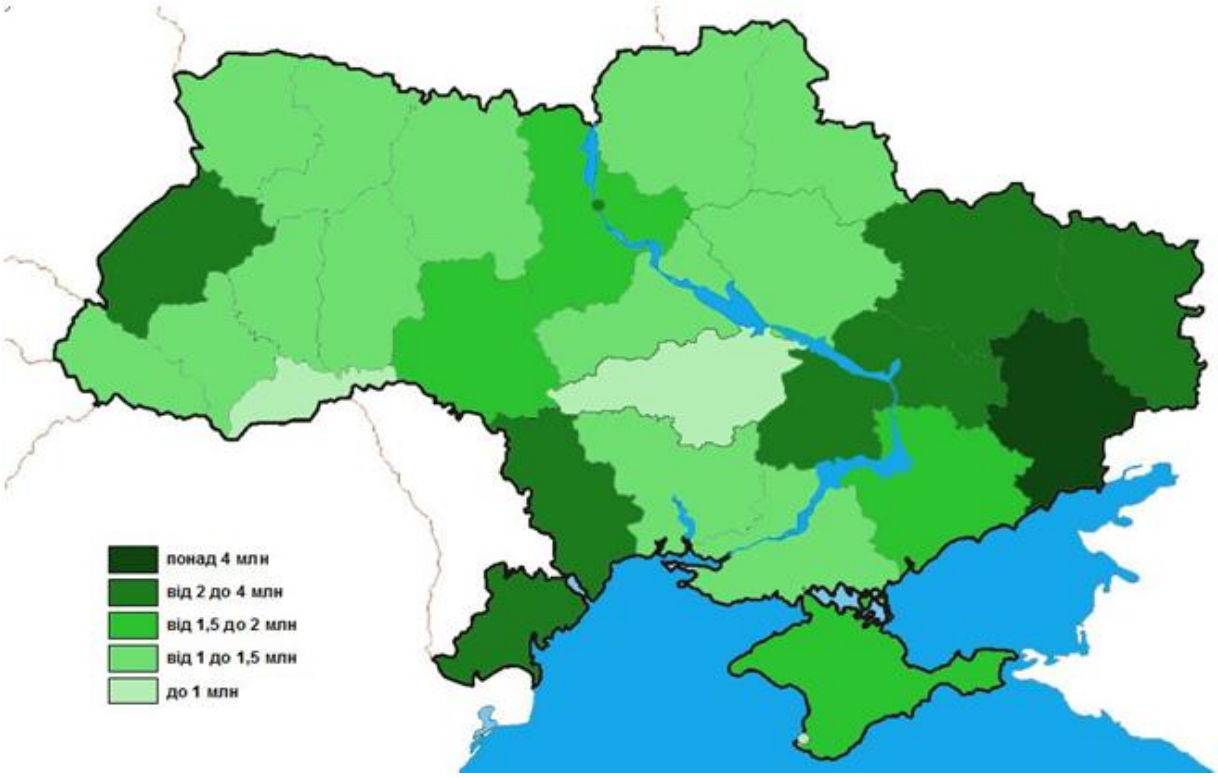


Рис. 2.49. Демографічна карта України

При цьому може використовуватися багатовимірне представлення даних. Наприклад, якщо потрібно показати на карті не тільки суми продажів, але і прибуток, то останню можна відобразити рельєфним виділенням відповідної області на карті. Також на карти можна наносити значки, форми, колір, розмір і взаємне положення яких відповідають властивостям досліджуваних об'єктів.

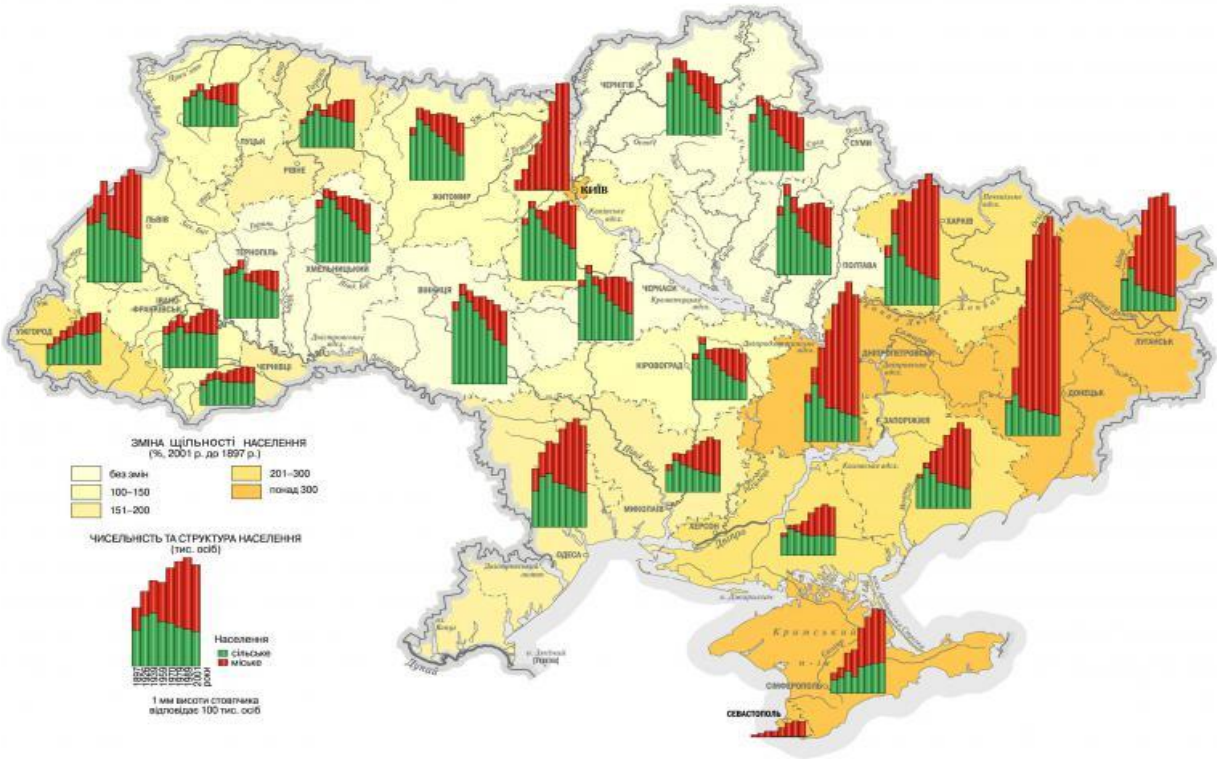


Рис. 2.50. Зміна темпів щільності, чисельності та структури населення України за період 1897–2001 рр.[24]

Втім карти не обов’язково повинні бути пов’язані з географією. У бізнес-аналітиці майже завжди доводиться мати справу з об’єктами, які описуються двома ознаками і більше. Тобто вибірки, утворені такими об’єктами, є багатовимірними. У цьому випадку може виникнути проблема з візуалізацією результатів, оскільки уявлення багатовимірних об’єктів на плоских візуалізаторах (графіках, діаграмах) не завжди зручно і коректно відображає результати. Тому актуальні візуалізатори, що дозволяють адекватно представляти багатовимірні дані. Зокрема, поширення набули двовимірні теплові карти, де кожному значенню ознаки відповідає один з відтінків у заздалегідь обраній кольоровій гамі (рис. 2.51).

[24] Бочковська А. І., Руденко Л. Г. Карти Національного атласу України як інформаційна база для дослідження формування чисельності населення. URL : https://igu.org.ua/uk/node/21095

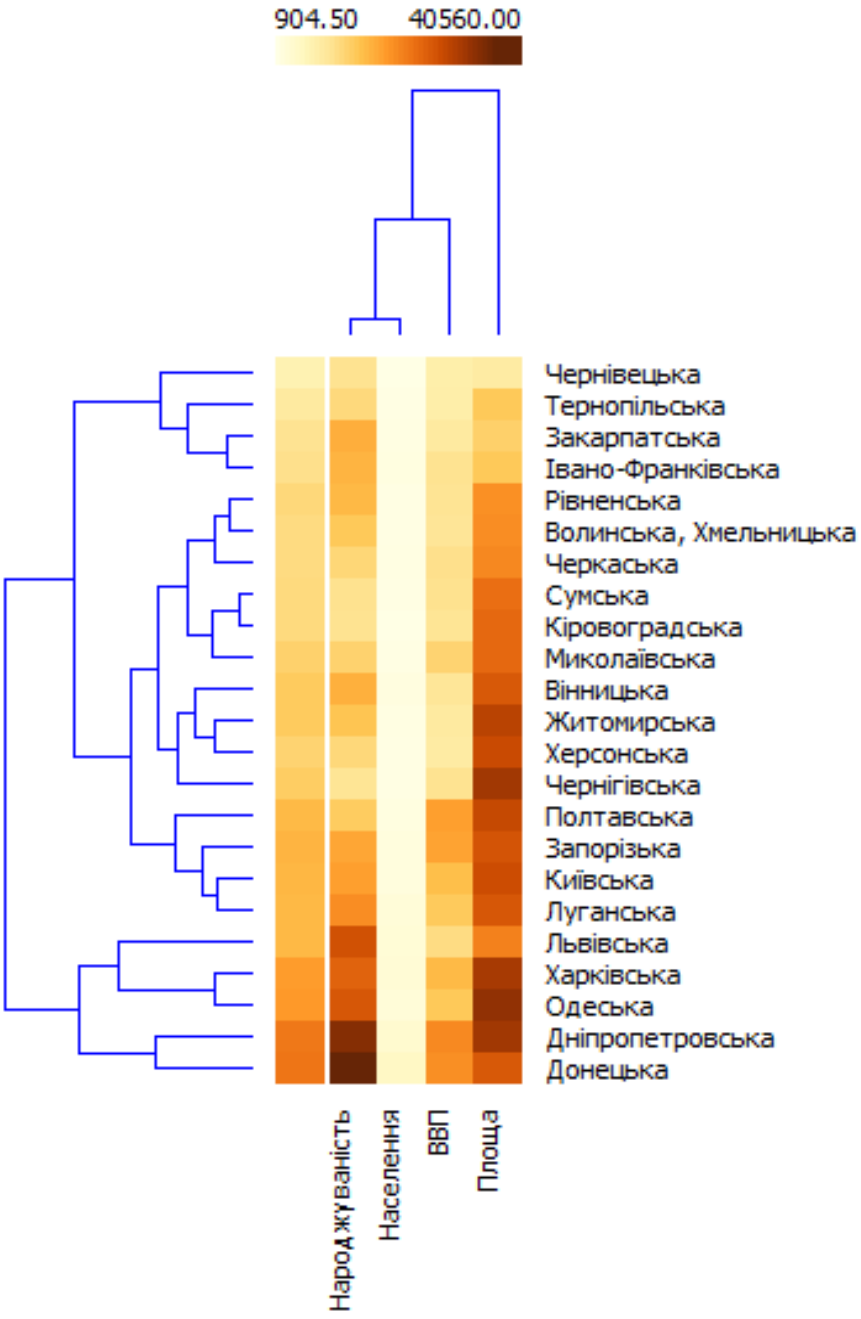


Рис. 2.51. Приклад теплової карти

До різновиду теплових карт відносять карти Кохонена. Принцип цього методу візуалізації заснований на тому, що багатовимірні дані представляються у вигляді плоских карт, у процесі побудови яких комбінується інформація про відстані між векторами об’єктів у багатовимірному просторі ознак і розфарбування залежно від значень ознак.

Запитання до параграфа 8

1. Що таке візуалізація даних?
2. Які є типи візуалізації даних?
3. Яка мета візуалізації даних?
4. Які переваги візуалізації даних?
5. Які цілі і завдання візуалізації на різних етапах аналітичного процесу?
6. Які є групи методів візуалізації?
7. Який є основний набір засобів візуалізації?
8. Що таке графіки?
9. Що таке діаграми?
10. Що таке гістограми?
11. Що таке деталізація даних?
12. Що таке візуалізатори для оцінки якості моделі?
13. Які є складові якості моделі?
14. Яка сутність ітераційного характеру моделювання?
15. Який типовий набір візуалізаторів для оцінки якості моделей?
16. Що таке візуалізатори, що застосовуються для інтерпретації результатів аналізу?
17. Які є візуалізатори, що застосовуються для інтерпретації результатів аналізу?
18. Які є варіанти діаграм зв'язків між об'єктами?
19. Для чого використовуються карти під час аналізу даних?
20. Що таке теплові карти?

§ 9. Очищення і попередня обробка даних

9.1. Оцінка якості даних

Якість даних, які збираються і консолідуються для аналізу з різних джерел, є однією з найбільших проблем бізнес-аналітики. Недостатня увага до цього питання може звести нанівець усі зусилля аналітиків і експертів із розробки аналітичних рішень. Аналітичні рішення, отримані на основі неякісних даних, можуть виявитися як завгодно далекими від дійсності, спотворити справжню картину досліджуваних бізнес-процесів, показати помилкові закономірності,

тенденції та зв'язки між бізнес-об'єктами. Як наслідок – розробка неправильних управлінських рішень. Тому моніторинг якості даних здійснюється на всіх етапах аналітичного процесу. Якщо важливість цього усвідомлюється й аналітичним технологіям на підприємстві приділяється належна увага, то контроль і підтримка якості даних проводиться вже на етапі збору даних у самих джерелах: облікових системах, корпоративних БД підприємства тощо.

Поняття «якість даних» досить широке та не має чіткого визначення. Воно може трактуватися дещо по-різному залежно від того, в якій сфері інформаційних технологій воно вживається. Так, спочатку під якістю даних розуміли кількість помилок під час введення і форматування даних. У контексті сучасних аналітичних технологій якість даних – сукупність їхніх властивостей і характеристик, що визначають ступінь придатності для аналізу.

У загальному якість інформації – сукупність властивостей, що відображають ступінь придатності конкретної інформації або даних про об'єкти та їхній взаємозв'язок для досягнення цілей, що стоять перед користувачем.

Також під якістю даних розуміють процес компонування даних (включаючи імена ресурсів, схеми даних, особливості доступу й опрацювання тощо), очищення та вдосконалення даних, а також об'єднання з усуненням дублювання та невизначеності[25]. Проте це визначення є невдалим, оскільки воно свідчить не про властивість даних, а про процес їхньої обробки (хоча і з метою підвищення якості даних). Якість, як і корисність даних для певного користувача чи групи користувачів, залежить також від джерела даних і ступеня довіри до нього.

Для підвищення якості даних використовується комплекс методів і алгоритмів, які отримали назву «очищення даних». Щоб правильно підготувати дані до аналізу, необхідно мати стратегію їхнього очищення. Вона розробляється на основі знання структури й особливостей джерел, з яких надходять дані, характеру самих даних, методики і цілі їхнього аналізу.

---

[25] Шаховська Н. Б., Пшеничний О. Ю., Чорней І. М. Проблеми якості консолідованих даних у просторах даних. Системи обробки інформації, 2011. Випуск 3 (93). С. 80–84.

Перед застосуванням певних методів очищення необхідно виконати оцінку якості даних з метою виявлення найбільш характерних проблем і рівня їхньої складності, а також для вироблення стратегії подальшого очищення. У результаті такої оцінки може бути прийнято рішення про відмову від очищення деяких даних, оскільки це занадто складно, дорого чи не приведе до бажаного результату.

Щоб уникнути ситуації, коли дані стали настільки «брудними», що їх аналіз став неможливим з технічних причин або через низьку достовірність отриманих результатів, необхідно підходити до проблеми якості даних комплексно, а саме розробити стратегію збору, консолідації та аналізу даних паралельно з оцінкою та моніторингом їхньої якості, виявленням факторів, що впливають на їхню якість, і виробленням стратегії виключення таких факторів.

У результаті розвитку комплексного підходу до підтримки якості даних на підприємстві в IT-технологіях з'явилася нова дисципліна – управління якістю даних на підприємстві (Enterprise Data Quality Management, EDQM). Розвиток EDQM стимулювався усвідомленням того факту, що багато інноваційних проєктів зазнало невдачі саме через низьку якість даних і неможливість розробити на їхній основі правильні управлінські рішення. EDQM де-факто стало частиною загального процесу управління якістю на підприємстві (Total Quality Management, TQM). Ключовою ланкою EDQM є оцінка якості даних, яка може бути реалізована трьома способами:

1. Одноразова оцінка якості. Проводиться для деякого масиву даних (файла або таблиці) один раз, після чого, залежно від результату, до даних застосовуються певні методи очищення. Іноді дані визнаються якісними і залишаються без змін.

2. Моніторинг. Потік даних, що надходять з різних джерел в аналітичну систему, безперервно сканується спеціальними програмами пошуку помилок у даних.

3. Візуальна оцінка. Використовується у виняткових випадках, коли спеціалізованим програмних засобам не вдалося підвищити якість даних до прийнятного рівня. Найчастіше застосовується під час попередньої обробки даних, завантажених в аналітичну

систему. У цій ситуації аналітик за допомогою наявних засобів візуалізації (таблиць, графіків і діаграм) визначає наявність у даних аномальних значень, пропусків, шумів тощо. Цей метод оцінки якості застосовується лише до невеликих за обсягами даних у разі вирішення окремих нетипових завдань.

У процесі розробки методики оцінки якості даних також необхідно вирішити, на якому етапі, де саме її потрібно робити. Існують такі варіанти:

1. Безпосередньо в джерелах даних. Можна ефективно виконувати пошук: орфографічних помилок, пропущених, аномальних, логічно неправильних і фіктивних значень, протиріч і дублікатів на рівні записів і таблиць. Перевага цього варіанта в тому, що за результатами оцінки відповідні методи очищення можуть бути задіяні вже на етапі ETL і в БД надійдуть очищені дані. При цьому варто враховувати, що в процесі ETL якість даних може знову погіршитися: з'являться нові дублікати і протиріччя, невідповідності форматів тощо. Це відбувається, тому що в процесі ETL здійснюється інтегрування даних з декількох окремих джерел.

2. У процесі ETL. Тут можна виявляти проблеми в даних і за результатами оцінювання якості оперативно задіяти певні методи очищення. Тоді буде забезпечено завантаження в сховище вже очищених даних. Недолік цього варіанта – збільшення навантаження на інформаційну систему.

3. В аналітичній системі. Використовується в процесі попередньої обробки даних перед застосуванням різних методів Data Mining. Оцінка може проводитися аналітиком візуально за допомогою таблиць, графіків і діаграм або на основі статистичних оцінок і характеристик.

У джерелах даних зручно оцінювати наявність помилок, пов'язаних з введенням даних. У процесі ETL виявляються чинники, пов'язані з порушеннями структури, цілісності і повноти даних, оскільки вони просто не дозволять завантажити дані в СД. В аналітичній системі дані оцінюються з точки зору їхньої відповідності конкретному завданню Data Mining і можливості застосування певних алгоритмів аналізу.

Мета оцінки якості даних – виявлення проблем у даних, визначення ступеня їхньої «забрудненості» і придатності для аналізу. Виявленням джерел проблем, і боротьбою з ними займається інша дисципліна – очищення даних.

Базовими характеристиками якості даних є: функціональна придатність до використання, коректність або достовірність, ресурсна економічність, практичність, супроводжувальність, мобільність. Щоб систематизувати проблеми, пов'язані з якістю даних, і визначити основні аспекти оцінки якості, виділяють декілька рівнів. На кожному з них існують свої специфічні проблеми, цілі, методи і критерії оцінки якості даних.

1. Технічний рівень. На якість даних впливають чинники, пов'язані з порушенням їхньої структури, цілісності та повноти, некоректні формати і способи представлення даних. Проблеми технічного плану заважають коректно виконати консолідацію та інтегрування даних, завантажити їх у сховище, застосувати до них певні алгоритми обробки.

2. Аналітичний рівень. Передбачає наявність факторів, що заважають виконувати коректний аналіз даних і отримати достовірні результати. До таких факторів належать: шуми в даних, аномальні значення, суперечливі та дублюючі записи, фіктивні значення, пропуски тощо. Для цього рівня характерна суб'єктивність оцінки якості даних. Так, наприклад, те, що для одного аналітика буде шумом, для іншого – цінні з точки зору аналізу дані. Що стосується аномальних значень, то потрібно зазначити, що певна аномалія може бути помилкою оператора або реальною подією. Виключивши таке дане, можемо втратити важливу інформацію. Розглядаючи два однакові записи, не можна зазначити, дублікати це чи інформація про дві ідентичні події. Із суперечливими записами теж не все так просто.

3. Концептуальний рівень. Відображає проблеми, пов'язані з тим, що стратегія збору даних на підприємстві була спочатку неправильною. Зібрані дані не містять достатньо інформації для всебічного опису предметної сфери та основних бізнес-процесів. Наприклад, у результаті аналізу може бути виявлено, що даних мало або дані не ті, які потрібні для розробки управлінського рішення.

Порівняльна характеристика різних рівнів якості даних наведена в табл. 2.16.

Таблиця 2.16

Рівні якості даних

Рівень	Фактори	Прояв	Місце боротьби
Технічний	Порушення в структурі даних; некоректне найменування таблиць і полів; некоректні формати і кодування даних; порушення повноти і цілісності даних; протиріччя і дублікати на рівні таблиць і файлів БД	Заважають інтегруванню даних, їх завантаженні в СД і аналітичні системи	ETL
Аналітичний	Пропуски; аномальні і фіктивні значення, друкарські помилки; шуми; протиріччя і дублікати на рівні записів	Знижують достовірність даних і спотворюють результати їхнього аналізу, не дозволяють використовувати деякі аналітичні методи	Джерела даних, ETL, аналітичні системи
Концептуальний	Зібрані і консолідовані дані в недостатній мірі відображають досліджувані бізнес-процеси	Відсутні або не достатньо даних для аналізу	Стадія розробки аналітичних проєктів і процесів

Отже, оцінка якості даних – це складне і трудомістке завдання, яке часом не піддається формалізації. Вона вимагає залучення всіх, хто бере участь у процесі збору та аналізу даних, починаючи від операторів, які здійснюють введення даних в OLTP, і закінчуючи остаточними користувачами.

9.2. Оцінка придатності даних до аналізу

Оцінка якості даних передбачає дві мети:

– контроль їхньої придатності до аналізу;

– виявлення проблем у функціонуванні систем збору і консолідації даних.

Обидві мети взаємопов'язані. Якість даних, що консолідуються для аналізу, в більшості випадків є індикатором якості функціонування відповідних інформаційних систем як на апаратному, так і програмному рівнях. Якщо стабільно поставляються якісні дані незалежно від їхнього походження і характеру, система їхнього збору та консолідації функціонує на високому організаційно-технічному рівні. Якщо якість даних стабільно низька, то, можливо, не лише їх очищенню приділяється мало уваги, а існує проблема, пов'язана з функціонуванням OLTP-систем, або існують прорахунки в процесі розробки ETL тощо.

Функціональна придатність визначається, у першу чергу, повнотою накопичених об'єктів – відносною кількістю об'єктів або документів, наявних у джерелах даних, до загальної кількості об'єктів, що потрапили у локальне сховище, яке містить консолідовані дані. Проте пріоритетною метою контролю якості даних є оцінка ступеня їхньої придатності до аналізу. За результатами аудиту якості даних можуть бути зроблені такі висновки:

1. Дані повністю придатні до аналізу і не потребують очищення. Це рідкісна ситуація, яка вказує на досить якісну та продуману стратегію збору і консолідації даних як на логічному, так і технічному рівнях.

2. Дані придатні до аналізу без очищення, але з певними обмеженнями. Такі дані містять помилки, які можуть вплинути на достовірність результатів аналізу, тому аналіз можливий, але з урахуванням ймовірного впливу помилок.

3. Дані для аналізу потребують очищення. Проблеми в даних не є критичними, тому після застосування відповідних процедур очищення якість даних може бути піднятою до прийнятного рівня.

4. Дані абсолютно непридатні до аналізу, і ніякі методи очищення ситуацію не виправлять.

Оцінка якості даних щодо їхнього походження. Приступаючи до оцінки якості даних, по-перше, необхідно враховувати інформацію про походження даних. Виділяють два типи джерел:

– джерела, які автоматично підтримують певну структуру, цілісність і несуперечність даних, OLTP-системи, корпоративні БД, деякі облікові системи та ін.;

– локальні файли окремих користувачів, де не передбачена підтримка структури, цілісності та несуперечності даних.

Під час використання даних, що надійшли з джерел першого типу, потрібно бути готовими до проблем, пов'язаних з помилками введення, пропусками, фіктивними й аномальними значеннями тощо. Що стосується джерел другого типу, то в них можна зустріти фактори технічного рівня, пов'язані з порушенням структури даних, використанням некоректних форматів, протиріччя тощо.

По-друге, необхідно враховувати, чи організований процес збору даних на підприємстві з прицілом на їх майбутній аналіз. Якщо цьому приділялася належна увага, то можна очікувати прийнятної якості даних, в іншому випадку – ні.

Коротко охарактеризуємо інші базові характеристики якості даних.

Коректність або достовірність даних – це ступінь відповідності даних про об'єкти в базах даних реальним об'єктам у цей момент часу, що визначається змінами самих об'єктів, некоректних записів про їх стан або некоректних розрахунків їхніх характеристик.

Ресурсна економічність відображається зайнятістю ресурсів центрального процесора, оперативної, зовнішньої та віртуальної пам'яті, каналів введення-виведення, терміналів і каналів зв'язку. Існують розроблені методи (наприклад, метод критичних робіт) та засоби визначення завантаженості ресурсів.

Практичність – визначає функціональну придатність і корисність застосування консолідованих даних для певних користувачів. У просторі даних оцінка практичності здійснюватиметься за допомогою функції корисності прийнятих рішень.

Супроводжуваність даних відображається зручністю й ефективністю виправлення, удосконалення або адаптації структури та змісту описів даних залежно від змін у зовнішньому середовищі застосування, а також у вимогах і функціональних специфікаціях замовника. Для оцінки супроводжуваності розроблені методи та засоби.

Мобільність характеризується тривалістю і трудомісткістю їхньої інсталяції, адаптації та заміщення під час перенесення на інші апаратні й операційні платформи. У просторах даних характеристика мобільності пов'язана зі зміною даних про джерела даних у каталозі.

Отже, оцінка якості даних – необхідний етап процесу підготовки до аналізу. Вона дає можливість своєчасно виявляти проблеми, які не дозволяють коректно аналізувати дані, знижують значимість і достовірність результатів аналізу. Контроль якості даних повинен проводитися на всіх етапах аналітичного процесу, оскільки для кожного етапу характерні свої проблеми.

9.3. Технології та методи оцінки якості даних

Нині технології оцінки якості даних широко використовуються не тільки в процесі аналізу даних, а й в інформаційних системах різноманітного. Розглянемо декілька практичних аспектів оцінки якості даних.

Одним з найбільш поширених методів оцінки якості даних і виявлення проблем є профайлінг. Він виконується на основі аналізу метаданих, що описують структуру даних. У процесі профайлінга аналізується така інформація.

Тип атрибута (поля). Якщо тип атрибута визначено як числовий, а у ході перевірки виявлено, що він має інший тип (строковий або дата / час), то з'ясовуються причини і проводиться, відповідно, перетворення даних.

Довжина значення. Визначається максимально допустима кількість символів у значенні атрибута, і якщо довжина значення перевищує максимальну, необхідно вживати заходи до її обмеження.

Дискретні значення. Зазвичай перевіряється частота їх появи й унікальність (у процентному або абсолютному вигляді).

Діапазон допустимих значень. Задаються мінімальне і максимальне значення, які може приймати атрибут. Якщо значення атрибута виходить за обмеження, то воно може змінюватися на деяке значення за замовчуванням. Наприклад, можна обмежити вік клієнта 18 і 70 роками, тоді будь-яке значення, що не потрапило в цей діапазон, буде визначено як неприпустиме і змінено на значення за замовчуванням.

Аналіз строкових шаблонів (наприклад, поштових індексів або номерів карт соціального страхування). Якщо рядок в осередку відповідає шаблону (наприклад, XXX-XXX-XXX), то помилки немає. В іншому випадку включається обробка рядка відповідно до деякого заздалегідь визначеного.

Профайлінг – досить розвинений метод виявлення проблем у даних. Проте він розрахований на аналіз «штатних», технічних проблем, які належать до найбільш типових і добре формалізованих. Тому це – метод попередньої оцінки якості даних з метою виявлення типових проблем.

Візуальна оцінка якості даних. Коли обсяг даних не досить великий, то для оцінки їхньої якості можна застосовувати візуальні методи. Для цього використовуються як вбудовані засоби візуалізації аналітичної платформи, так і сторонні програмні засоби, в яких можна переглядати дані в табличній і графічній формах, розраховувати статистичні характеристики.

Оцінка якості даних за допомогою таблиць. Табличне представлення дозволяє робити загальні висновки про наявність у даних структурних порушень, пропусків, аномальних і фіктивних значень. Погляду на таблицю здебільшого досить, щоб оцінити кількість неповних рядків і стовпців, порожніх клітинок. Аномальні числові значення можуть різко виділятися на фоні навколишніх даних кількістю цифр.

На загальному тлі добре виділяються невідповідності типів даних в одному стовпці. Так, наприклад, числові дані зазвичай вирівнюються по правому краю комірки таблиці, а строкові дані – по лівому краю. Тому якщо у стовпці з числовими даними виявиться строкове значення, його можна легко виявити щодо вирівнювання.

Важче виявити такі проблеми, як дублікати і протиріччя, оскільки на фоні оточуючих даних вони виглядають цілком правдоподібно. У деяких випадках дублюючі записи вдається виявити шляхом сортування за зростанням або спаданням, коли ідентичні записи виявляються поруч.

Оцінка якості даних за допомогою графіків. Графіки і діаграми – ефективні засоби для виявлення проблем у даних. Наприклад, за допомогою графіків можна легко оцінити наявність шумів у даних, а також пропущених і аномальних значень. Так, наприклад, аномальні значення будуть відображатися у вигляді різких відхилень у певний бік щодо основної тенденції. Наявність проблем у даних можна виявити за допомогою гістограм розподілу числових значень, наприклад сум продажів. Поява у гістограмі додаткових піків (мод) може свідчити про присутність нехарактерних для цього ряду значень, які можуть бути аномаліями, помилками та ін.

Виявлення важко формалізованих помилок. Існує безліч факторів, що знижують якість даних і не піддаються формальній оцінці. Один з них – людський фактор. Сотні і тисячі працівників безперервно формують потоки «брудних» даних. При цьому елементарні помилки введення і реєстрації даних здатні призвести до великих спотворень інформації. З цим можна «боротися», наприклад, організувавши процес реєстрації так, щоб мінімізувати ручне введення, коли найменування клієнтів і товарів вибираються зі списку. Це допоможе домогтися деякого поліпшення якості даних, проте не обов'язково вирішить проблему кардинально.

Для прикладу розглянемо проблему накопичення «брудних» даних у соціальних мережах. В одному російськомовному інтернет-проєкті було дозволено користувачам «створювати» міста, вулиці, інститути та інші об'єкти. У результаті упродовж кількох місяців база даних виявилася засмічена дублікатами, помилками і різними варіантами найменувань одних і тих самих об'єктів. Так, в одному і тому ж місті були присутні «школа 17» і одночасно «шк. № 17»,« середня школа № 17», що насправді означало один і той самий об'єкт.

Основне завдання оцінки якості даних полягає в тому, щоб відрізнити «брудні» дані від «чистих», тобто виявити, які дані містять помилки й оцінити, яку частку «брудні» дані становлять загальним обсягом даних. Для глибокого аналізу причин «забруднення» даних, як правило, потрібні методи, засновані на знаннях про те, якими мають бути якісні дані. Такі знання містять правила і шаблони, яким повинні відповідати якісні дані. Чим більше даних не відповідає шаблонах і правилам, тим сильніше вони «забруднені» і більше уваги потрібно приділяти очищенню даних.

Інструментальні засоби аналізу якості даних містять десятки тисяч шаблонів і правил для виявлення «брудних» даних.

Одна зі сфер, де особливо гостро стоїть проблема «брудних» даних, – мережі роздрібної торгівлі. Це обумовлено такими причинами, як:

– територіальна розподіленість – безліч філій, розкиданих на великій території;

– великі обсяги даних, що консолідуються, відображаються сотні тисяч торговельних операцій на день;

– висока ступінь деталізації даних – сума кожної окремої торгової операції незначна порівняно із загальними обсягами продажів.

Усе це породжує безліч помилок у даних, які здатні суттєво знизити достовірність результатів аналізу. Виділяються такі основні класи помилок:

1. Помилки введення – виникають у процесі введення даних на торгових терміналах і в касових пунктах. Звичайно, в більшості сучасних супермаркетів касові пункти максимально автоматизовані, щоб уникнути ручного введення даних і пов’язаних з цим помилок. Однак повністю виключити помилки не вдається, оскільки неможливо уникнути ручного введення, коли, наприклад, не зчитується штрих-код, товар надійшов у продаж, але не внесений до БД тощо. Крім того, сканер може неправильно «прочитати» штрих-код.

2. Помилки в програмному забезпеченні торгового термінала.

3. Помилки в базі даних системи обліку торгових операцій.

4. Помилки консолідації даних, які виникають у процесі перенесення даних з філій у централізоване СД торговельної організації.

Завдання оцінки якості даних полягає у виявленні помилок і визначенні їхньої критичності щодо результатів аналізу. Виявлення помилок здійснюється за допомогою спеціальних процедур, кожна з яких розробляється для виявлення певного роду помилок. При цьому необхідно скласти пріоритет помилок і забезпечити відповідний порядок роботи процедур. Результатом роботи такої системи буде звіт про виявлені помилки та оцінки рівня забрудненості даних. Звіт може використовуватися аналітиками для розробки комплексу заходів, спрямованих на підвищення якості даних.

Контроль якості даних повинен здійснюватися безперервно в процесі їхнього збору та консолідації. Для цього доцільно організувати моніторинг даних. Контроль якості даних і виявлення можливих проблем здійснюються як на етапі збору даних в OLTP-системи, так і під час їхньої консолідації і перетворення в процесі ETL. Коли встановлено, що дані не містять помилок, вони безперешкодно завантажуються в СД. Якщо у ході перевірки були виявлені помилки, то проводиться оцінка їхньої критичності щодо аналізу даних. Якщо помилки некритичні, то дані також завантажуються в СД. Якщо виявлено, що помилки критичні, то виробляється комплекс заходів з очищення даних.

Оцінка якості даних, контроль помилок і аналіз виявлених проблем – необхідні ланки будь-якого проєкту у бізнес-аналітиці. Оцінка якості даних не тільки дозволяє визначити ступінь їхньої придатності до аналізу, але і слугує індикатором оптимальності та продуманості роботи таких систем збору і консолідації даних, як OLTP-системи, ETL і СД.

9.4. Очищення і попередня обробка

Незважаючи на багатоступеневу процедуру очищення даних (в OLTP-системах, процесі ETL, СД), навіть на етапі їхнього безпосереднього аналізу виникають серйозні проблеми, що перешкоджають ефективному і коректному аналізу даних. Ці проблеми не завжди пов’язані лише із забрудненістю даних. Існують інші причини, які не дозволяють ефективно використовувати дані для конкретної аналітичної роботи.

З точки зору аналітики «якість даних» і «чистота» – це не одне і те саме. Якщо чистота даних має на увазі відсутність у них помилок введення, структурних порушень, некоректних форматів та інших причин, які заважають аналізу даних взагалі, то якість даних тісно пов'язана з конкретними цілями і задачами аналізу, моделями, методами й алгоритмами, які використовуються. Це означає, що дані, які можна вважати якісними з точки зору одного завдання аналізу (наприклад, прогнозування), виявляться малопридатними для вирішення іншої (наприклад, класифікації). Так, для побудови прогнозу може бути цілком достатньо спостережень за розвитком досліджуваного процесу в деякому часовому інтервалі. А для класифікації об'єкта швидше за все буде потрібно його всебічний опис. Тому не можна достовірно стверджувати, чи є дані якісними, доки вони не будуть пов'язані з конкретною аналітичною задачею. Тому як тільки вибірка даних завантажується в аналітичний додаток з метою вирішення конкретного завдання, аналітик повинен оцінити ступінь їхньої відповідності вимогам, які дозволять отримати якісне аналітичне рішення.

Успіх обробки залежить від таких факторів, як:

– кількість проблем, «успадкованих» від попередніх етапів життєвого циклу даних;

– набору інструментальних засобів обробки, що надаються аналітичною платформою;

– вміння аналітика використовувати інструменти для приведення даних відповідно до вимог конкретного завдання аналізу.

З метою підготовки даних до вирішення конкретного завдання аналізу використовується комплекс засобів, який отримав назву «попередня обробка даних».

Попередня обробка даних – комплекс методів і алгоритмів, що застосовуються в аналітичному додатку з метою підготовки даних до вирішення конкретного завдання і приведення їх відповідно до вимог, визначеним специфікою завдання і способами його вирішення.

Очищення даних в аналітичному додатку – лише один з аспектів попередньої обробки, хоча ці два процеси часто ототожнюються. Однак очищення не є синонімом попередньої обробки.

Більш того, якщо в даних, завантажених в аналітичний додаток, відсутні проблеми, що вимагають очищення, або їх вплив на якість рішення оцінюється як мінімальний, то очищення даних у процесі їхньої попередньої обробки може взагалі не проводитися. Водночас попередня обробка здійснюється в будь-якому випадку.

Характер впливу окремих аспектів якості даних на процес аналізу є різним. Одні проблеми в даних можуть викликати тільки технічні складності, наприклад неможливість інтегрування даних або завантаження їх у сховище. Інші не дозволяють виконувати коректний аналіз даних і перешкоджають отриманню достовірних результатів, наприклад шуми, дублікати і протиріччя. Технічні проблеми в основному виявляються і за можливості усуваються під час консолідації даних. Аналітичні проблеми домінують на етапі безпосередньої обробки даних, коли дані вже прив'язані до визначеної методики і задачі аналізу.

Аномальні значення можуть бути штучними або природними. Штучне аномальне значення є наслідком помилки оператора, збою в роботі системи, тобто воно відображає подію, яка насправді не мала місця і лише спотворює опис предметної сфери. Природна аномалія відображає реальну подію, викликана винятковими обставинами. Зауважимо, що вирішуючи проблему аномальних значень, можна породити інші. Так, пригнічуючи аномальні значення, аналітик позбавляє себе можливості в майбутньому аналізувати роботу підприємства в умовах різкого зростання і падіння попиту, їхнього передбачення.

Аномальні значення не є технічною проблемою, оскільки це просто числа, які в декілька разів більші, ніж сусідні значення. Водночас залежно від специфіки аналізу даних їхня наявність або відсутність може вплинути на якість аналізу і достовірність отриманих результатів, тому проблема аномальних значень є аналітичною. Оптимальним є перенесення обробки аномальних значень в аналітичний додаток, де користувач може самостійно приймати рішення про те, які значення в контексті виконуваного аналізу варто вважати аномальними, чи потрібно їх пригнічувати, якщо так, то якою мірою.

Технічні проблеми впливають як на процес консолідації даних, так і на результати аналізу. Аналітичні проблеми впливають тільки на аналіз даних. Якщо дані в певному вигляді вдалося завантажити в аналітичний додаток, то в них залишилися тільки проблеми аналітичного плану.

Основна відмінність очищення даних у разі попередньої обробки від очищення в OLTP-системах і у процесі ETL полягає в тому, що вона проводиться з урахуванням планованої методики аналізу, його цілей, необхідної достовірності результатів і може коригуватися залежно від отриманих попередніх результатів.

Водночас як розробка очищення даних в OLTP-системах і в процесі ETL є результатом спільної діяльності програміста й аналітика, очищення даних у разі їхньої безпосередньої підготовки до аналізу повністю покладається на плечі користувача аналітичної програми. Цілі і методи попередньої обробки даних повинні повністю визначатися користувачем і обмежуватися лише комплексом засобів, які надані йому системою.

Щоб система могла виявити помилку, необхідно вказати правило або шаблон, якому повинні відповідати «правильні» дані. Помилкою вважатиметься все, що не відповідає певним шаблонам або правилам. Передбачити шаблони і правила для всіх можливих проблем у даних неможливо.

1. Не всі виявлені проблеми вдається усунути. Існують проблеми, усунути які можна, тільки вилучивши самі дані. Тому приймається рішення, що використовувати дані в «брудному» вигляді краще, ніж взагалі позбутися їх.

2. Відмова від усунення помилки. Процес усунення деяких виявлених помилок досить трудомісткий, тому якщо витрати на усунення помилки перевищують вигоду, то може бути прийнято рішення відмовитися від очищення.

3. Програми очищення даних самі можуть бути джерелами помилок. Іноді процедури для очищення даних тільки погіршують ситуацію. Якщо оператор OLTP-системи створює не досить багато помилок, то алгоритм очищення може обробляти одночасно тисячі записів у спробі виправити неіснуючу помилку, що призведе до надмірного навантаження на систему.

Усі помилки, пов'язані з якістю даних, поділяються на два типи. Помилки першого типу мають місце в тому випадку, коли програма очищення даних виявить помилку там, де її немає.

Наприклад. У ларьок, в якому продається морозиво, було завезено товару на 500 грн. До обіду морозиво було продано, і продавець запросив ще одну партію на таку саму суму. До закриття кіоску вона також була розпродана. На складі торгової фірми в систему реєстрації замовлень були введені два записи про відпустку однакових товарів на одну і ту саму суму та однієї і тієї самої дати. Процедура очищення даних, якщо під час її розробки ця ситуація не була передбачена, вважатиме записи дублікатами і видалить одну з них.

Помилки другого типу називають «втраченими». Вони з'являються, якщо процедура очищення даних не змогла виявити проблему, тобто дані, що містять помилку, були розпізнані як правильні. Це відбувається тоді, коли дані виглядають правдоподібно, навіть якщо містять помилку. Також пропуск помилки можливий, якщо відповідне правило або шаблон не були передбачені в програмі очищення.

Замість відсутньої інформації про вік клієнта оператор вказує послідовність чисел 00, вважаючи, що система розпізнає її як фіктивне значення і застосує відповідну обробку. Проте якщо набір символів відповідає заданим шаблонам, то у системи немає формального приводу підозрювати помилку і ці дані можуть безперешкодно проникнути в СД або аналітичну систему.

Таким чином, очищення даних як частина їхньої попередньої обробки в аналітичному додатку є «останньою інстанцією», де аналітик може провести оцінку якості даних і вжити необхідні заходи до їхнього очищення на основі власної думки з урахуванням конкретних цілей та завдань аналізу.

Попередню обробку даних можна розглядати як комбінацію методів очищення і спеціальних методів оптимізації даних для вирішення конкретного завдання.

9.5. Типовий набір інструментів попередньої обробки в аналітичному додатку

У блок попередньої обробки аналітичного додатку, як зазначалося вище, входить набір методів очищення й оптимізації даних для конкретного завдання аналізу. Якщо організація використовує інформаційно-аналітичну підсистему власної розробки, то комплекс засобів попередньої обробки даних у ній може бути створений з урахуванням специфіки вирішуваних завдань. Однак розробити власну аналітичну платформу мають можливість тільки великі підприємства, більшість використовує продукти сторонніх виробників, тому виконувати підготовку даних до аналізу доводиться тими засобами, які закладені розробниками системи.

Типовий набір інструментів попередньої обробки і підготовки даних до аналізу, що є в більшості аналітичних платформ, такі.

Очищення від шумів і згладжування рядів даних. Часто ряди даних містять швидкі випадкові зміни значень, які можна розглядати як шум. Шум заважає виконувати аналіз даних, робить нестійкою роботу алгоритмів, не дозволяє виявляти в даних приховані закономірності, структури, тенденції тощо. Згладжування необхідно в тому випадку, коли ряд даних виявився нерівномірним, містить велику кількість дрібних структур (які не є шумом у звичайному розумінні), що перешкоджають дослідженню більш значних об'єктів і закономірностей. Хоча згладжування й очищення від шумів за своїми цілями подібні, це різні методи обробки даних, що використовують різні математичні підходи.

Відновлення пропущених значень. Порожні значення викликають невизначеність під час роботи багатьох алгоритмів Data Mining. Навіть одне пропущене значення може призвести до неможливості моделювання. Якщо ж пропущених даних багато, то обсягу інформації у вибірці може виявитися недостатньо.

Редагування аномальних значень. Аномальні значення також вимагають уваги у процесі підготовки даних до аналізу. У більшості випадків вони є просто помилками введення. Якщо ж аномальні значення – дійсні події, викликані винятковими обставинами, то вони все одно є перешкодою, оскільки заважають виявленню закономірностей у досліджуваному процесі. Однак дослідження

аномальної поведінки даних дозволить прогнозувати умови, що викликають аномальні події, і їх наслідки, вивчати реакцію бізнес-систем на аномальні зміни умов. Отже, відповіді на питання, які значення вважати аномальними, чи приймати заходи щодо їхнього усунення, далеко неоднозначні і вимагають додаткових досліджень у контексті цілей і методів аналізу.

Обробка дублікатів і протиріч. Дублікати і протиріччя в даних – досить поширені явища, особливо коли дані надходять не з СД, де виконується автоматична підтримка їхньої унікальності та несуперечності, а безпосередньо з різноманітних джерел. При цьому підходи до обробки дублікатів і протиріч істотно різняться.

Дублікати – однакові дані (записи). Вони можуть дублювати інформацію про одну й ту саму подію, а можуть містити ідентичну інформацію про дві різні, але подібні події. У першому випадку дублікати повинні бути видалені, а в другому – потрібна більш тонка обробка. З протиріччями теж не все так просто. Вони виникають там, де порушується логіка причинно-наслідкового зв'язку.

Протиріччя істотно заважають аналізу даних, особливо у разі їхнього використання для побудови моделей, заснованих на навчанні: нейронних мереж, дерев рішень тощо. Використання суперечливих даних різко погіршує процес налаштування параметрів моделей.

Зниження розмірності вхідних даних. В основі роботи більшості моделей Data Mining полягає принцип узагальнення. Тобто щоб отримати на виході моделі навіть єдине значення, потрібно подати на її вхід деякий набір значень, на основі співвідношень між якими і буде визначено вихідний. У процесі побудови моделі спочатку намагаються привернути максимум зібраної інформації про досліджуваний об'єкт: клієнта, постачальника, товар та ін. У результаті набір вхідних змінних розростається, що призводить до ускладнення моделі, робить її вразливою для неякісних даних, збільшує час, необхідний на її побудову. Тому одним із важливих завдань підготовки даних до аналізу є зниження їхньої розмірності. Для цього проводиться пошук вхідних ознак (атрибутів, показників), які мають високий ступінь статистичної взаємозалежності. Такі дані можуть бути виключені з перегляду без істотного збитку для результатів аналізу.

Усунення незначних факторів. Не всі наявні в розпорядженні аналітики дані однаково важливі з точки зору цілей аналізу. Фактор, що включається в розгляд, повинен вносити достатній внесок у рішення задачі, брати участь у поясненні причинно-наслідкового зв'язку між вихідними даними і результатом, тобто між вхідними та вихідними факторами повинна бути висока ступінь взаємозалежності. Якщо між будь-яким вхідним фактором і вихідним результатом такий зв'язок малий або відсутній, то використання цього вхідного фактора безглуздо або навіть шкідливо, оскільки може відвести рішення в хибний напрям. Тому під час підготовки даних до аналізу необхідно виявити вхідні фактори, які слабо або зовсім не пов'язані із шуканим аналітичним рішенням, і виключити їх з простору вхідних факторів. Найчастіше критерієм для визначення значущості вхідних факторів є певний рівень (показник) значущості, який узгоджується зі ступенем залежності (кореляції) шуканого рішення від цього чинника. Однак такий підхід приховує в собі певний ризик, оскільки поріг значимості задається аналітиком на основі суб'єктивних оцінок і може бути помилковим.

У цілому останні два завдання подібні: з аналізу виключаються дані, які в контексті розв'язуваної задачі є зайвими (надлишковими). Принципова різниця полягає в такому.

У першому випадку як критерій для виключення факторів використовується ступінь взаємної залежності (кореляції) між вхідними факторами. Тобто чим вище ця ступінь, тим більше підстав для виключення факторів.

У другому випадку використовується ступінь зв'язку між вхідними факторами і результатом. Тобто чим нижче ця ступінь, тим більше підстав для виключення фактора.

Варто зазначити, що зменшення розмірності аналізованих даних не пов'язано з їх очищенням, оскільки дані можуть розглядатися як надлишкові тільки в контексті розв'язуваної задачі.

9.6. Фільтрація даних

Під час підготовки вибірки даних до аналізу часто виникає ситуація, коли деякі записи потрібно виключити з вибірки і не використовувати. Це може бути потрібно в таких випадках:

1. Значення, що містяться в записі, можуть негативно вплинути на результати аналізу.
2. Значення записів небажано використовувати в цій аналітичній задачі.
3. Запис пов'язаний з будь-яким об'єктом або подією, який небажано розглядати під час аналізу.
4. Запис містить незначну інформацію тощо.

Для виключення записів, присутність яких у вихідній вибірці з будь-якої причини небажано, використовується фільтрація. Фільтрація є багаторозрахованим засобом, який дозволяє виконувати очистку даних від факторів, які зменшують якість аналізу, знижувати розмірність вихідної множини даних, відбирати найбільш важливі дані, спрощувати візуальний аналіз вихідної вибірки та ін.

В основі фільтрації полягає використання умов, які відіграють роль фільтрів, що дозволяють залишати у вибірці одні дані і виключати інші. У деяких випадках застосування фільтрації як альтернатива таким методам підготовки даних, як очищення і зниження розмірності, може дати певні переваги.

По-перше, фільтрація досить проста з обчислювальної точки зору, тому в ситуаціях, коли обробка більш складними алгоритмами займає занадто багато часу, фільтрація дозволяє впоратися з проблемою значно швидше.

По-друге, рішення про доцільність фільтрації та налаштування її параметрів виробляються безпосередньо користувачем, що дає можливість діяти більш тонко, ніж деякі алгоритми очищення і скорочення розмірності, а також уникати проблем, які іноді створюються цими алгоритмами. Крім того, результати фільтрації більш передбачувані і легше інтерпретуються.

Алгоритми фільтрації істотно відрізняються за своєю складністю. У найпростішому випадку це може бути відбір записів,

значення певної ознаки в яких не перевищує заданого. Однак за необхідності використовуються складні багатоступеневі алгоритми фільтрації за декількома полями й умовами одночасно.

В основі технології фільтрації полягають три базових поняття: умова, значення і ставлення.

Умова. Визначається порядок відбору даних, наприклад «більше або дорівнює», «менше» тощо.

Значення. За допомогою заданої умови перевіряється конкретне значення.

Ставлення. Використовується під час фільтрації за допомогою декількох умов і встановлює логічні відносини між цими полями.

Набір умов, які використовуються у разі фільтрації, може бути різним для даних різного типу. Для числових даних використовуються такі умови:

1. «Так само», «нерівно», «більше», «менше», «менше або дорівнює», «більше чи одне». Відбираються тільки ті записи, в яких значення деякої ознаки задовольняє заданій умові. Таким простим способом можна виключати аномалії або значення, величина яких не відповідає цілям аналізу.

2. «Пусте» / «непорожнє». Відбираються тільки ті записи, які в цьому полі містять порожнє (або, навпаки, непорожнє) значення. Цей метод фільтрації дозволяє виключити з розгляду записи з пропусками даних або, навпаки, відібрати з усієї множини тільки повні записи. Значення при цьому не задається.

3. «Містить» / «не містить». Не включаються (виключаються) всі записи, які містять або, навпаки, не містять деяке задане значення.

4. «В інтервалі» / «поза інтервалом». Відбираються тільки ті записи, значення яких для певного поля полягають у деякому заданому діапазоні або поза ним.

5. «У списку» / «поза списком».

6. «Починається на» / «не почалося на». Відбираються записи, в яких значення певної ознаки починається (або, навпаки, не починається) із заданого набору символів. Умова застосовується тільки до строкових даних.

7. «Закінчується на» / «не закінчується на».

За необхідності можна використовувати складні алгоритми фільтрації, які мають кілька умов, які перевіряють дані за декількома полями. Основне в цьому випадку – правильно встановити відносини між умовами. Відносини встановлюються за допомогою логічних операцій «І» та «АБО».

Обробка дублікатів і протиріч. Наявність у даних дублікатів і протиріч часто знижує ефективність роботи моделей і достовірність результатів аналізу. Тому під час підготовки даних до аналізу велика увага приділяється усуненню дублікатів і протиріч, а інструментарій для цього включений у комплекс засобів попередньої обробки багатьох аналітичних платформ.

У більшості випадків вибірка даних, на основі якої будується модель, має деякий набір ознак (атрибутів, показників), що описують бізнес-процес або об'єкт. Ці ознаки можна розділити на дві групи:

– вхідні – подаються на вхід моделі, і на основі їхніх значень модель розраховує вихідний результат. Вхідні ознаки відіграють роль незалежних перемінних;

– вихідні (цільові) – їхні значення формуються на виході моделі як відгук на подачу на її вхід набору вхідних ознак. Вихідні ознаки відіграють роль залежних змінних моделі.

Два записи або більше називаються дублікатами, якщо вони містять ідентичні набори значень усіх ознак.

Два записи або більше є суперечливими, якщо вони містять однакові набори значень вхідних ознак і різні набори значень вихідних.

Дублікати – це однакові записи. А в суперечливих записах порушується однозначність причинно-наслідкового зв'язку: однаковими вихідними даними відповідають різні результати.

Під час обробки дублікатів і протиріч потрібен неформальний підхід, який враховує їхні походження і логіку аналізованих даних. У багатьох випадках це дозволяє уникнути втрати інформації через неправильну інтерпретацію дублюючих і суперечливих записів. Непродумана методика обробки дублікатів і протиріч може призвести до втрати інформації, що зробить негативний вплив на результати аналізу.

Узагальнена модель дублікатів і протиріч даних. У процесі обробки дублікатів і протиріч можуть виникнути різночитання в тому, що вважати дублюючими і суперечливими даними. Наприклад, якщо всі ознаки числові, чи можна вважати дублікатами записи в табл. 2.17.

Таблиця 2.17

Приклад набору даних

№ запису	Ціна	Кількість	Сума	Націнка
1.	1 000	250	250 000	120,02
2.	1 000	250	250 000	120,01
3.	999,9	249,8	249 775,02	120,05
4.	1 100	255	280 500	122,2

Формально три перші записи не є дублікатами, але з точки зору логіки бізнесу відображають ідентичні дані. При цьому неважливо, звідки взялася «зайва копійка» (можливо, це просто помилка округлення).

Щоб визначити, чи варто вважати формально різні записи дублікатами з точки зору логіки бізнесу, можна скористатися їх поданням у багатовимірному просторі ознак, а як міра відмінності вибрати відстань між точками, відповідну кожному запису.

При цьому розмір сфери вибирається з допустимої міри близькості, в межах якої записи можна вважати дублікатами.

Таким чином, якщо два записи або більше потрапили всередину сфери, в межах якої відстань між ними не перевищує заданого значення, то з точки зору логіки бізнесу вони можуть вважатися дублікатами, хоча нормально такими і не є.

Розглянемо аналогічний аналіз для протиріч (табл. 2.18).

Таблиця 2.18

Приклад набору даних

№ запису	Ціна	Кількість	Сума	Націнка
1.	1 001	250	250 250	120,02
2.	1 000	250	250 000	120,01
3.	1 000	250	250 000	130,05

Якщо два записи є дублікатами, то значення всіх їхніх ознак повинні бути ідентичні (або майже ідентичні), а відповідні точки в просторі ознак повинні збігатися або розташовуватися на досить малій відстані один від одного. Якщо ж значення вихідних ознак відрізнятимуться і записи будуть суперечливими, то, незважаючи на близькість вхідних значень, їхні точки в просторі ознак виявляться на великій відстані.

Таким чином, якщо записи мають однакові значення вхідних і вихідних полів і, отже, є дублікатами, то їхні точки в просторі ознак збігаються або близькі одна до одної. Якщо записи суперечливі, то, незважаючи на однакові значення вхідних полів, їхні точки в багатовимірному просторі ознак виявляться віддаленими через відмінності значень вихідних полів. Якщо відстань між записами, що мають однакові значення вхідних полів і різні значення вихідних, досить мала, то це дублікати. Якщо відстань між ними значна, то це протиріччя.

Дублікати викликають надмірність даних, збільшують обсяг вибірки, при цьому абсолютно не підвищуючи інформативність даних. Так, однакові елементи в навчальній вибірці не сприяють процесу навчання, оскільки в них міститься одна і та сама інформація. Разом з тим вони збільшують обсяг вибірки і час, необхідний для навчання. Крім того, якщо обсяг вибірки обмежений, дублікати «забирають» у ній місце у корисних прикладів, що погіршує результати навчання.

Протиріччя призводять до спотворення результатів аналізу і знижують якість моделей, оскільки порушують загальні закономірності в даних, які є метою аналізу.

9.7. Обробка дублікатів і протиріч

Обробка дублікатів і протиріч повинна проводитися з урахуванням особливостей вихідних даних і логіки розв'язуваної задачі. Існує кілька підходів до вирішення проблеми наявності в даних дублікатів і протиріч.

1. Обробка дублікатів і протиріч не проводиться взагалі. Це можливо в таких випадках. По-перше, коли формально дублікати і протиріччя присутні, але при цьому коректність опису динаміки

процесів не порушується. По-друге, якщо дублікати і протиріччя були введені в дані штучно. Нарешті, в деяких завданнях дублікати і протиріччя можуть бути індикаторами відхилень, пошук яких і є метою аналізу.

2. Видалення дубльованих і суперечливих записів. Підхід застосовується в тому випадку, якщо встановлено, що з набору дублюючих і суперечливих записів тільки один відображає дійсне спостереження або подію. Тоді з усіх дублюючих один запис залишають, а інші видаляють. З протиріччями складніше, оскільки потрібно визначити, який із суперечливих записів правильний.

3. Об'єднання (злиття) дублікатів і суперечливих записів. Цей метод застосовується тоді, коли дублікати і протиріччя відображають дійсні події. Наприклад, якщо реєстрація надходження двох однакових партій товару привела до появи двох однакових записів, то можна або видалити один запис, або залишити все без змін. У першому випадку буде втрачена інформація про реальну подію, в другому – з'явиться надмірність у даних, тому оптимальним буде об'єднати два записи в один, виконавши агрегацію кількості одиниць товару і відповідних сум. При цьому ціни, очевидно, агрегувати не варто (рис. 2.52).

Товар	Ціна	К-сть, кг	Сума
Банани	35,00	500	17 500
Банани	35,00	500	17 500

↓

Товар	Ціна	К-сть, кг	Сума
Банани	35,00	100	35 000

Рис. 2.52. Об'єднання дублікатів

Товар	Код товару	К-сть, м	Ціна за 1 м	Сума
Шпалери № 15	2365	100	180	18 000
Шпалери № 15	2395	150	180	27 000

↓

Товар	Код товару	К-сть, м	Ціна за 1 м	Сума
Шпалери № 15	2365	250	180	45 000

Рис. 2.53. Об'єднання протиріч

Залежно від ситуації під час об'єднання записів можуть використовуватися різні функції агрегування: підсумовування, вибір середнього, максимального або мінімального значень тощо.

З усіх видів попередньої обробки даних обробка дублікатів і протиріч найменшою мірою піддається формальному підходу. Це пов'язано з тим, що за наявності суперечливих або дублюючих один одного записів неможливо вибрати метод їхньої обробки, доки не будуть з'ясовані їх характер і походження. За великим обсягом вибірки даних і наявності значної кількості дублікатів і протиріч їхня обробка може стати тривалою процедурою.

Виявлення аномальних значень. Часто у великих наборах даних зустрічаються значення, які не вкладаються в загальну модель поведінки аналізованого процесу. Такі значення, які сильно відрізняються від тих, що оточують дані або несумісні з ними, називаються аномальними значеннями. Аномалії можуть бути викликані помилками вимірювань або введення даних. Якщо, наприклад, у базі даних зазначено вік клієнта, який дорівнює 0, це значення некоректно. Помилка може бути пов'язана з тим, що комп'ютерна програма за замовчуванням встановила це значення, якщо вік клієнта не було вказано взагалі (тобто має місце фіктивне значення, яке також є аномальним). Але якщо в базі даних зазначено, що у клієнта 19 дітей, це значення буде швидше підозрілим і його потрібно перевірити.

Під час підготовки даних до аналізу необхідно виконувати пошук і коригування аномальних значень, оскільки вони є одним з факторів, що істотно зменшує якість даних і достовірність результатів їхнього аналізу. Аномальні дані здатні значно погіршити роботу моделей Data Mining.

Усі аномальні значення можна розділити на два класи:

– штучні – пов'язані з помилками введення даних, некоректною роботою програм або систем реєстрації (наприклад, сканера штрих-коду);

– природні – відображають факти і події, що мали місце в дійсності, але викликані винятковими обставинами, які зустрічаються досить рідко або в одиничних випадках.

Так, аномально велика сума виручки придорожнього кафе може бути викликана помилкою оператора, який у разі введення

вказав зайвий нуль (штучна аномалія). Але можливо, що паралельна дорога була закрита через ремонт, і потік машин (а отже, і кількість клієнтів) зріс у кілька разів. У цьому випадку має місце природна аномалія.

Відрізнити штучне аномальне значення від природного, крім очевидних випадків (типу Вік = 350), практично неможливо. Але якщо виникають сумніви, то краще провести додаткову перевірку. Втім такі дії можливі тільки тоді, коли обсяг вибірки даних і кількість аномальних значень невеликі і з кожним з них можна розібратися окремо.

Завдання обробки аномальних значень складається з двох етапів:

1. Виявлення. Проводиться пошук аномальних значень і, якщо можливо, перевіряється, чи є аномалія штучною або природною.

2. Коригування. Виявлені аномальні значення виключаються або корегуються залежно від логіки й особливостей завдання аналізу.

Якщо вихідна множина не надто велика і не містить великої кількості вимірив даних, то ефективним способом виявлення аномалій є візуальний аналіз таблиці або графіка. У таблиці аномальне значення може бути виявлено за кількістю цифр у числі або за першою цифрою. Більш наочним, звичайно, слугує графічне представлення.

Найбільш простим є візуальний аналіз одновимірних рядів даних. У них аномалії виглядають як викиди значень у певний бік (рис. 2.54).



Рис. 2.54. Ряд даних

Інформація, отримана за допомогою графіка, може використовуватися двояко. З точки зору аналізу має сенс або виключити аномальні значеня, або привести їх у відповідність із загальною моделлю даних, щоб вони не впливали на результати аналізу. Проте з точки зору керівництва кредитної установи ці два значення – привід до перевірки. Особи, які взяли великий кредит, можуть бути або перспективними клієнтами, до яких потрібен особливий підхід (наприклад, спеціальні умови кредитування), щоб підвищити їхню лояльність до банку, або шахраями. Варто зазначити, що якби система автоматично придушувала аномальні значення, то потенційно цінна інформація була б втрачена.

Пошук аномалій і їхнє потенційне видалення з наборів даних можуть бути описані як процес відбору значень, які значно відрізняються від тих, що їх оточують, вибиваються із загального ряду значень або несумісні з іншими даними. Завдання виявлення аномалій особливо в багатовимірних спостереженнях є складним. Методи візуалізації, корисні під час пошуку аномалій в разі використання 1–3 вимірювань, виявляються неспроможними за багатовимірними даними, оскільки немає адекватної методології візуалізації таких просторів.

Найбільш простим підходом до виявлення аномалій в одновимірних рядах даних є статистичний. Якщо припустити, що розподіл значень ряду відомо, потрібно визначити такі основні статистичні параметри, як середнє значення і дисперсія. На основі цих характеристик можна встановити певний поріг, величина якого залежить від дисперсії ряду. Усі елементи ряду, значення яких перевищили цей поріг, можуть розглядатися як за потенційною аномалією.

Методи коригування аномальних значень. Якщо виявлене аномальне значення дійсно спотворює інформацію про досліджуваний бізнес-процес і може вплинути на якість моделі і достовірність результатів аналізу, то необхідно виконати його коригування. Для цього залежно від логіки й особливостей розв’язуваної задачі можна використовувати декілька способів.

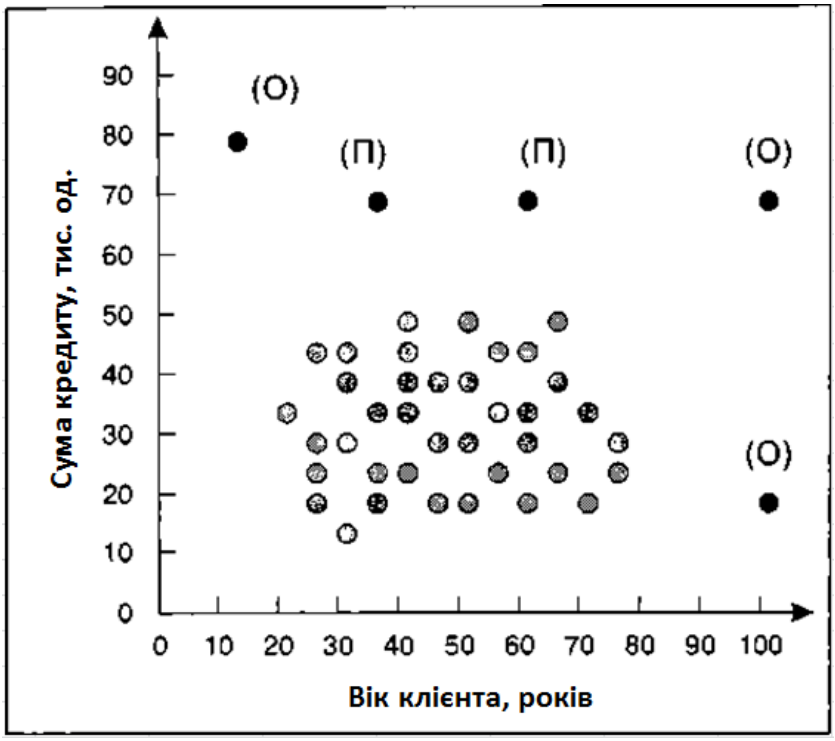


Рис. 2.55. Очевидні і підозрілі аномальні значення

Видалення запису з аномальним значенням. Якщо кількість записів у вибірці даних істотно перевищує мінімум, необхідний для аналізу, то аномальні значення можна просто видалити.

Ручна заміна аномальних значень. Застосовується, якщо кількість аномальних значень невелика і вони можуть бути оброблені вручну. При цьому аналітик замінює аномальні значення на інші, більш відповідні моделі поведінки даних.

Згладжування і фільтрація даних. Для обробки аномальних значень можна використовувати методи частотної або просторової фільтрації, що застосовуються під час згладжування даних і очищення від шумів. При цьому варто враховувати, що в результаті обробки будуть змінені не тільки аномальні значення, але і всі значення ряду.

Інтерполяція даних. Аномальні значення замінюються іншими, обчисленими на основі декількох найближчих сусідів.

Заміна на найбільш імовірне значення. Будується гістограма розподілу значень ряду, і за нею визначається значення, відповідне моді гістограми, яке і буде статистично найбільш імовірним.

9.8. Відновлення пропущених значень

Одним з тих чинників, які знижують якість аналізованих даних і достовірність результатів аналізу, є наявність у них пропущених значень. Пропущене значення виглядає як порожня клітинка в таблиці, де, відповідно до логіки і структури даних, має перебувати деяке значення.

Самі по собі пропущені значення (якщо їх не надто багато) не загрожують інформативності даних і не спотворюють їх. Навпаки, до спотворення може призвести некоректне застосування процедур заповнення пропущених значень. Проте відновлювати пропущені значення необхідно. Основна причина цього полягає в тому, що багато алгоритмів Data Mining не можуть обробляти пропуски.

Підстановка значень за замовчуванням не завжди вирішує проблему, оскільки при цьому не враховуються особливості даних. Припустимо, як значення, підставляються замість пропущених даних за замовчуванням середні продажі за місяць. Але якщо внаслідок зовнішніх факторів обсяги продажів стали піддаватися частим змінам, то середні значення можуть не відповідати реальній поведінці даних.

Отже, необхідно приділяти велику увагу відновленню пропущених значень, а використовувані для цього підходи повинні обиратися виходячи з особливостей самих даних і розв'язуваної задачі аналізу.

Щоб виробити методику боротьби з пропусками в даних, спочатку потрібно зрозуміти, звідки вони беруться. У зв'язку з цим можна виділити два основних моменти появи пропущених значень.

У процесі введення даних, коли потрібні значення були відсутні спочатку. Так, якщо оператор вводить дані з анкет клієнтів, а анкета заповнена неправильно і деякі дані в ній відсутні або вказані нерозбірливим почерком, то оператору нічого не залишається, крім як пропустити відсутні значення, сподіваючись, що пізніше про це хтось подбає. Крім того, пропуски в даних можуть з'являтися через збої в роботі автоматичних систем реєстрації, наприклад під час роботи сканера штрих-коду.

У процесі завантаження даних в аналітичний додаток. У цьому випадку пропуски можуть виникати на місці значень, що мають некоректний тип або формат (рис. 2.56).

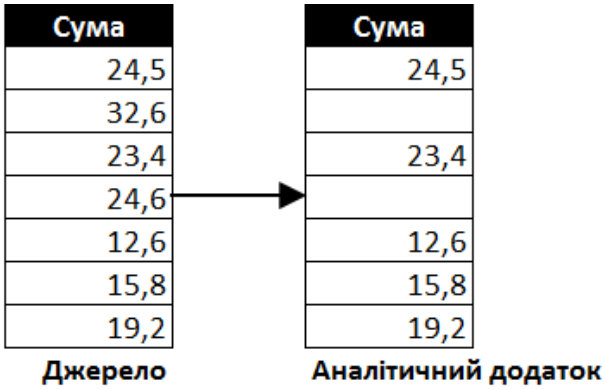


Рис. 2.56. Приклад появи пропусків у даних на етапі завантаження

Вибір методів відновлення пропусків залежить від того, чи є дані впорядкованими. Під впорядкованістю в цьому випадку варто розуміти ступінь взаємної залежності значень даних. Якщо можна припустити, що декілька сусідніх значень даних пов'язані між собою, то такі дані вважатимуться впорядкованими. Найбільш характерна впорядкованість для часових рядів, які складаються зі спостережень за зміною параметрів деякого об'єкта або процесу.

Як правило, реальні бізнес-процеси не схильні до різких змін. Знаючи одне зі значень ряду, можна припустити, що попереднє і наступне значення будуть приблизно такими самими. Тобто в упорядкованих даних значення певною мірою «пояснюють» одне одного.

Якщо дані не впорядковані, то їхні значення ніяк не пов'язані. У цьому випадку намагатися оцінити деяке значення ряду на основі його сусідів безглуздо.

Причини, за якими методики відновлення пропущених значень в упорядкованих і невпорядкованих даних відрізнятимуться, досить очевидні. Якщо дані впорядковані, то отримати хорошу оцінку значення для відновлення пропуску можна на основі сусідніх значень. У разі невпорядкованих даних це зробити неможливо і доводиться застосовувати більш складні алгоритми.

Відновлені за допомогою певного методу значення повинні максимально відповідати загальній моделі поведінки даних. Обчислювальна складність методу не повинна збільшувати час обробки даних до неприйнятної величини, що особливо важливо у процесі аналізу великих масивів даних. Зазвичай вибір методу,

використовуваного для відновлення пропущених значень, – пошук компромісу між перерахованими вимогами. У будь-якому випадку у разі відновлення пропущених значень варто враховувати, що автоматичні процедури, що реалізують цей процес, генерують нові значення даних. Це може призвести до виникнення інших проблем: дублікатів, протиріч і аномальних значень. Тому відновлення пропусків повинно бути першим етапом попередньої обробки даних.

Ручна обробка пропусків. Це найпростіший і ефективний спосіб відновлення пропущених значень, який може використовуватися як для упорядкованих, так і невпорядкованих даних. Аналітик спільно з експертом у предметній сфері перевіряє наявність пропусків у даних і підставляє найбільш імовірні або правдоподібні значення, ґрунтуючись на особистих знаннях і досвіді.

Підстановка констант. Замість пропущених значень підставляється деяке фіксоване значення (константи). Значення констант визначаються такими методами. Заміна пропущеного значення деякою глобальною константою, яка обирається виходячи із специфіки даних і розв'язуваної задачі аналізу. Глобальність має на увазі, що відповідне значення підставляється замість усіх пропусків незалежно від конкретної ситуації. Недолік методу полягає в тому, що якщо під впливом деяких факторів типові обсяги продажів зміняться, а значення, що використовуються для підстановки замість пропущених, не будуть вчасно скориговані, то вони можуть виявитися такими, що суперечать загальній поведінці даних, і негативно вплинути на результати аналізу.

Заміна відсутнього значення найбільш характерним для цього класу. У разі коли дані не впорядковані, для підвищення достовірності оцінки пропущених значень можна виконати попереднє групування даних за класами. Оскільки всередині кожного класу значення ознак будуть подібними, то спочатку у неврегульованій множині даних з'явиться кілька підмножин, у кожній з яких значення можна буде вважати впорядкованими. Аналогічно товари можна розкласти по полицях: на одній – коштовні, а на іншій – дешеві. Якщо відомо, що поряд з товаром, на якому не зазначена ціна, розміщуються дорогі товари, логічно і на нього встановити високу ціну.

Підстановка середнього значення. Ще одним часто використовуваним підходом до відновлення пропущених значень в упорядкованих даних є підстановка замість пропущеного значення середнього за рядом даних або за його найближчими сусідами.

Підстановка найбільш імовірних значень. Одним з популярних методів відновлення пропусків є підстановка замість них найбільш імовірних зі статистичної точки зору значень. Цей метод можна застосовувати як до впорядкованих ним, так і до невпорядкованих даних. У його основі полягає побудова гістограми розподілу значень ознаки, і для підстановки обирається значення з інтервалу, відповідного моді розподілу. Хоча повної коректності відновленого значення метод не гарантує, його використання часто дає хороші результати. Крім того, цей метод має дві незаперечні переваги:

– зрозумілість користувачеві і хорошу візуальну інтерпретацію результатів;

– адаптивність до даних – для кожної вибірки можна побудувати свою гістограму і визначити найбільш імовірне значення.

9.9. Введення в скорочення розмірності

Під час вирішення завдань аналізу часто доводиться мати справу з двома полярними проблемами – надмірністю і недостатністю даних. Варто зазначити, що надмірність і недостатність далеко не завжди пов’язані з кількістю наявних даних. Цілком можливо, що обсяг вихідних даних великий, але для вирішення конкретного завдання аналізу їх все одно недостатньо через відсутність у них необхідної інформації. Можливий і прямо протилежний випадок, коли обсяг вихідних даних невеликий, але вони підібрані настільки вдало, що забезпечують високу ефективність аналізу. Таким чином, інформативність даних не завжди пропорційна їх обсягу.

Практика показує, що краще мати дані із запасом, тобто збирати для аналізу якомога більше даних, що описують предметну сферу. Дійсно, у великих масивах даних є ймовірність «нашкребти» достатню кількість інформації для вирішення певного завдання. Крім того, технічний рівень сучасних носіїв інформації, які можуть зберігати сотні гігабайт даних, знімає проблему зберігання даних.

Коли справа доходить до аналізу, постає проблема: з усієї наявної множини даних, яка може містити десятки ознак, атрибутів і показників, що описують аналізований процес або об'єкт, необхідно вибрати підмножник, який забезпечить оптимальні результати аналізу.

Метою аналізу є пошук закономірностей і структур у даних, тому перед його проведенням варто виділити тільки ті дані, які можуть описати ці закономірності та структури. Це досить важке і неочевидне завдання. Його виконання можна розділити на два етапи.

1. Попередній відбір даних. Здійснюється вручну, коли аналітик вибирає дані для вирішення конкретного завдання на основі особистого досвіду, знань та інтуїції. Основна мета цього етапу – відсіяти дані, які не сприяють отриманню рішення.

2. Скорочення обсягу вихідної вибірки. Як правило, під час первинного відбору даних аналітики перестраховуються. Якщо відсутність або наявність зв'язку між будь-якими даними і цілями аналізу є неочевидною, то в більшості випадків аналітики включають ці дані у вихідну множину, вважаючи, що втрата важливої інформації обійдеться дорожче, ніж залучення зайвої. Але в результаті навіть після попереднього відбору вихідна множина виявляється надмірною, містить досить багато даних, у тому числі слабо або взагалі не пов'язаних з цілями аналізу. Це здатне викликати серйозні проблеми – не менші, ніж у разі втрати релевантної інформації.

Більшість реальних об'єктів і бізнес-процесів досить складна, і для їхнього вичерпного опису потрібно багато ознак і спостережень (записів). Тому розмірність вхідного вектора може виявитися великою – декілька десятків і сотень ознак.

Це викликає ряд серйозних проблем під час аналізу даних:

– зростання обчислювальних витрат і часу, необхідного на обробку даних, до абсолютно неприйнятних значень;

– складність побудови моделі, труднощі розуміння її користувачем;

– складність інтерпретації результатів аналізу й оцінки їхньої достовірності;

– зниження якості результатів аналізу.

У вихідному наборі можуть міститися дані, не пов'язані з досліджуваним процесом або об'єктом. Якщо такі дані не будуть вилучені перед аналізом, то вони можуть відвести рішення задачі в неправильний напрям.

Щоб виключити ці проблеми, в процесі попередньої обробки даних в аналітичних додатках виконується зниження їхньої розмірності.

Вихідні сукупності даних представляються в аналітичному додатку у вигляді «плоских» таблиць, тому коли зазначають про зниження розмірності даних, мають на увазі скорочення даних за такими трьома напрямами, як:

– скорочення кількості ознак (стовпців таблиці);

– скорочення кількості спостережень (записів таблиці);

– скорочення кількості значень вимірювання (при цьому сама кількість значень не змінюється, зменшуються тільки кількості різних варіацій значення). Скорочення даних може здійснюватися в двох режимах.

1. Режим відбору. Визначається значущість кожної ознаки вихідної сукупності для вирішення конкретного завдання. Потім ознаки відбираються в порядку зменшення їхньої значимості. Як тільки трапляється ознака, значимість якої менше деякого порога, відбір припиняється. Поріг значущості встановлюється або на основі статистичного аналізу, або виходячи з досвіду.

2. Режим виключення. Розмір вихідної вибірки скорочується шляхом відкидання незначущих і надлишкових даних. Наприклад, для кожної ознаки вихідної вибірки визначається коефіцієнт значимості, а потім виключаються всі ознаки, значимість яких нижче деякої межі.

Вимоги до алгоритмів зниження розмірності даних. Нині використовується велика кількість різних підходів до зменшення розмірності даних. Одні з цих підходів належать до евристичних, інші засновані на статистичному аналізі даних, треті поєднують у собі і те й інше. Не існує методу зниження розмірності, який однаково добре працював би для всіх завдань аналізу і видів вихідних даних. Рішення про вибір методу ґрунтується на знанні про поставлену задачу й очікувані результати з урахуванням тимчасових обмежень.

Алгоритм повинен надавати користувачеві можливість контролювати хід процесу, а також оцінювати вплив скорочення ознак, записів і окремих значень на очікуваний результат аналізу. Найбільша ефективність повинна досягатися на перших ітераціях алгоритму і з часом зменшуватися (тобто спочатку повинні відкидатися найменш значущі дані або відбиратися найбільш значущі). Це дозволить не чекати закінчення роботи алгоритму, а зупинити його, як тільки будуть досягнуті результати. Алгоритм повинен забезпечувати можливість припинення в будь-який момент часу зі збереженням проміжних результатів, їхньої оцінки, а також відновлення роботи.

Скорочення кількості ознак. Основних переваг від скорочення розмірності вихідної сукупності даних можна домогтися, зменшивши кількість ознак, які використовуються для аналізу. При цьому якість результатів аналізу не повинна істотно знизитися, а в деяких випадках – підвищитися. Коли свідчать про зменшення розмірності вихідних даних, у першу чергу, мають на увазі скорочення кількості ознак. Зменшення кількості спостережень (записів) і окремих значень ознак зазвичай розглядається як допоміжний елемент. Це цілком справедливо, оскільки саме скорочення ознак дозволяє:

– відразу виключити з розгляду велику кількість;

– спростити майбутню модель, зробити її більш зрозумілою;

– виключити надлишкові і незначні дані, які здатні повести рішення в бік, знизити вірогідність результатів аналізу.

Це дозволяє підвищити ефективність вирішення таких завдань Data Mining, як класифікація і кластеризація, заснованих на визначенні ступеня близькості векторів спостережень. Крім того, зменшуючи кількість ознак, можна знизити розмірність OLAP-кубів. Відомо, що багатовимірні поділи з розмірністю більше 7–9 досить погано піддаються візуальному аналізу й осмисленню користувачем. Тому якщо розмірність моделі вдається скоротити без істотної втрати і спотворення інформації, робота з моделлю буде більш ефективною.

Ознаки, які можна скоротити, поділяються на дві групи.

Незначущі – ознаки, які слабо або ніяк не пов’язані з результатами аналізу. Використовувати такі ознаки недоцільно з двох

причин. По-перше, вони не несуть у собі інформації, яка може бути використана під час аналізу, не містять закономірностей і структур, які можуть являти собою інтерес у процесі вирішення задачі. По-друге, використання незначних даних може завдати шкоди результатам аналізу. Такі дані можуть містити закономірності, які не мають відношення до розв'язуваної задачі, але можуть бути виявлені моделлю замість шуканих.

Надлишкові – ознаки, які містять корисну інформацію, але без яких можуть бути отримані хороші результати аналізу. Тобто інформація, яку несуть такі ознаки, є зайвою. Уся необхідна інформація міститься в одній або декількох інших вхідних ознаках, які також використовуються для аналізу, а надлишкова ознака в певному вигляді дублює її.

Основна відмінність незначущих ознак від надлишкових полягає в тому, що незначні ознаки мають низький ступінь зв'язку з результатами аналізу, а надлишкові – високий ступінь зв'язку з іншими вхідними полями вихідної сукупності даних.

Більшість алгоритмів і методів скорочення кількості ознак використовують статистичні або евристичні критерії, на основі яких приймаються рішення про включення певної ознаки в модель. Зі скороченням набору ознак пов'язані дві стандартні задачі:

– відбір ознак – ґрунтується на знанні предметної сфери і цілей аналізу;

– композиція ознак – відбувається формування однієї ознаки на основі злиття декількох інших з використанням деякого перетворення. Хоча такі алгоритми скорочення ознак досить складні, вони дозволяють отримати нові ознаки, які суттєво впливають на результати аналізу.

Різні методи відбору ознак забезпечують різний ступінь зменшення вихідної сукупності даних. Можна класифікувати ці методи на дві групи: алгоритми впорядкування ознак (ранжування) й алгоритми мінімальної підмножини.

Алгоритми впорядкування ознак засновані на обчисленні для кожної ознаки деякого оціночного показника (коефіцієнт кореляції тощо). Потім ознаки ранжуються в порядку зростання або зменшення цього показника. У більшості випадків ці алгоритми

не використовуються безпосередньо для формування множин скороченої розмірності, вони тільки показують значимість цієї ознаки порівняно з іншими, що дозволяє аналітику прийняти рішення про їхнє скорочення.

Алгоритми мінімальної підмножини, навпаки, після їх запуску видають мінімальну підмножину ознак, необхідних для ефективного аналізу даних. Ознаки, що потрапили в підмножину, є значущими для аналізу, інші – ні.

В обох видах алгоритмів важливо встановити схему оцінки ознаки, тобто спосіб, за допомогою якого ознаки оцінюються, ранжуються і додаються в результуючу підмножину.

Відбір ознак на основі статистичних показників. Одна з методик відбору ознак заснована на порівнянні їхніх середніх значень і дисперсій. Недолік цього підходу в тому, що розподіл значень ознак не завжди відомий. Якщо він передбачається нормальним, то статистичні оцінки можуть працювати досить добре, але якщо характер розподілу невідомий, то на основі статистичних оцінок не можна побудувати точну модель.

Якщо одна ознака містить значення, що описують різні класи об'єктів, то повинні бути досліджені елементи з різних класів. Для цього значення ознаки розподіляються за класами, обчислюються їхні середні значення і дисперсія. Потім середні значення нормалізуються дисперсією і порівнюються. Якщо середні значення для різних класів істотно розрізняються, то інтерес до ознаки, що містить такі значення, з точки зору поділу класів збільшується. Якщо середні значення розрізняються мало, то очікувана ефективність ознаки з точки зору поділу класів невелика. Цей евристичний підхід не є оптимальним під час відбору ознак, але він цілком заможний у разі вирішення деяких завдань аналізу.

У результаті скорочення розмірності даних інформаційний зміст результуючої вибірки не повинен суттєво змінитися щодо вихідної. Але це не більше ніж побажання, до того ж досить очевидне. Проте воно слугує відправною точкою для групи методів скорочення надлишкових і незначущих ознак. Іншими словами, скорочувати кількість ознак потрібно таким чином, щоб характеристики, що відображають інформаційний зміст вихідної вибірки, змінилися не більше ніж на задану величину.

Для оцінки інформаційного змісту даних може слугувати спеціальна міра теорії інформації – ентропія. Це поняття вперше було введено як міра різноманітності значень даних. Чим більш гладким є ряд даних, тим менше його ентропія. Якщо ряд даних складається з однакових значень, його ентропія дорівнює 0. Очевидно, що ні про яку інформативність даних у цьому випадку свідчити не доводиться. Зі збільшенням різноманітності значень ознаки зростають його ентропія та очікуваний інформаційний зміст. Таким чином, ентропія може слугувати мірою інформаційної насиченості даних. Чим вище ентропія набору даних, тим вище очікувана інформаційна насиченість, і навпаки.

В основі методики полягає порівняння ентропії вихідної вибірки даних та ентропії, отриманої після зниження розмірності. Якщо ентропія результуючої вибірки близька до ентропії вихідної, то вона за своїм інформаційним змістом близька до вихідної. Метод, заснований на ентропії, відносно простий, але зі зростанням кількості ознак його обчислювальна складність істотно зростає.

Кореляційний аналіз. Для кількісної оцінки зв'язку між двома випадковими величинами використовується кореляційний аналіз. Принцип кореляційного аналізу полягає в пошуку таких факторів, які найменшою мірою корельовані (взаємопов'язані) з вихідним результатом. Ці фактори можуть бути виключені з результуючого набору даних практично без втрати корисної інформації. Метод також застосовується для пошуку серед вхідних ознак таких пар, між якими існує функціональна лінійна залежність. У цьому випадку з двох ознак доцільно залишити тільки одну. На практиці вважається, що якщо кореляція більше 0,6, то зв'язок між двома ознаками досить високий, якщо менше 0,3 – залежність відсутня.

Метод головних компонент. Іншим популярним методом скорочення розмірності вихідних даних є метод головних компонент. Він полягає в перетворені вихідної множини даних X, представлених векторами значень ознак, в іншу множину Y, де дані будуть представлені векторами меньшої розмірності. Тобто метод можна розглядати як скорочення розмірності векторів ознак, що описують об'єкти класифікації, кластеризації та інших задач аналізу. Мета цього перетворення – сконцентрувати всю інформацію про відмінності між спостереженнями в невеликій кількості ознак.

Перетворення з X в Y ґрунтується на припущенні, що більш інформативними є ознаки, що забезпечують найбільшу мінливість вихідних даних. Мірою мінливості слугує дисперсія векторів ознак. Необхідно знайти ознаку, яка «бере» на себе більшу частину дисперсії набору даних. При цьому передбачається, що вона успадкує і більшу частину інформації. Така ознака називається головною компонентою.

Завдання полягає в послідовному пошуку головних компонент, максимальна кількість яких дорівнює кількості ознак вихідної множини. Тоді перша головна компонента буде займати перше місце з інформаційного навантаження, друга – друге тощо. Таким чином, останні головні компоненти будуть найменш інформативні і можуть бути виключені з розгляду як незначущі, що приведе до скорочення простору ознак з мінімальною втратою інформації.

Проте кожна головна компонента являє собою нову ознаку, яка не пов'язана ні з одною з вихідних ознак, а є лише їхньою лінійною комбінацією. Тому у разі подальшого аналізу, можливо, виникне необхідність інтерпретувати нові ознаки, отримані на основі головних компонент, наповнити їх новою логікою і здоровим глуздом. Наприклад, інтерпретація нових ознак може бути проведена на основі їхньої кореляції з вихідними ознаками.

Процес визначення головних компонент полягає в побудові нової осі багатовимірного простору ознак, що збігається з напрямком, уздовж якого дисперсія векторів ознак максимальна. Як тільки ця вісь буде побудована, базова система координат повертається таким чином, щоб одна з її осей збіглася з новою віссю. Вона і утворює першу головну компоненту (рис. 2.57).

Закономірним є питання: скільки головних компонент стане необхідно для опису вихідних даних? Для відповіді найпростіше проаналізувати частку дисперсії, яку вносить кожна компонента в загальну дисперсію. Виразивши частку дисперсії у відсотках, можна встановити поріг, який називається порогом значущості. Усі ознаки, частка яких у загальній дисперсії менше порога значимості, вважаються такими, що мають низьку значимість і підлягають скороченню.

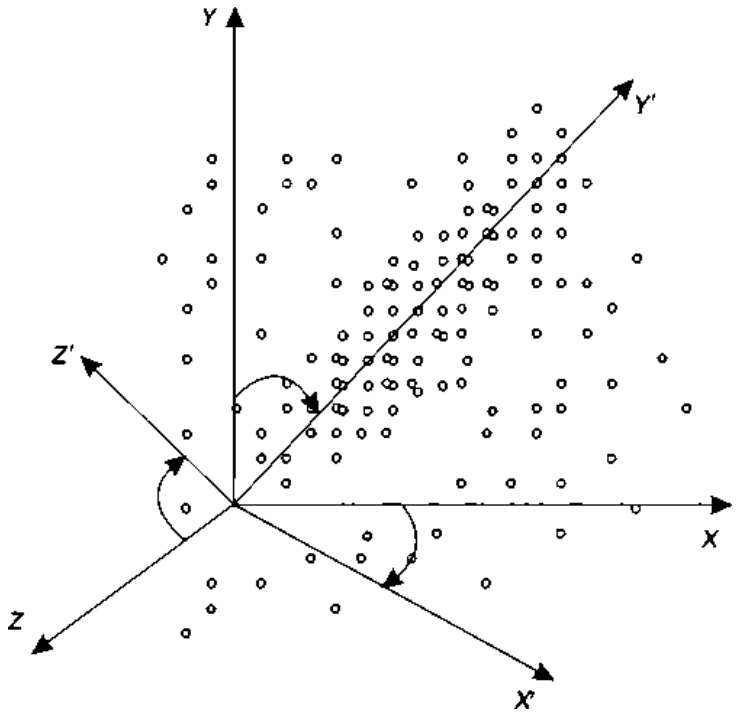


Рис. 2.57. Метод головних компонент

На практиці поріг значимості вибирається експериментально виходячи з необхідної точності результатів.

У разі скорочення розмірності вихідних даних основна увага приділяється зменшеню кількості ознак, які використовуються для аналізу. І це виправдано, оскільки саме виключення надлишкових і незначущих ознак дозволяє найбільш ефективно спростити аналітичну модель, зробити її зрозумілою для користувача і підвищити інтерпретованість результатів аналізу.

Скорочення кількості різноманітних значень ознак зазвичай пов'язане зі створенням інтервалу дискретизації значень. Завдання полягає в зменшенні кількості значень вихідного набору даних за рахунок їхнього об'єднання в межах деякого інтервалу і присвоєння цьому інтервалу певного символу або значення. У результаті такого перетворення кількість значень ознаки, які потрібно обробити в процесі аналізу, має зменшитися без істотного збитку для інформативності даних. У багатьох випадках додатковим виграшем від такої обробки є спрощення опису досліджуваних об'єктів.

Звернемо увагу, що зменшується не сама кількість значень ознаки, а кількість його унікальних значень. Наприклад, якщо ознака містить 1 000 значень у діапазоні від 9,5 до 10,4, то просте

округлення дозволить перетворити всі ці значення до 10. Тобто сама кількість значень ознаки залишиться без змін, а ось кількість унікальних значень зменшиться.

Зазвичай така операція проводиться вручну на основі деяких апріорних знань про досліджуваний об'єкт або процес. Наприклад, вік людини, заданий як безперервний ряд в інтервалі від 0 до 100, може бути розділений на інтервали за категоріями: дитина, підліток, зрілий вік, середній вік, похилий вік тощо.

Скорочення кількості значень ознак зазвичай не шкодить практичним завданням аналізу і водночас дозволяє домогтися істотного зменшення обчислювальних і часових витрат на аналітичну обробку даних.

Якщо кількість різних значень ознаки може бути зменшено, то якість аналізу, проведеного за допомогою деяких методів Data Mining, особливо заснованих на виявленні логічних правил (наприклад, дерева рішень), може покращитися.

Основна проблема цього методу – пошук точок зрізу для груп. Теоретично розбиття на групи не може бути виконано незалежно від інших ознак. Проте евристичне рішення для кожної ознаки окремо у багатьох задачах аналізу дає хороші результати.

Завдання скорочення кількості значень може розглядатися як завдання оптимізації у виборі до груп: дано кількість груп I, яка повинна розподіляти значення в групах таким чином, щоб мінімізувати відстань від значення до середнього або медіани групи, в яку це значення попало. Ця відстань зазвичай визначається як квадрат відстані для середнього по групі і як модуль відстані для медіани. Спрощена модифікація такого алгоритму містить такі кроки.

1. Упорядковано всі значення цього показника.

2. Задається приблизно однакова кількість суміжних значень x_i для кожної групи; кількість груп задана заздалегідь.

3. Граничний елемент i переміщається зі своєї групи в наступну або в попередню, якщо це зменшує загальну помилку відстані – суму відстаней від значення x_i до середньої або моди відповідної групи.

9.10. Семплінг

Під час збору і консолідації даних у сховищах даних можуть накопичуватися великі масиви інформації. Зазвичай вона стосується бізнес-процесів, які відбуваються в компанії та іноді у зовнішньому її оточенні. Коли виникає необхідність виконати аналіз деяких даних з метою отримання нових знань про пов'язані з цими процесами або об'єктами, аналітик стикається з проблемою вибору: що саме використовувати для аналізу. Тут можливі два варіанти:

– використовувати для аналізу всі наявні дані;

– вибрати зі всієї множини наявних даних деяку підмножину і на основі аналізу цієї підмножини спробувати зробити висновки про властивості досліджуваного процесу або об'єкта.

Обидва підходи мають свої переваги і недоліки. Використання всієї множини наявних даних дозволить забезпечити максимальну достовірність результатів аналізу, оскільки при цьому буде врахована вся доступна інформація. Але на практиці цей підхід реалізувати проблематично, оскільки обсяг даних може виявитися настільки великим, що час і обчислювані ресурси, необхідні для аналізу, вийдуть за всі можливі межі. Нікому не сподобається, якщо інформаційна система компанії виявиться паралізована на декілька годин через те, що сервер, на якому встановлена аналітична платформа, займається аналізом набору даних, що складається із сотень мільйонів записів. Однак варто зазначити, що якщо цінність очікуваних результатів аналізу досить висока, а тимчасові обмеження на його проведення некритичні, то, можливо, варто використовувати всі доступні дані, незважаючи на витрати.

Основна перевага другого підходу полягає в тому, що аналізується відносно незначна підмножина даних. При цьому часові й обчислювальні витрати невисокі, що особливо важливо для оперативного управління і планування, коли управлінські рішення, які генеруються на основі результатів аналізу, потрібно отримувати практично в режимі реального часу. Недолік цього підходу в тому, що оскільки під час аналізу використовується тільки частина доступної інформації, то ймовірне зниження достовірності результатів аналізу.

На практиці в бізнес-аналітиці застосовується другий метод, коли аналізу підлягає тільки деяка підмножина даних. Це обумовлено такими причинами.

1. Стратегічні рішення, для обґрунтування яких необхідно залучити максимальну кількість інформації, приймаються досить рідко. Набагато більш затребувані рішення у сфері оперативного управління і планування, для яких потрібна інформація про окремі аспекти і в певних інтервалах діяльності компанії. Тому для аналізу з метою підтримки прийняття таких рішень повинні бути обрані лише дані, які безпосередньо стосуються досліджуваної проблеми.

2. З точки зору аналізу вирішальну роль відіграє не кількість даних, а їхня інформаційна насиченість. Наприклад, можна проаналізувати досить великий масив даних, витративши на це багато часу. Але якщо ці дані слабо відображають шукані закономірності, то результати аналізу виявляться некорисними. Водночас підмножина даних може бути компактною, але при цьому добре відображати поведінку досліджуваного процесу. У такому випадку аналіз не займе багато часу і дасть якісні результати.

3. Під час вирішення деяких завдань аналізу використання всіх накопичених даних неможливо. Особливо це характерно для прогнозування, коли мають використовуватися тільки актуальні дані, а застарілі – відкидатися.

Існують різні стратегії вилучення підмножин спостережень з вихідних множин даних. Комплекс таких методів отримав назву «семплінг». Прийнятний розмір вибірки визначається виходячи з обчислювальних можливостей, доступного обсягу пам’яті, необхідної точності та інших характеристик алгоритму і даних.

Таким чином, у більшості випадків у разі підготовки даних до аналізу з усієї наявної в розпорядженні аналітики множини даних вибирають деяку підмножину, на основі якої і намагаються отримати потрібне рішення. Семплінг забезпечує оперативність виконання аналізу.

Поняття «семплінг» виходить за рамки технологій аналізу. У більш загальному сенсі під семплінгом розуміють відбір деякого набору об’єктів (товарів, виробів, респондентів тощо), який

виділяється в окрему групу за будь-якої мети. Наприклад, респонденти відбираються з метою опитування, товари – в групу найбільш популярних тощо.

Сукупність – множина всіх даних, що описують деякий об'єкт або процес, які можуть бути використані для аналізу з метою отримання нових знань про цей об'єкт або процес.

Вибірка – підмножина даних, що формується шляхом відбору із сукупності з метою аналізу. Для забезпечення достовірності результатів аналізу отримана вибірка повинна бути репрезентативною, тобто достатньою мірою відображати властивості і характеристики сукупності.

Семплінг – процес відбору з вихідної сукупності даних вибірки, з метою проведення деякого аналізу. Під час реалізації семплінгу використовуються спеціальні методи відбору, які повинні забезпечити репрезентативність вибірки з точки зору розв'язуваної задачі.

Отже, семплінг являє собою процес відбору з початкової множини даних (сукупності) деякої підмножини (вибірки) для аналізу. Кінцевою метою семплінгу є зведення задачі аналізу всіх доступних даних до аналізу обмеженої вибірки, яка достатньою мірою відображає властивості всієї досліджуваної сукупності.

Рівномірний випадковий семплінг. Усі записи вихідної сукупності поділяються на групи, в кожній з яких міститься однакова кількість записів (наприклад, 100). Потім з кожної групи випадковим чином обирається один запис і вставляється в результатну вибірку. Вибірка, отримана в результаті семплінгу, складатиметься із записів, випадковим чином відібраних з кожної групи.

Кількість груп, на які потрібно розбити вихідну сукупність, залежить від загальної кількості записів у ній і кількості записів, з яких повинна складатися результатна вибірка. Розмір результатної вибірки, в свою чергу, залежить від особливостей розв'язуваної задачі.

Іноді застосовується більш простий варіант випадкового семплінгу, коли результатна сукупність не розбивається на підгрупи, а вибірка проводиться випадковим чином з усієї сукупності (але в ряді випадків це робить результуючу вибірку менш рівномірною).

Основні переваги цього методу – простота й ефективність. Головний недолік полягає в тому, що якщо вихідна сукупність не є гомогенною, тобто містить різного розміру групи, в кожній з яких дані мають різний розподіл, це може призвести до отримання зміщеної вибірки.

Якщо вихідна сукупність істотно неоднорідна, випадковий семплінг працює погано і кращих результатів вдається досягти, якщо використовувати інші методи семплінгу, такі як стратифікаційний і кластерний.

Стратифікаційний семплінг. Коли вихідна сукупність істотно неоднорідна, найкращих результатів вдається досягти, якщо проводити вибірку з кожної групи незалежно від інших груп. Такі групи називаються стратами, або верствами (strata – шар). Стратифікаційний семплінг виконується в два етапи.

1. Стратифікація – групування елементів вихідної сукупності у відносно однорідні підгрупи, які і називаються шарами.

2. Випадковий семплінг – випадкова вибірка з кожного шару окремо.

Шари повинні бути взаємовиключними, тобто кожен елемент вихідної сукупності входить тільки в один шар. Крім того, жоден з елементів сукупності не повинен бути пропущеним. Основною проблемою стратифікації цінного семплінгу є визначення складу шарів і розміру вибірки, що вилучається з кожного шару.

Як правило, для визначення складу шарів необхідно використовувати деяку апріорну інформацію про вихідну сукупність, а також враховувати характер завдання, для вирішення якого формується вибірка. Що стосується розмірів вибірок з кожного шару, то тут можливі дві стратегії:

1. Пропорційний розподіл.

2. Оптимальний розподіл. Розмір вибірки, яка вилучається з шару, пропорційна стандартному відхиленню розподілу деякої змінної. Найбільші вибірки беруться в шарі з більшою варіабельністю.

Тести до розділу 2

1. Виділяють три основні завдання СППР:

а) введення даних, зберігання даних, аналіз даних;

б) використання даних, зберігання даних, обробка даних;

в) введення даних, використання даних, аналіз даних;

г) введення даних, зберігання даних, видалення даних.

2. За ступенем «інтелектуальності» обробки даних під час аналізу виділяють три класи завдань:

а) інформаційно-аналітичний, інтелектуальний, пошуковий;

б) інформаційний, оперативний, розподільчий;

в) інформаційно-пошуковий, оперативно-аналітичний, інтелектуальний;

г) інформаційно-аналітичний, інформаційний.

3. Транзакція – це:

а) набір операцій над базою даних, який розглядають як початок, з точки зору користувача, дії над деякою інформацією;

б) набір операцій над базою даних, який розглядають як початок, з точки зору користувача, дії над деякою операцією;

в) набір операцій над базою даних, який розглядають як початок і кінець, з точки зору користувача, дія над деякою інформацією, зазвичай пов’язана зі звертанням до бази даних;

г) деякий набір операцій над базою даних, який розглядають як єдину завершену, з точки зору користувача, дія над деякою інформацією, зазвичай пов’язана зі зверненням до бази даних.

4. Принцип незмінності передбачає:

а) що в СД дані після завантаження можуть піддаватися будь-яким змінам і додаються нові дані;

б) що в СД дані після завантаження не повинні піддаватися будь-яким змінам, за винятком додавання нових даних;

в) що в СД дані після завантаження не повинні піддаватися будь-яким змінам, за винятком їх видалення;

г) що в СД дані після завантаження можуть виправлятися, постійно оновлюватися, видалятися та додаватися нові дані.

5. Найважливішим елементом СД є:

а) сховище даних;

б) переробний процес даних;

в) семантичний шар;

г) симетричний шар.

6. Дані в СД зберігаються:

а) як в оперативному вигляді;

б) як в агрегатному, так і в оперативному вигляді;

в) як в деталізованому, так і в оперативному вигляді;

г) як в деталізованому, так і в агрегованому вигляді.

7. Щоб наблизити СД до умов і специфіки конкретної організації, на сьогодні розроблено декілька архітектур сховищ:

а) реляційні, багатовимірні, гібридні і віртуальні;

б) регресивні, багатогранні, віртуальні;

в) реляційні, багатогранні, гібридні;

г) гібридні і віртуальні.

8. OLAP-система повинна надавати можливість:

а) кільком користувачам працювати одночасно з різними моделями;

б) кільком користувачам працювати за чергою з однією аналітичною моделлю;

в) кільком користувачам працювати спільно з однією аналітичною моделлю;

г) правильної відповіді немає.

9. OLAP-куб може бути реалізований і за допомогою звичайної:

а) реляційної моделі;

б) регресивної моделі;

в) геометричної моделі;

г) філігранної моделі.

10. Реляційна база даних – це:

а) сукупність відносин, що містять усю інформацію, яка повинна зберігатися в базі;

б) сукупність відносин, яка зберігає інформацію тимчасово;

в) сукупність відносин, що містить частину цієї інформації, яка зберігається певний період;

г) правильної відповіді немає.

11. Під час вибору джерел даних для завантаження в СД необхідно враховувати такі фактори:

а) процес отриманих даних, можливе порушення цілісності та достовірності даних, обсяг даних у джерелі;

б) значимість даних з точки зору аналізу, обсяг даних у джерелі, процес отримання і завантаження даних, процес переробки і зберігання даних;

в) процес сортування даних, їх переробка і збереження;

г) значимість даних з точки зору аналізу, складність отримання даних з джерел, можливе порушення цілісності та достовірності даних, обсяг даних у джерелі.

12. Очищення даних обов'язкове під час їх:

а) перевантаження в сховище;

б) переробки;

в) аналізу;

г) сортування.

13. Некритичні помилки можуть бути виправлені у процесі:

а) переробки даних засобами аналітичної системи;

б) сортування даних засобами аналітичної системи;

в) аналізу даних засобами аналітичної системи;

г) скачування з джерел даних засобами аналітичної системи.

14. Фіктивні значення – це:

а) дані, які не мають сенсу, ніяк не пов'язані з описуваним ними бізнес-процесом;

б) дані, які постійно змінюються і тісно пов'язані з описуванням бізнес-процесу;

в) дані, які постійно змінюються і ніяк не пов'язані з описуванням бізнес-процесу;

г) дані, які відіграють важливу роль у процесі описування бізнес-процесу.

15. Закодовані значення – це:

а) скорочення, абревіатури, заміна найменувань числовими кодами;

б) повне роз'яснення скорочень і абревіатур;

в) часткове роз'яснення скорочень і абревіатур;

г) правильної відповіді немає.

16. Крім достовірності та повноти даних, існує ще один фактор, що безпосередньо впливає на ефективність їхнього аналізу:

а) багатовимірність даних;

б) інформаційна насиченість;

в) багатогранність даних;

г) правильної відповіді немає.

17. Трансформація даних – це:

а) комплекс методів і алгоритмів, спрямованих на оптимізацію уявлення і форматів даних з точки зору завдань, які необхідно вирішити, і цілей аналізу;

б) досить широке поняття, яке не має чітко окреслених меж;

в) однин з найважливіших процесів в аналізі;

г) усі відповіді правильні.

18. Системи оперативної реєстрації даних (OLTP) зазвичай:

а) доступні всім працівникам;

б) постійно доступні аналітику;

в) взагалі недоступні аналітику;

г) доступні в екстремальних ситуаціях.

19. Групування даних – це:

а) комплекс методів і алгоритмів, спрямованих на оптимізацію уявлення і форматів даних з точки зору вирішуваних завдань і цілей аналізу;

б) корисний інструмент, застосування якого в процесі підготовки даних до аналізу дозволяє більш ефективно використовувати інформацію;

в) корисний інструмент, застосування якого в процесі підготовки даних до переробки дозволяє більш ефективно використовувати інформацію;

г) комплекс методів і алгоритмів, спрямованих на переробку даних.

20. Розрізняють два основні методи квантування:

а) пропорційне і непропорційне;

б) рівноцінне і нерівноцінне;

в) рівномірне і нерівномірне;

г) пряме і кругле.

21. Семплінг – це:

а) процес відбору з вихідної сукупності даних вибірки з метою проведення деякого аналізу;

б) процес відбору з початкової множини даних (сукупності) деякої підмножини (вибірки) для аналізу;

в) відбір деякого набору об’єктів (товарів, виробів, респондентів тощо), який виділяється в окрему групу з будь-якою метою;

г) усі відповіді правильні.

22. Ознаки, які можна скоротити:

а) незначущі, надлишкові;

б) якісні;

в) кількісні;

г) немає правильної відповіді.

23. Розмірність вхідного вектора може викликати ряд серйозних проблем під час аналізу даних, до яких відносять:

а) зростання обчислювальних витрат і часу, необхідного на обробку даних до абсолютно неприйнятних значень;

б) складність побудови моделі, труднощі розуміння її користувачем;

в) складність інтерпретації результатів аналізу й оцінки їхньої достовірності;

г) усі відповіді правильні.

РОЗДІЛ 3
МЕТОДИ БІЗНЕС-АНАЛІТИКИ ТА МОДЕЛЮВАННЯ

§ 10. Задачі Data Mining

10.1. Класифікація задач Data Mining у бізнес-аналітиці

Методи Data Mining у бізнес-аналітиці допомагають вирішити багато завдань, з якими має справу аналітик. З них основними є:

– класифікація;

– регресія;

– пошук асоціативних правил;

– послідовність;

– прогнозування;

– визначення відхилень або викидів, аналіз відхилень або викидів;

– оцінювання;

– аналіз зв'язків;

– візуалізація;

– кластеризація.

Завдання класифікації зводиться до визначення класу об'єкта за його характеристиками. При цьому кількість класів, до яких може бути віднесений об'єкт, відомо заздалегідь.

Завдання регресії, подібне задачі класифікації, дозволяє визначити за відомими характеристиками об'єкта значення деякого його параметра. На відміну від завдання класифікації значенням параметра є не кінцева кількість класів, а множина дійсних чисел. Тобто основна відмінність задачі класифікації від регресії полягає в тому, що у першому випадку залежна змінна є дискретною, категоріальною, а в другому – неперервною. Водночас за допомогою звичайної регресії можна виконувати завдання класифікації, розбивши сферу значень на певну кількість інтервалів, кожний з яких відповідає певному класу (категорії).

Під час пошуку асоціативних правил метою є знаходження часових залежностей (або асоціацій) між об'єктами або подіями. Знайдені залежності представляються у вигляді правил і можуть бути використані як для кращого розуміння природи аналізованих даних, так і для передбачення появи подій.

Послідовність дозволяє знайти тимчасові закономірності між транзакціями. Завдання послідовності подібне асоціації, але її метою є встановлення закономірностей не між одночасно наступаючими подіями, а між подіями, пов'язаними в часі (тобто відбуваються з деяким певним інтервалом у часі). Іншими словами, послідовність визначається високою ймовірністю ланцюжка пов'язаних у часі подій. Фактично асоціація є окремим випадком послідовності з часовим лагом, рівним нулю. Це завдання Data Mining також називають завданням знаходження послідовних шаблонів. Правило послідовності: після події X через певний час відбудеться подія Y.

Приклад. Після покупки квартири мешканці в 60 % випадків упродовж двох тижнів купують холодильник, а упродовж двох місяців у 50 % випадків – телевізор. Рішення цього завдання широко застосовується в маркетингу та менеджменті, наприклад у процесі управління циклом роботи з клієнтом.

У результаті рішення задачі прогнозування на основі особливостей історичних даних оцінюються пропущені або ж майбутні значення цільових чисельних показників. Для вирішення таких завдань широко застосовуються методи математичної статистики, нейронні мережі та ін.

Мета рішення завдання визначення відхилень або викидів, аналіз відхилень або викидів – виявлення та аналіз даних, які найбільш відрізняються від загальної множини даних, виявлення так званих нехарактерних шаблонів.

Завдання оцінювання зводиться до передбачення безперервних значень ознаки.

Аналіз зв'язків – задача знаходження залежностей у наборі даних.

У результаті візуалізації створюється графічний образ аналізованих даних. Для рішення задачі візуалізації використовуються графічні методи, що показують наявність закономірностей у даних.

Завдання кластеризації полягає в пошуку незалежних груп (кластерів) і їхніх характеристик у всій множині аналізованих даних. Вирішення цього завдання допомагає краще зрозуміти дані. Крім того, групування однорідних об'єктів дозволяє скоротити їхню кількість, а отже, і полегшити подальший аналіз. Основною відмінністю класифікації від кластеризації є те, що під час кластеризації наперед не відома кількість класів, на які буде розбита досліджувана сукупність.

Підведення підсумків – задача, мета якої – опис конкретних груп об'єктів з аналізованого набору даних.

Перераховані завдання за призначенням поділяються на описові і прогностичні. Описові завдання приділяють увагу поліпшенню розуміння аналізованих даних. Ключовий момент у таких моделях – легкість і прозорість результатів для сприйняття аналітиком. Можливо, виявлені закономірності будуть специфічною рисою саме конкретних досліджуваних даних і більше не зустрінуться, але це все одно може бути корисно. До такого виду завдань належать кластеризація і пошук асоціативних правил.

Рішення прогностичних завдань розбивається на два етапи. На першому етапі на підставі набору даних з відомими результатами будується модель. На другому етапі вона використовується для передбачення результатів на підставі нових наборів даних. До цього виду завдань відносять завдання класифікації і регресії. Сюди можна віднести і завдання пошуку асоціативних правил, якщо результати її рішення можуть бути використані для передбачення появи деяких подій.

За способами вирішення завдання поділяють на навчання з учителем і навчання без учителя. Така назва походить від терміна Machine Learning (машинне навчання), який часто використовують в англомовній літературі, він означає всі технології Data Mining.

Під час навчання з учителем завдання аналізу даних вирішується в декілька етапів. Спочатку за допомогою будь-якого алгоритму Data Mining будується модель аналізованих даних – класифікатор. Потім класифікатор піддається навчанню. Іншими словами, перевіряється якість його роботи, і якщо вона незадовільна, відбувається додаткове навчання класифікатора. Так триває

доти, доки не буде досягнутий необхідний рівень якості або не стане зрозуміло, що обраний алгоритм не працює коректно з даними, або ж самі дані не мають структури, яку можна виявити. До цього типу завдань відносять завдання класифікації і регресії.

Навчання без учителя об'єднує завдання, де будуються описові моделі, наприклад закономірності в покупках, зроблених клієнтами великого магазину. Очевидно, що якщо ці закономірності є, то модель повинна їх представити, і недоречно свідчити про її навчання. Перевагою таких завдань є можливість їхнього вирішення без будь-яких попередніх знань про аналізовані дані. До цих завдань належать кластеризація і пошук асоціативних правил.

10.2. Завдання класифікації і регресії

Під час аналізу часто потрібно визначити, до якого з відомих класів належать досліджувані об'єкти, тобто класифікувати їх. Наприклад, коли людина звертається в банк за наданням їй кредиту, банківський службовець повинен прийняти рішення: кредитоспроможний або потенційний клієнт чи ні. Очевидно, що таке рішення приймається на підставі даних про досліджуваний об'єкт (у цьому випадку про людину): його місце роботи, розмір заробітної плати, вік, склад сім'ї тощо. У результаті аналізу цієї інформації банківський службовець повинен віднести людину до одного з двох відомих класів: «кредитоспроможний» і «некредитоспроможний».

У загальному випадку кількість класів у завданнях класифікації може бути більше двох. Наприклад, у завданні розпізнавання образу цифр таких класів може бути 10 (за кількістю цифр у десятковій системі числення). У такому завданні об'єктом класифікації є матриця пікселів, що представляє образ цифри. При цьому колір кожного пікселя є характеристикою аналізованого об'єкта.

У Data Mining завдання класифікації розглядають як завдання визначення значення одного з параметрів аналізованого об'єкта на підставі значень інших параметрів. Параметр часто називають залежною змінною, а параметри, які беруть участь у його

визначенні, – незалежними змінними. У розглянутих прикладах незалежними змінними були:

– зарплата, вік, кількість дітей тощо;

– значення кольору пікселів матриці.

Залежними змінними в цих самих прикладах були, відповідно:

– кредитоспроможність клієнта (можливі значення цієї змінної – «так» і «ні»);

– цифра образу (можливі значення цієї змінної – 0, 1, ..., 9).

Необхідно звернути увагу, що у всіх розглянутих прикладах незалежна змінна приймала значення із скінченої множини значень. Якщо значеннями незалежних і залежною змінних є дійсні числа, то завдання називається завданням регресії. Прикладом завдання регресії може слугувати задача визначення суми кредиту, яка може бути видана банком клієнту.

Завдання класифікації і регресії вирішується в два етапи. На першому виділяється навчальна вибірка. У неї входять об'єкти, для яких відомі значення як незалежних, так і залежних змінних. В описаних раніше прикладах такими навчальними вибірками можуть бути:

– інформація про клієнтів, яким раніше видавалися кредити на різні суми, й інформація про їхнє погашення;

– розпізнані раніше матриці образів цифр.

На підставі навчальної вибірки будується модель визначення значення залежної змінної. Її часто називають функцією класифікації або регресії. Для отримання максимально точної функції до навчальної вибірки ставляться такі основні вимоги:

– кількість об'єктів, що входять до вибірки, має бути досить великою. Чим більше об'єктів, тим точніше буде побудована на її основі функція класифікації або регресії;

– у вибірку повинні входити об'єкти, що являють собою всі можливі класи в разі завдання класифікації або всю сферу значень у разі завдання регресії;

– для кожного класу в завданні класифікації або для кожного інтервалу сфери значень у завданні регресії вибірка повинна містити достатню кількість об'єктів.

На другому етапі побудовану модель застосовують до аналізованих об’єктів (до об’єктів з невизначеним значенням залежної змінної).

Основні проблеми, з якими стикаються під час вирішення завдань класифікації і регресії, – це незадовільна якість вихідних даних, в яких зустрічаються як помилкові дані, так і пропущені значення, різні типи атрибутів: числові і категоріальні, різна значимість атрибутів, а також так звані проблеми overfitting і underfitting. Сутність першої з них полягає в тому, що класифікаційна функція у разі побудови «занадто добре» адаптується до даних, і тому помилки й аномальні значення, які зустрічаються в них, намагаються інтерпретувати як частину внутрішньої структури даних. Очевидно, що в подальшому така модель функціонуватиме неналежним чином з іншими даними, де характер помилок буде дещо іншим. Терміном underfitting позначають ситуацію, коли занадто велика кількість помилок під час перевірки класифікатора на навчальній множині. Це означає, що особливих закономірностей у даних не було виявлено, і або їх немає взагалі, або необхідно вибрати інший метод їхнього виявлення.

10.3. Задача пошуку асоціативних правил

Сутність завдання полягає у визначенні наборів об’єктів, які часто зустрічаються. Це завдання є окремим випадком задачі класифікації. Спочатку вона вирішувалася під час аналізу тенденцій в поведінці покупців у супермаркетах. Аналізувалися дані здійснених ними покупок, які покупці складають у візок (кошик). Це і зумовило іншу назву – аналіз ринкових кошиків. Аналізуючи ці дані, інтерес насамперед становила інформація про те, які товари купуються разом, у якій послідовності, які категорії споживачів, які товари краще купують, у які періоди тощо. Така інформація дозволяє більш ефективно планувати закупівлю товарів, проведення рекламної кампанії та ін.

Наприклад, з набору покупок, що здійснюються в магазині, можна виділити набори товарів, які купують разом: {чіпси, пиво}; {вода, горіхи}. Отже, можна зробити висновок, що якщо

купують чіпси або горіхи, то, як правило, купують і пиво або воду, відповідно. Володіючи такими знаннями, можна розмістити ці товари поруч, об'єднати їх в один пакет зі знижкою або вжити інші дії, що стимулюють покупця придбати товар.

Завдання пошуку асоціативних правил актуальне не тільки у сфері торгівлі. Наприклад, у сфері обслуговування інтерес являє собою інформація про те, якими послугами клієнти воліють користуватися в сукупності. Для отримання цієї інформації завдання вирішується стосовно до даних про послуги, якими користується один клієнт упродовж певного часу (місяці, роки). Це допомагає визначити, наприклад, як найбільш вигідно скласти пакети послуг, пропонованих клієнту.

У процесі аналізу часто виникає інтерес послідовність подій, що відбуваються. Під час виявлення закономірностей у таких послідовностях можна з деякою ймовірністю прогнозувати появу подій у майбутньому, що дозволяє ухвалювати більш правильні рішення. Таке завдання є різновидом завдання пошуку асоціативних правил і називається сіквенціальним аналізом.

Основною відмінністю завдання сіквенціального аналізу від пошуку асоціативних правил є встановлення відносин порядку між досліджуваними наборами. Це відношення може бути визначено різними способами. Під час аналізу послідовності подій, що відбуваються в часі, об'єктами таких наборів є події, а відношення порядку відповідає хронології їхньої появи.

10.4. Задача кластеризації

Задача кластеризації полягає в поділі досліджуваної множини об'єктів на групи «подібних» об'єктів, які називаються кластерами (cluster). Слово cluster перекладається з англійської як згусток, пучок, група. Споріднені поняття, використовувані в літературі, – клас, таксон, згущення. Часто вирішення задачі розбиття множини елементів на кластери називають кластерним аналізом.

Кластеризація може застосовуватися практично в будь-якій сфері, де необхідні дослідження експериментальних або статистичних даних.

Концептуально сегментування засноване на передумові, що всі споживачі різні. У них різні потреби, різні вимоги до товару, вони ведуть себе по-різному: в процесі вибору товару, придбання товару, використання товару, формування реакції на товар. У зв'язку з цим необхідно по-різному підходити до роботи зі споживачами: пропонувати їм різні за своїми характеристиками товари, по-різному просувати і продавати товари. Для того щоб визначити, чим відрізняються споживачі один від одного і як ці відмінності відбиваються на вимогах до товару, і проводиться сегментація споживачів. У маркетингу критеріями (характеристиками) сегментації є: географічне розташування, соціально-демографічні характеристики, мотиви здійснення покупки тощо. На підставі результатів сегментації маркетолог може визначити, наприклад, такі характеристики сегментів ринку, як реальна і потенційна ємність сегмента, групи споживачів, чиї потреби не задовольняються повною мірою жодним виробником, працюючим на цьому сегменті ринку тощо. На підставі цих параметрів маркетолог може зробити висновок про привабливість роботи фірми в кожному з виділених сегментів ринку.

Кластеризація відрізняється від класифікації тим, що для проведення аналізу не потрібно мати виділену залежну змінну. Це завдання вирішується на початкових етапах дослідження, коли про дані мало що відомо. Його рішення допомагає краще зрозуміти дані, і з цієї точки зору завдання кластеризації є описовим.

Для завдання кластеризації характерна відсутність будь-яких відмінностей як між змінними, так і між об'єктами, тут шукають групи найбільш близьких, подібних об'єктів. Методи автоматичного розбиття на кластери рідко використовуються самі по собі, а тільки для отримання груп подібних об'єктів. Після визначення кластерів використовуються інші методи Data Mining, щоб спробувати встановити, що означає таке розбиття, чим воно викликане. Кластерний аналіз дозволяє розглядати досить великий обсяг інформації і різко скорочувати, стискати великі масиви інформації, робити їх компактними і наочними.

Зазначимо ряд особливостей, властивих завданню кластеризації.

По-перше, рішення сильно залежить від природи об’єктів даних (і їх атрибутів). Так, з одного боку, це можуть бути однозначно визначені, кількісно окреслені об’єкти, а з іншого – об’єкти, які мають розподіл або нечіткий опис.

По-друге, рішення значною мірою залежить і від уявлення кластерів і передбачуваних відносин об’єктів даних і кластерів. Так, необхідно враховувати такі властивості, як можливість або неможливість приналежності об’єктів до кількох кластерів. Необхідне визначення самого поняття приналежності кластеру: однозначна (належить / не належить), імовірнісна (ймовірність приналежності), нечітка (ступінь приналежності).

10.5. Моделі Data Mining

Мета технології Data Mining – знаходження в даних таких моделей, які не можуть бути знайдені звичайними методами. Існують два види моделей: прогностичні й описові.

Прогностичні моделі будуються на підставі набору даних з відомими результатами. Вони використовуються для передбачення результатів на підставі інших наборів даних. При цьому потрібно, щоб модель працювала максимально точно, була статистично значуща, ефективна тощо.

До таких моделей належать:

– моделі класифікації;

– моделі послідовностей.

Описові моделі приділяють увагу сутності залежностей у наборі даних, взаємного впливу різних чинників, тобто побудови емпіричних моделей різних систем. Ключовий момент у таких моделях – легкість і прозорість для сприйняття аналітиком. Можливо, виявлені закономірності будуть специфічною рисою саме конкретних досліджуваних даних і більше ніде не зустрінуться, але це все одно може бути корисно, і тому повинно бути відомо.

До таких моделей належать:

– регресивні моделі;

– підсумкові моделі – виявлення обмежень на дані аналізованого масиву;

– моделі кластерів;

– асоціативні моделі;

– моделі винятків – описують виняткові ситуації в записах (наприклад, окремих пацієнтів), які різко відрізняються чимось від основної множини записів (групи хворих).

Знання винятків може бути використано двояким чином. Можливо, ці записи є випадковим збоєм, наприклад помилки операторів, що вводили дані в комп'ютер. З іншого боку, окремі, виняткові записи можуть являти собою самостійний інтерес для дослідження, оскільки вони можуть вказувати на деякі рідкісні, але важливі аномальні явища. Навіть сама ідентифікація цих записів, не зазначаючи про їх подальший аналіз і детальний розгляд, може виявитися досить корисною для розуміння сутності досліджуваних об'єктів або явищ.

Для побудови розглянутих моделей використовуються різні методи й алгоритми Data Mining. З огляду на те, що технологія Data Mining розвивалася і розвивається на стику таких дисциплін, як статистика, теорія інформації, машинне навчання і теорія баз даних, цілком закономірно, що більшість алгоритмів і методів Data Mining були розроблені на основі різних технологій і концепцій.

10.6. Процес виявлення знань

Основні етапи аналізу. Для виявлення знань у даних недостатньо просто застосувати методи Data Mining, хоча, безумовно, цей етап є основним у процесі інтелектуального аналізу. Весь процес складається з кількох етапів. Розглянемо основні з них, щоб продемонструвати, що без спеціальної підготовки аналітика методи Data Mining самі по собі не вирішують існуючих проблем. Отже, весь процес можна розбити на такі етапи:

– розуміння і формулювання завдання аналізу;

– підготовка даних для автоматизованого аналізу;

– застосування методів Data Mining і побудова моделей;

– перевірка побудованих моделей;

– інтерпретація моделей людиною.

На першому етапі виконується осмислення поставленого завдання й уточнення цілей, які повинні бути досягнуті методами

Data Mining. Важливо правильно сформулювати цілі і вибрати необхідні для їхнього досягнення методи, оскільки від цього залежить подальша ефективність усього процесу.

Другий етап полягає у приведенні даних до форми, придатної для застосування конкретних методів Data Mining. Цей процес далі буде описаний більш детально, тут зауважимо тільки, що вид перетворень, які здійснюються над даними, багато в чому залежить від використовуваних методів, обраних на попередньому етапі.

Третій етап – це, власне, застосування методів Data Mining. Сценарії цього застосування можуть бути різними і включати складну комбінацію будь-яких методів, особливо якщо використовуються методи, які дозволяють проаналізувати дані з різних точок зору.

Наступний етап – перевірка побудованих моделей. Досить простий і часто використовуваний спосіб полягає в тому, що всі наявні дані, які необхідно аналізувати, розбиваються на дві групи. Як правило, одна з них більшого розміру, інша – меншого. На більшій групі, застосовуючи певні методи Data Mining, отримують моделі, а на меншій – перевіряють їх. За різницею в точності між тестовою та навчальною групами можна робити висновок про адекватність побудованої моделі.

Останній етап – інтерпретація отриманих моделей людиною з метою їхнього використання для прийняття рішень, додавання одержаних правил і залежностей у бази знань тощо. Цей етап часто має на увазі використання методів, які перебувають на стику технологій Data Mining і експертних систем. Від того, наскільки ефективним він буде, значною мірою залежить успіх вирішення поставленого завдання.

Цим етапом завершується цикл Data Mining. Остаточна оцінка цінності видобутого нового знання виходить за рамки аналізу автоматизованого або традиційного і може бути проведена тільки після втілення в життя рішення, прийнятого на основі здобутого знання, після перевірки нового знання практикою. Дослідження досягнутих практичних результатів завершує оцінку цінності видобутого засобами Data Mining нового знання.

Запитання до параграфа 10

1. Які існують основні методи Data Mining?
2. У чому сутність класифікації даних?
3. Яка сутність регресії?
4. У чому відмінність класифікації від регресії?
5. Які етапи вирішення завдань класифікації і регресії?
6. Яка сутність завдання пошуку асоціативних правил?
7. Яка сутність завдання кластеризації?
8. Які є моделі Data Mining?
9. Які основні етапи аналізу даних?

§ 11. Пошук асоціативних правил

11.1. Формальна постановка задачі

Однією з найбільш поширених задач аналізу даних є визначення наборів об'єктів, які часто зустрічаються великими.

Опишемо цю задачу в узагальненому вигляді. Для цього позначимо об'єкти, що утворюють досліджувані набори (itemsets), такою множиною:

$$I = \{i_1; i_2; \ldots; i_n\},$$

де i_j – об'єкти, що входять в аналізовані набори; n – загальна кількість об'єктів.

У сфері торгівлі, наприклад, такими об'єктами є товари, подані в прайс-листі (табл. 3.1).

Таблиця 3.1

Прайс-лист

Ідентифікатор	Найменування товару	Ціна
0	Шоколад	30.00
1	Чіпси	12.00
2	Банани	10.00
3	Вода	4.00
4	Пиво	14.00
5	Горіхи	15.00

Вони відповідають такій множині об’єктів:

$$I = \{шоколад, чіпси, банани, вода, пиво, горіхи\}.$$

Набори об’єктів з множини I, що зберігаються в БД і піддаються аналізу, називаються транзакціями. Опишемо транзакцію як підмножину множини I:

$$T = \{i_j | i_j \in I\}.$$

Такі транзакції в магазині відповідають наборам товарів, що купуються споживачем і зберігаються в БД у вигляді товарного чека або накладної. У них перераховуються придбані покупцем товари, їхня ціна, кількість тощо. Наприклад, такі транзакції відповідають покупкам, що здійснюються споживачами в супермаркеті:

$$T_1 = \{чіпси, вода, пиво\};$$
$$T_2 = \{банани, вода, горіхи\}.$$

Набір транзакцій, інформація про які доступна для аналізу, позначимо такою множиною:

$$D = \{T_1, T_2, ..., T_r, ..., T_m\},$$

де m – кількість доступних для аналізу транзакцій.

Наприклад, у магазині такою множиною буде:

$$D = \{\{чіпси, вода, пиво\},$$
$$\{банани, вода, горіхи\},$$
$$\{горіхи, банани, чіпси, банани, вода\},$$
$$\{банани, горіхи, банани\}\}.$$

Для використання методів Data Mining множина D може бути надана у вигляді табл. 3.2.

Таблиця 3.2

Множина *D*

Номер транзакції	Номер товару	Найменування товару	Ціна
0	1	Чіпси	12,00
0	3	Вода	4,00
0	4	Пиво	14,00
1	2	Банани	10,00
1	3	Вода	4,00
1	5	Горіхи	15,00
2	5	Горіхи	15,00
2	2	Банани	10,00
2	1	Чіпси	12,00
2	2	Банани	10,00
2	3	Вода	4,00
3	2	Банани	10,00
3	5	Горіхи	15,00
3	2	Банани	10,00

Множину транзакцій, в які входить об’єкт, опишемо таким чином:

$$D_{i_j} = \{T_r | i_j \in T_r; j = 1 \ldots n; r = 1 \ldots . m\} \subseteq D$$

.

У цьому прикладі множиною транзакцій, що містять об’єкт Вода, є така множина:

$D_{вода}$ = {{чіпси, вода, пиво},
{банани, вода, горіхи},
{горіхи, банани, чіпси, банани, вода}}.

Деякий довільний набір об’єктів позначимо таким чином:

$$F = \{i_j | i_j \in I; j = 1 \ldots n\}.$$

Наприклад, *F = {банани, вода}*. Набір, який складається із *k* об’єктів, називається k-елементним набором (у цьому прикладі це є 2-елементний набір).

Множина транзакцій, в які входить набір *F*, позначимо таким чином:

$$D_F = \{T_r | F \in T_r; r = 1 \ldots . m\}.$$

У цьому прикладі:

$$D_{банани,\ вода} = \{\{банани,\ вода,\ горіхи\},$$
$$\{горіхи,\ банани,\ чіпси,\ банани,\ вода\}\}.$$

Відношення кількості транзакцій, в які входить набір F, до загальної кількості транзакцій називається підтримкою (support) набору F і позначається Supp (F):

$$Supp(F) = \frac{|D_F|}{|D|}$$

Для набору *{банани, вода}* підтримка дорівнює 0,5, оскільки цей набір входить у дві транзакції (з номерами 1 і 2), а всього транзакцій 4.

Під час пошуку аналітик може вказати мінімальне значення підтримки цікавих йому наборів $Supp_{min}$. Набір називається частим, якщо значення його підтримки більше мінімального значення підтримки, заданого користувачем:

$$\text{Supp (F)} > \text{Supp}_{\text{min}}.$$

Таким чином, у разі пошуку асоціативних правил потрібно знайти множину всіх частих наборів:

$$\text{L} = \{\text{F} \mid \text{Supp (F)} > \text{Supp}_{\text{min}}\}.$$

У цьому прикладі частими наборами при $\text{Supp}_{\text{min}} = 0{,}5$ є такі:

{чіпси} $Supp_{min} = 0{,}5$;
{чіпси, вода} $Supp_{min} = 0{,}5$;
{банани} $Supp_{min} = 0{,}75$;
{банани, вода} $Supp_{min} = 0{,}5$;
{банани, вода, горіхи} $Supp_{min} = 0{,}5$;
{банани, горіхи} $Supp_{min} = 0{,}75$;
{вода} $Supp_{min} = 0{,}75$;
{вода, горіхи} $Supp_{min} = 0{,}5$;
{горіхи} $Supp_{min} = 0{,}75$.

11.2. Секвенційний аналіз

Під час аналізу часто викликає інтерес послідовність подій, що відбуваються. У процесі виявлення закономірностей у таких послідовностях можна з деякою часткою ймовірності прогнозувати появу подій у майбутньому, що дозволяє приймати більш правильні рішення.

Послідовністю називається впорядкована множина об'єктів. Для цього на множині має бути задано відношення порядку.

Тоді послідовність об'єктів можна описати в такому вигляді:

$$S = \{..; i_p; \ldots; i_q\}.$$

Наприклад, у випадку з покупками в магазинах таким відношенням порядку може виступати час покупок. Тоді послідовність

S = {(вода, 02.03.2003), (чіпси, 05.03.2003), (пиво, 10.03.2003)}

можна інтерпретувати як покупки, що здійснюються однією людиною в різний час (спочатку була куплена вода, потім чіпси, а потім пиво).

Розрізняють два види послідовностей: з циклами і без циклів. У першому випадку допускається входження в послідовність одного і того самого об'єкта на різних позиціях:

$$S = \{..; i_p; \ldots; i_q; \ldots\}\ p < q,\ i_q = i_p.$$

Зазначають, що транзакція T містить послідовність S, якщо S входить у T і об'єкти, що входять у S, входять і в множину T зі збереженням відносини порядку. При цьому допускається, що в множині T між об'єктами з послідовності можуть перебувати інші об'єкти.

Підтримкою послідовності S називається відношення кількості транзакцій, в які входить послідовність S, до загальної кількості транзакцій. Послідовність є частою, якщо її підтримка перевищує мінімальну підтримку, задану користувачем:

$$\text{Supp}\ (S) > \text{Supp}_{\min}.$$

Завданням секвенційного аналізу є пошук усіх частих послідовностей:

$$L = \{S \mid Supp\ (S) > Supp_{min}\}.$$

Основною відмінністю завдання секвенційного аналізу від пошуку асоціативних правил є встановлення відносин порядку між об'єктами множини *I*. Це встановлення може бути визначено різними способами. Під час аналізу послідовності подій, що відбуваються в часі, об'єктами множини *I* є події, а відношення порядку відповідає хронології їхньої появи.

Наприклад, у разі аналізу послідовності покупок у супермаркеті наборами є покупки, що здійснюються в різний час одними і тими самим покупцями, а відношенням порядку в них є хронологія покупок:

D = {{(вода), (пиво)},
{(банани, вода), (пиво), (вода, шоколад, банани)},
{(пиво, чіпси, вода), (пиво)}}.

Природно, що в цьому випадку виникає проблема ідентифікації покупців. На практиці вона вирішується введенням дисконтних карт, що мають унікальний ідентифікатор. Цю множину можна уявити у вигляді табл. 3.3.

Таблиця 3.3

Множина покупок

ID покупця	Послідовність покупок
0	(пиво), (вода)
1	(банани, вода), (пиво), (вода, шоколад, банани)
2	(пиво, чіпси, вода), (пиво)

Інтерпретувати таку послідовність можна таким чином: покупець з ідентифікатором 1 спочатку купив банани і воду, потім купив пиво, в останній раз він купив воду, шоколад і банани.

Підтримка, наприклад, послідовності {(пиво), (вода)} становить 2/3, оскільки вона зустрічається у покупців з ідентифікаторами 0 і 1. В останнього покупця також зустрічається набір {(пиво), (вода)}, але не зберігається послідовність (він купив спочатку воду, а потім пиво).

Секвенціальний аналіз актуальний і для телекомунікаційних компаній. Основна проблема, для вирішення якої він використовується, – це аналіз даних про аварії на різних вузлах телекомунікаційної мережі. Інформація про послідовність здійснення аварій може допомогти у виявленні неполадок та попередження нових аварій. Дані для такого аналізу можуть бути надані, наприклад, у вигляді табл. 3.4.

Таблиця 3.4

Інформація про послідовність здійснення аварій

Дата	Час	Джерело помилки	Код джерела	Код помилки	...
01.01.03	15:04:23	Станція 1	1001	a	…
01.01.03	16:45:46	Станція 1	1001	f	…
01.01.03	18:32:26	Станція 4	1004	z	…
01.01.03	20:07:11	Станція 5	1005	н	…
01.01.03	20:54:43	Станція 1	1001	q	…
…	…	…	…	…	…

У цій задачі об’єктами множини I є коди помилок, що виникають у процесі роботи телекомунікаційної мережі. Послідовність S_{sid} містить збої, що відбуваються на станції з ідентифікатором *sid*. Їх можна надати у вигляді пар (eid, t), де eid – код помилки, а t – час, коли вона сталася. Таким чином, послідовність збоїв на станції з ідентифікатором sid матиме такий вигляд:

$$S_{sid} = \{(eid_1, t_1), (eid_2, t_2), (eid_n, t_n)\}.$$

Для даних, наведених у табл. 3.4, транзакції будуть такі:

T_{1001} = {(a, 15:04:23 01.01.03), (f, 16:45:46 01.01.03),
(q, 20:54:43 01.01.03), ...};
T_{1004} = {(z, 18:32:26 01.01.03), ...};
T_{1005} = {(h, 20:07:11 01.01.03), ...}.

Під час аналізу такої послідовності важливим є визначення інтервалу часу між збоями. Воно дозволяє передбачити момент і характер нових збоїв, а отже, провести профілактичні заходи. З цієї причини у процесі аналізу даних інтерес викликає не просто послідовність подій, а збої, які відбуваються один за одним.

11.3. Різновиди задач пошуку асоціативних правил

У багатьох прикладних сферах об'єкти множини *I* природним чином поєднуються в групи, які, в свою чергу, також можуть об'єднуватися в більш загальні групи. Таким чином, виходить ієрархічна структура об'єктів, представлена на рис. 3.1.

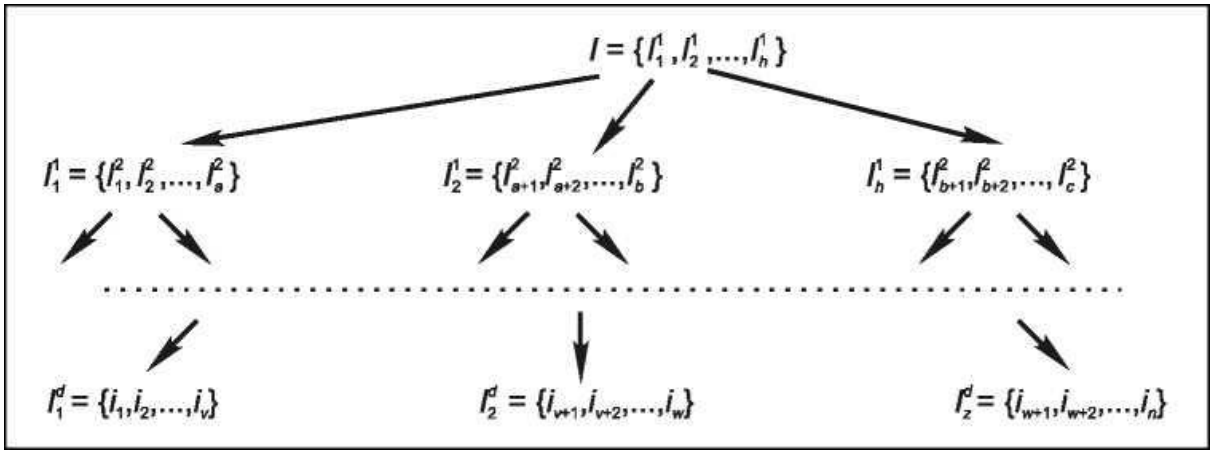


Рис. 3.1. Ієрархічне представлення об'єктів множини *I*

Для прикладу, наведеного в табл. 3.5, такою ієрархією може бути така категоризація товарів:

Таблиця 3.5

Напої	Їжа
• алкогольні: – пиво	• шоколад
	• чіпси
• безалкогольні: – вода	• банани
	• горіхи

Наявність ієрархії змінює уявлення про те, коли об'єкт присутній у транзакції *T*. Очевидно, що підтримка не окремого об'єкта, а групи, в яку він входить, більша:

$$Supp(I_q^g) \geq Supp(i_j),$$

де $i_j \in I_q^g$.

Це пов'язано з тим, що під час аналізу груп підраховуються не тільки транзакції, в які входить окремий об'єкт, але і транзакції, що містять усі об'єкти аналізованої групи. Наприклад, якщо підтримка Supp {банани, вода} = 2/4, то підтримка Supp {їжа, напої} = 2/4, оскільки об'єкти груп їжа і напої входять у транзакції з ідентифікаторами 0, 1 і 2.

Використання ієрархії дозволяє визначити зв'язки, що входять у більш високі рівні ієрархії, оскільки підтримка набору може збільшуватися, якщо підраховується входження групи, а не її об'єкта. Пошук наборів часто зустрічається в транзакціях, що складаються з об'єктів F = $\{i|i \in I\}$ або груп одного рівня ієрархії F = $\{I^g|I^g \in I^{g+1}\}$, можна розглядати також змішані набори об'єктів і груп F = $\{i, I^g|i \in I^g \in I^{g+1}\}$. Це дозволяє розширити аналіз і отримати додаткові знання.

За ієрархічною структурою об'єктів можна змінювати характер пошуку, змінюючи аналізований рівень. Очевидно, що чим більше об'єктів у множині *I*, тим більше об'єктів у транзакціях *T* і частих наборах. Це, в свою чергу, збільшує час пошуку й ускладнює аналіз результатів. Зменшити або збільшити кількість даних можна за допомогою ієрархічного уявлення аналізованих об'єктів. Переміщаючись вгору по ієрархії, узагальнюємо дані і зменшуємо їхню кількість, і навпаки.

Недоліком узагальнення об'єктів є менша корисність отриманих знань, оскільки в цьому випадку вони належать до груп товарів, що не завжди прийнятно. Для досягнення компромісу між аналізом груп і аналізом окремих об'єктів часто використовують такий спосіб: спочатку аналізують групи, а потім, залежно від отриманих результатів, досліджують об'єкти, які зацікавили аналітика груп. У будь-якому випадку можна стверджувати, що наявність ієрархії в об'єктах і її використання в задачі пошуку асоціативних правил дозволяє виконувати більш гнучкий аналіз та отримувати додаткові знання.

У розглянутій задачі пошуку асоціативних правил наявність об'єкта в транзакції визначалася тільки його присутністю в ній ($i_j \in T$) або відсутністю ($i_j \notin T$). Часто об'єкти мають додаткові атрибути, як правило, чисельні. Наприклад, товари в транзакції мають атрибути: ціна і кількість. При цьому наявність об'єкта в наборі може визначатися не просто фактом його присутності, а виконанням умови щодо певного атрибуту. Наприклад, під час аналізу транзакцій, що здійснюються покупцем, може цікавити не просто наявність товару, що купується, а товару, що купується за деякою ціною.

Для розширення можливостей аналізу за допомогою пошуку асоціативних правил у досліджувані набори можна додавати додаткові об’єкти. У загальному випадку вони можуть мати природу, відмінну від основних об’єктів. Наприклад, для визначення товарів, що мають більший попит залежно від місця розташування магазину, в транзакції можна додати об’єкт, що характеризує район.

11.4. Надання результатів

Рішення задачі пошуку асоціативних правил, як і будь-якого завдання, зводиться до обробки вихідних даних та отримання результатів. Обробка над вихідними даними виконується за певним алгоритмом Data Mining.

Результати, одержані під час вирішення цього завдання, прийнято надавати у вигляді асоціативних правил. У зв’язку з цим у разі їхнього пошуку виділяють два основних етапи:

– знаходження всіх частих наборів об’єктів;

– генерація асоціативних правил із знайдених частих наборів об’єктів. Асоціативні правила мають такий вигляд:

якщо (умова), то (результат),

де *умова* – зазвичай не логічний вираз (як у класифікаційних правилах), а набір об’єктів з множини *I*, з якими пов’язані (асоційовані) об’єкти, включені в результат цього правила.

Наприклад, асоціативне правило:

якщо (банани, вода), то (горіхи)

означає, що якщо споживач купує банани і воду, то він купує і горіхи.

Як уже зазначалося, в асоціативних правилах умова і результат є об’єктами множини *I*:

якщо X, то У, де $X \in Y, X \in I, Y \in I, X \cup Y = \varphi$.

Асоціативне правило можна уявити як імплікації над множиною *X*, де X є I, У є I, X і У = ф.

Основною перевагою асоціативних правил є їх легке сприйняття людиною і проста інтерпретація мовами програмування. Однак вони не завжди корисні. Виділяють три види правил:

– корисні правила – містять дійсну інформацію, яка раніше була невідома, але має логічне пояснення. Такі правила можуть бути використані для прийняття рішень, що приносять користь;

– тривіальні правила – містять дійсну та легко зрозумілу інформацію, яка вже відома. Такі правила, хоча і можна пояснити, але не можуть принести будь-якої користі, оскільки відображають або відомі закони в досліджуваній сфері, або результати минулої діяльності. Іноді такі правила можуть використовуватися для перевірки виконання рішень, прийнятих на підставі попереднього аналізу;

– незрозумілі правила – містять інформацію, яка не може бути пояснена. Такі правила можуть бути отримані або на основі аномальних значень, або глибоко прихованих знань. Безпосередньо такі правила не можна використовувати для прийняття рішень, оскільки їх нез'ясовність може призвести до непередбачуваних результатів. Для кращого розуміння потрібен додатковий аналіз.

Асоціативні правила будуються на основі частих наборів. Так, правила, побудовані на підставі набору F ($X \cup Y = F$), є усіма можливими комбінаціями об'єктів, що входять у нього.

Наприклад, для набору *{банани, вода, горіхи}* можуть бути побудовані такі правила:

якщо (банани), то (вода);
якщо (банани), то (горіхи);
якщо (банани), то (вода, горіхи);
якщо (вода, горіхи), то (банани);
якщо (вода), то (банани);
якщо (вода), то (горіхи);
якщо (вода), то (банани, горіхи);
якщо (банани, горіхи), то (вода);
якщо (горіхи), то (вода);
якщо (горіхи), то (банани);
якщо (горіхи), то (вода, банани);
якщо (вода, банани), то (горіхи).

Таким чином, кількість асоціативних правил може бути досить великою і важко сприймається людиною. До того ж, не всі з побудованих правил несуть у собі корисну інформацію. Для оцінки їхньої корисності вводяться такі величини.

Підтримка – показує, який відсоток транзакцій підтримує це правило. Оскільки правило будується на підставі набору, то правило $X \rightarrow Y$ має підтримку, яка дорівнює підтримці набору F, який складають X і Y:

$$Supp_{X \rightarrow Y} = Supp_F = \frac{|D_{F=X \cup Y}|}{|D|}.$$

Очевидно, що правила, побудовані на підставі одного і того самого набору, мають однакову підтримку, наприклад підтримка

$\text{Supp}_{\text{якщо(вода, банани), то (горіхи)}} \text{Supp}_{\{\text{вода, банани, горіхи}\}} = 2/4.$

Достовірність – показує ймовірність того, що з наявності в транзакції набору X випливає наявність у ній набору Y. Достовірністю правила $X \rightarrow Y$ є ставлення кількості транзакцій, що містять набори X і Y, до кількості транзакцій, що містять набір X:

$$Conf_{X \rightarrow Y} = \frac{|D_{F=X \cup Y}|}{|D|} = \frac{Supp_{X \cup Y}}{Supp_X}.$$

Очевидно, що чим більше вірогідність, тим правило краще, причому у правил, побудованих на підставі одного і того самого набору, достовірність буде різна, наприклад:

$\text{Conf}_{\text{якщо (вода, банани), то (горіхи)}} = 1;$
$\text{Conf}_{\text{якщо(вода), то (горіхи, банани)}} = 2/3.$

На жаль, достовірність не дозволяє оцінити корисність правила. Якщо відсоток наявності в транзакціях набору Y за умови наявності в них набору X менше, ніж відсоток безумовної наявності набору Y, тобто:

$$Conf_{X \rightarrow Y} = \frac{Supp_{X \cup Y}}{Supp_X} < Supp_Y.$$

Це означає, що ймовірність випадково вгадати наявність у транзакції набору *Y* більше, ніж передбачити це за допомогою правила $X \rightarrow Y$. Для виправлення такої ситуації вводиться міра – поліпшення.

Поліпшення показує, чи корисніше правило випадкового вгадування. Поліпшення правила є відношенням кількості транзакцій, що містять набори *X* і *Y*, до добутку кількості транзакцій, що містять набір *X*, і кількості транзакцій, що містять набір *Y*:

$$impr_{X \rightarrow Y} = \frac{|D_{F=X \cup Y}|}{|D_X| \cdot |D_Y|} = \frac{Supp_{X \cup Y}}{Supp_X \cdot Supp_Y}.$$

Наприклад,

$$impr_{\text{якщо (вода, банани) то (горіхи)}} = 0{,}5/(0{,}5*0{,}5)=2.$$

Якщо поліпшення більше одиниці, то це означає, що за допомогою правила передбачити наявність набору *У* найімовірніше, ніж випадкове вгадування, якщо менше одиниці, то навпаки.

В останньому випадку можна використовувати правило заперечення, тобто правило, яке передбачає відсутність набору *Y*:

$$X \rightarrow \text{не } Y.$$

У такого правила поліпшення буде більше одиниці, оскільки $Supp_{\text{не У}} = 1 - Supp_Y$.

Таким чином, можна отримати правило, яке передбачає результат краще, ніж випадковим чином. Щоправда, на практиці такі правила мало застосовні. Наприклад, правило: *якщо (вода, горіхи), то не пиво* мало корисно, оскільки слабо відображає поведінку покупця.

Ці оцінки використовуються за генерації правил. Аналітик під час пошуку асоціативних правил задає мінімальні значення перерахованих величин. У результаті ті правила, які не задовольняють цим умовам, відкидаються і не включаються до рішення задачі. З цієї точки зору не можна об'єднувати різні правила, хоча і мають загальне смислове навантаження. Наприклад, такі правила:

$$X = \{i_1, i_2\} \rightarrow Y = \{i_3\};$$
$$X = \{i_1, i_2\} \rightarrow Y = \{i_4\}$$

не можна об’єднати в одне:

$$X = \{i_1, i_2\} \rightarrow Y = \{i_3, i_4\},$$

оскільки достовірності їх будуть різні, отже, деякі з них можуть бути виключені, а деякі – ні.

Якщо об’єкти мають додаткові атрибути, які впливають на склад об’єктів у транзакціях, а отже, і в наборах, то вони повинні враховуватися в генеруючі правила. У цьому випадку умовна частина правил міститиме не тільки перевірку наявності об’єкта в транзакції, але і більш складні операції порівняння: більше, менше, включає та ін. Результатна частина правил також може містити твердження щодо значень атрибутів. Наприклад, якщо у товарів розглядається ціна, то правила можуть мати такий вигляд:

якщо пиво.ціна < 10, то чіпси.ціна < 7.

Це правило свідчить про те, що якщо купується пиво за ціною менше 10 грн, то ймовірно будуть куплені чіпси за ціною менше 7 грн.

11.5. Алгоритм Apriori

Виявлення частих наборів об’єктів – операція, що вимагає значної кількості обчислень, а отже, і часу. Алгоритм Apriori описаний 1994 р. Він використовує одну з властивостей підтримки, з якого випливає: підтримка будь-якого набору об’єктів не може перевищувати мінімальної підтримки будь-якої з її підмножин:

$$Supp_F \leq Supp_E \, , E \subset F.$$

Наприклад, підтримка 3-об’єктного набору {пиво, вода, чіпси} буде завжди менше або дорівнюватиме підтримці 2-об’єктних наборів {пиво, вода}, {вода, чіпси}, {пиво, чіпси}. Це пояснюється тим, що будь-яка транзакція, яка має {пиво, вода, чіпси}, містить також і набори {пиво, вода}, {вода, чіпси}, {пиво, чіпси}, причому зворотне неправильно.

Алгоритм Apriori визначає, чи часто зустрічаються набори за кілька етапів. На *i*-му етапі визначаються всі *i*-елементні набори, які часто зустрічаються. Кожен етап складається з двох кроків: формування кандидатів і підрахунку підтримки кандидатів.

Розглянемо *i*-й етап. На кроці формування кандидатів алгоритм створює множину кандидатів з *i*-елементних наборів, чия підтримка досі не обчислюється. На кроці підрахунку кандидатів алгоритм сканує множину транзакцій, обчислюючи підтримку наборів-кандидатів. Після сканування відкидаються кандидати, підтримка яких менше певного заданого користувачем мінімуму, і зберігаються тільки ті, які часто зустрічаються 1-елементні набори. Під час 1-го етапу обрана множина наборів-кандидатів містить усі 1-елементні часті набори. Алгоритм обчислює їхню підтримку під час етапу підрахунку кандидатів.

Описаний алгоритм можна записати у вигляді такого псевдокоду:

L_1 = {*часто зустрічаються 1-елементні набори*}
для ($k = 2$; $L_{k-1} <> \boldsymbol{\phi}$; k ++)
C_k = *Apriorigen* (F_{k-l}) // *генерація кандидатів*
для всіх транзакцій $t \in D$ *виконати*
C_t = *subset* (C_k, t) // *видалення надлишкових правил*
для всіх кандидатів $c \in C_t$ *виконати*
c.count ++
кінець для всіх
кінець для всіх
$L_k = \{c \in C_k \mid c.count > = Supp_{min}\}$ // *відбір кандидатів*
кінець для
Результат $= \bigcup_k L_k$,

де:

L_k – множина k-елементних частих наборів, чия підтримка не менше заданої користувачем. Кожен член множини має набір упорядкованих ($i_j < i_p$, якщо $j < p$) елементів F і значення підтримки набору $Supp_F > Supp_{min}$;

$L_k = \{(F_1 , Supp_l), (F_2 , Supp_2), \ldots (F_q , Supp_q)\}$, де $F_j = \{i_1 , i_2 \ldots, i_k\}$;

C_k – множина кандидатів k-елементних наборів потенційно частих. Кожен член множини має набір упорядкованих ($i_j < i_p$, якщо $j < p$) елементів F і значення підтримки набору Supp.

Опишемо цей алгоритм за кроками.

Крок 1. Присвоїти k = 1 і виконати відбір всіх 1-елементних наборів, у яких підтримка більше мінімально заданої користувачем $Supp_{min}$.

Крок 2. k = k +1.

Крок 3. Якщо не вдається створювати k-елементні набори, то завершити алгоритм, інакше виконати наступний крок.

Крок 4. Створити множину k-елементних наборів кандидатів у часті набори. Для цього необхідно об'єднати в k-елементні кандидати (k – 1) – елементні часті набори. Кожен кандидат $c \in C_k$ буде формуватися шляхом додавання до (k – 1) – елементному частому набору – p елемента з іншого (k – 1) – елементного частого набору – q. Причому додається останній елемент набору q, який за порядком вище, ніж останній елемент набору p ($p.item_{k-1} < q.item_{k-1}$). При цьому перші все (k – 2) елемента обох наборів однакові ($p.item_1 = q.item_1$, $p.item_2 = q.item_2$, ..., $p.item_{k-2} = q.item_{k-2}$).

Це може бути записано у вигляді наступного SQL-подібного запиту:

insert into Ck
select $p.item_1, p.item_2, ..., p.item_{k-1}, q.item_{k-1}$
from $L_{k-1} p, L_{k-1} q$
where $p.item_1 = q.item_1, p.item_2 = q.item_2, ..., p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$.

Крок 5. Для кожної транзакції T з множини D вибрати кандидатів C_t з множини C_k, присутніх у транзакції T. Для кожного набору з побудованої множини C_k видалити набір, якщо хоча б одне з його (k – 1) підмножин не часто зустрічається, тобто відсутній у множині L_{k-1}. Це можна записати у вигляді такого псевдокоду:

для всіх наборів $c \in C_k$ *виконати*
для всіх (k–1) – піднаборів s з c виконати
якщо $(s \notin L_{k-1})$, *то*
видалити c з C_k.

Крок 6. Для кожного кандидата з множини C_k збільшити значення підтримки на одиницю.

Крок 7. Вибрати тільки кандидатів L_k з множини C_k, у яких значення підтримки більше заданої користувачем $Supp_{min}$. Повернутися до кроку 2.

Результатом роботи алгоритму є об'єднання всіх множин L_k.

Розглянемо роботу алгоритму на прикладі, наведеному в табл. 3.5, при $Supp_{min} = 0,5$. На першому кроці маємо таку множину кандидатів C_1 (вказуються ідентифікатори товарів) (табл. 3.6).

Таблиця 3.6

Ідентифікатори товарів

№	Набір	Supp
1	{0}	0
2	{1}	0,5
3	{2}	0,75
4	{4}	0,25
5	{3}	0,75
6	{5}	0,75

Заданою мінімальною підтримкою задовольняються лише кандидати 2, 3, 5 і 6, отже:

$$L_1 = \{\{1\}, \{2\}, \{3\}, \{5\}\}.$$

На другому кроці збільшуємо значення k до двох. Оскільки можна побудувати 2-елементні набори, то отримуємо множину C_2 (табл. 3.7).

Таблиця 3.7

Множина C_2

№	Набір	Supp
1	{1, 2}	0,25
2	{1, 3}	0,5
3	{1, 5}	0,25
4	{2, 3}	0,5
5	{2, 5}	0,75
6	{3, 5}	0,5

З побудованих кандидатів заданою мінімальною підтримкою задовольняються лише кандидати 2, 4, 5 і 6, отже:

$$L_2 = \{\{1, 3\}, \{2, 3\}, \{2, 5\}, \{3, 5\}\}.$$

На третьому кроці перейдемо до створення 3-елементних кандидатів і підрахунку їхньої підтримки. У результаті отримаємо таку множину C_3 (табл. 3.8).

Таблиця 3.8

Множина C_3

№	Набір	Supp
1	{2, 3, 5}	0,5

Цей набір задовольняє умову мінімальної підтримки, отже:

$$L_3 = \{\{2, 3, 5\}\}.$$

Оскільки 4-елементні набори створити не вдається, то результатом роботи алгоритму є множина:

$$L = \{\{1\}, \{2\}, \{3\}, \{5\}, \{1, 3\}, \{2, 3\}, \{2, 5\}, \{3, 5\}, \{2, 3 , 5\}\}.$$

Для підрахунку підтримки кандидатів потрібно порівняти кожну транзакцію з кожним кандидатом. Очевидно, що кількість кандидатів може бути досить велика і потрібен ефективний спосіб підрахунку. Набагато швидше й ефективніше використовувати підхід, заснований на зберіганні кандидатів у хеш-дереві. Внутрішні вузли дерева містять хеш-таблиці з покажчиками на нащадків, а листя – на кандидатів. Це дерево використовується під час швидкого підрахунку підтримки для кандидатів.

Хеш-дерево будується щоразу, коли формуються кандидати. Спочатку дерево складається тільки з кореня, який є листом, і не містить ніяких кандидатів-наборів. Щоразу, коли формується новий кандидат, він заноситься в корінь дерева, і так доти, доки кількість кандидатів у корені-листі не перевищить певного порога. Як тільки це відбувається, корінь перетворюється у хеш-таблицю, тобто стає внутрішнім вузлом, і для нього створюються нащадки-листя. Усі кандидати розподіляються за вузлами-нащадками згідно з хеш-значеннями елементів, що входять у набір. Кожен новий кандидат хешірується на внутрішніх вузлах, доки не досягне

першого вузла-листа, де він і буде зберігатися, доки кількість наборів знову ж таки не перевищить порога.

Після того, як хеш-дерево з кандидатами-наборами побудовано, легко підрахувати підтримку для кожного кандидата. Для цього потрібно «пропустити» кожну транзакцію через дерево і збільшити лічильники для тих кандидатів, чиї елементи також містяться і в транзакції, $C_к \cap T_i = C_к$. На кореневому рівні хеш-функція застосовується до кожного об'єкта з транзакції. Далі, на другому рівні, хеш-функція застосовується до інших об'єктів тощо. На *к*-му рівні хешірується до Елемент, і так доти, доки не досягнемо листа. Якщо кандидат, що зберігається в листі, є підмножиною цієї транзакції, збільшуємо лічильник підтримки цього кандидата на одиницю.

Після того, як кожна транзакція з вихідного набору даних «пропущена» через дерево, можна перевірити, чи задовольняють значення підтримки кандидатів мінімального порога. Кандидати, для яких ця умова виконується, переносяться в розряд і часто зустрічаються. Крім того, варто запам'ятати і підтримку набору, яка стане в нагоді під час добування правил. Ці самі дії застосовуються для знаходження $(k + 1)$-елементних наборів та ін.

11.6. Різновиди алгоритму Apriori

Алгоритм Apriori Tid є різновидом алгоритму Apriori. Відмінною рисою цього алгоритму є підрахунок значення підтримки кандидатів не під час сканування множини *D*, а за допомогою множини C_k, що є множиною кандидатів (*k*-елементних наборів) потенційно частих, у відповідність яким ставиться ідентифікатор TID транзакцій, в яких вони містяться.

Кожен член множини C_k є парою виду <TID, F_k>, де кожен F_k є потенційно частим *k*-елементним набором, наданим у транзакції з ідентифікатором TID. Множина C = D відповідає множині транзакцій, хоча кожен об'єкт у транзакції відповідає однооб'єктному набору в множини *C*, що містить цей об'єкт. Для k > 1 множина C_k генерується відповідно до алгоритму, описаному нижче. Член множини C_k, відповідний транзакції *T*, є парою такого вигляду:

$$<T.TID, \{ c \in C_k \mid c \in T\}>.$$

Підмножина наборів у C_k з однаковими TID називається записом. Якщо транзакція не містить жодного k-елементного кандидата, то C_k не матиме записів для цієї транзакції. Тобто кількість записів у C_k може бути менша, ніж у D, особливо для великих значень k. Крім того, для великих значень k кожен запис може бути меншим, ніж відповідна йому транзакція, оскільки в транзакції міститиметься мало кандидатів. Однак для малих значень k кожен запис може бути більшим, ніж відповідна транзакція, оскільки C_k включає всіх кандидатів k-елементних наборів, що містяться в транзакції.

Іншим різновидом алгоритму Apriori є алгоритм MSAP (Mining Sequential Alarm Patterns), спеціально розроблений для виконання секвенційного аналізу збоїв телекомунікаційної мережі.

Алгоритм MSAP для пошуку подій, що наслідують одна за одну, використовує поняття «термінове вікно» (Urgent Window). Це дозволяє виявляти не просто однакові послідовності подій, а наступні одна за одною. В іншому цей алгоритм працює за тим самим принципом, що і Apriori.

Запитання до параграфа 11

1. Яка формальна постановка задачі пошуку асоціативних правил?
2. Що називають підтримкою набору?
3. Що називають частим набором?
4. Що таке послідовність?
5. Які є види послідовностей?
6. Яка відмінність послідовностей від наборів?
7. Яка основна відмінність завдання секвенційного аналізу від пошуку асоціативних правил?
8. Які є задачі пошуку асоціативних правил?
9. Який вигляд мають асоціативні правила?
10. Що таке підтримка?
11. Що таке достовірність?
12. Що таке поліпшення?
13. Яка сутність алгоритму Apriori?
14. Які є різновиди алгоритму Apriori?

§ 12. Методи класифікації і прогнозування дерева рішень

12.1. Загальна характеристика

Метод дерев рішень є одним з найбільш популярних методів вирішення завдань класифікації та прогнозування. Іноді цей метод також називають деревами вирішальних правил, деревами класифікації і регресії.

Якщо залежна, тобто цільова, змінна набуває дискретних значень, за допомогою методу дерева рішень вирішується завдання класифікації.

Якщо ж залежна змінна приймає неперервні значення, то дерево рішень встановлює залежність цієї змінної від незалежних змінних, тоді йдеться про регресію і прогнозування.

Вперше дерева рішень були запропоновані Ховілендом і Хантом наприкінці 50-х років минулого століття.

Дерева рішень належать до кількості найбільш популярних і потужних інструментів Data Mining, що дозволяють ефективно вирішувати завдання класифікації і регресії. На відміну від методів, які використовують статистичний підхід, таких як класифікатор Байєса, лінійна і логістична регресія, дерева рішень засновані на машинному навчанні і в більшості випадків не вимагають припущень про статистичний розподіл значень ознак. В основі дерев рішень полягають вирішальні правила виду «якщо ..., то ...», які можуть бути сформульовані природною мовою.

У найбільш простому вигляді дерево рішень – це спосіб надання правил в ієрархічній, послідовній структурі. В основі такої структури містяться відповіді «Так» або «Ні» на ряд питань.

Для зручності наведемо базові поняття теорії дерев рішень у табл. 3.9.

Атрибутами в теорії дерев рішень називаються ознаки, що описують об’єкти, які класифікуються.

В основі роботи дерев рішень полягає процес рекурсивного розбиття вихідної множини спостережень або об’єктів на підмножини, асоційовані з класами. Розбиття проводиться за допомогою вирішальних правил, в яких здійснюється перевірка значень атрибутів за заданою умовою. Рекурсивними називаються

алгоритми, які працюють у покроковому режимі, при цьому на кожному наступному кроці використовуються результати, отримані на попередньому кроці.

Таблиця 3.9

Поняття, що зустрічаються в теорії дерев рішень

Назва	Опис
Об'єкт	Приклад, шаблон, спостереження, запис
Атрибут	Ознака, незалежна змінна, властивість, вхідне поле
Мітка класу	Залежна змінна, цільова змінна, вихідне поле
Вузол	Внутрішній вузол дерева
Лист	Кінцевий вузол дерева, вузол рішення
Перевірка	Умова у вузлі

На рис. 3.2 наведено приклад дерева рішень, завдання якого – відповісти на запитання: «Грати чи в гольф?» Щоб вирішити задачу, тобто прийняти рішення, варто віднести поточну ситуацію до одного з відомих класів (у цьому випадку – «грати» або «не грати»). Для цього потрібно відповісти на ряд питань, які містяться у вузлах цього дерева, починаючи з його кореня.

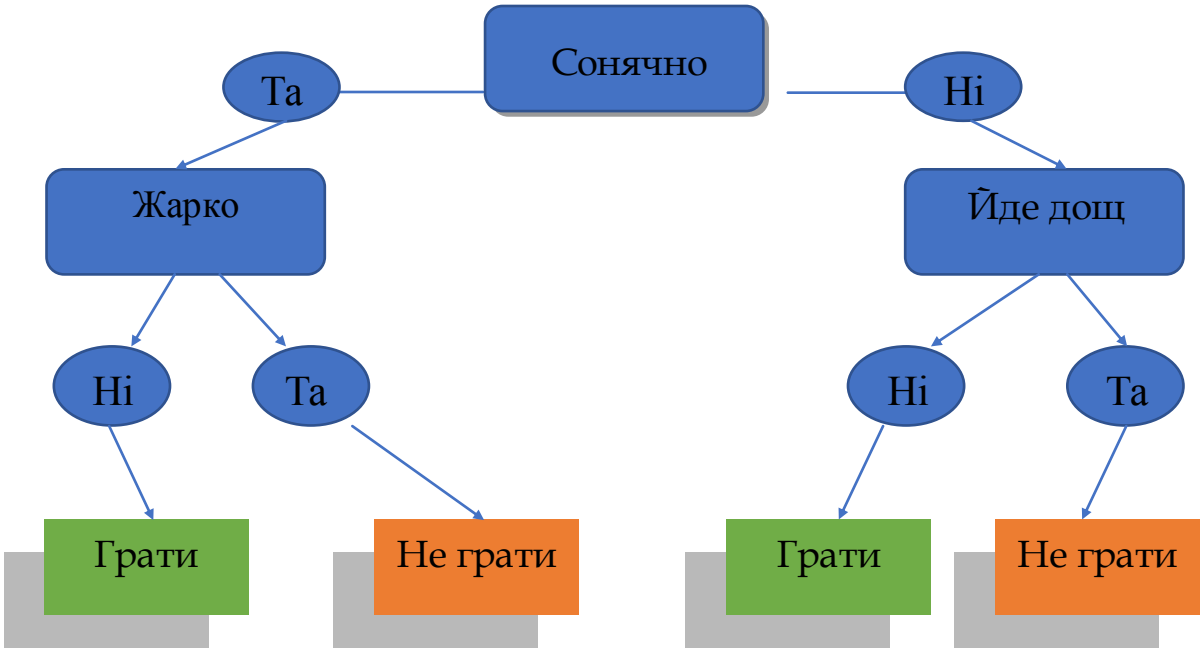


Рис. 3.2. Дерево рішень «Грати чи в гольф?»

Перший вузол дерева «Сонячно?» є вузлом перевірки, тобто умовою. З позитивної відповіді на запитання здійснюється перехід до лівої частини дерева, що називається лівою гілкою, з негативної – до правої частини дерева. Таким чином, внутрішній вузол дерева є вузлом перевірки певної умови. Далі йде наступне

запитання та ін., доки не буде досягнутий кінцевий вузол дерева, що є вузлом рішення. Для нашого дерева існує два типи кінцевого вузла: «Грати» і «Не грати» у гольф. У результаті проходження від кореня дерева до його вершини вирішується завдання класифікації, тобто обирається один з класів – «Грати» і «Не грати» у гольф.

Метою побудови дерева рішення в нашому випадку є визначення значення категоріальною залежною змінною. Отже, для нашої задачі основними елементами дерева рішень є:

1. Корінь дерева: «Сонячно?».
2. Внутрішній вузол дерева або вузол перевірки: «Температура повітря висока?», «Чи йде дощ?».
3. Лист, кінцевий вузол дерева, вузол рішення або вершина: «Грати», «Не граймо».
4. Гілка дерева (випадки відповіді): «Так», «Ні».

У розглянутому прикладі вирішується завдання бінарної класифікації, тобто створюється класифікаційна модель. Приклад демонструє роботу так званих бінарних дерев.

У вузлах бінарних дерев розгалуження може вестися тільки в двох напрямах, тобто існує можливість тільки двох відповідей на поставлене запитання («так» і «ні»).

Бінарні дерева є найпростішим окремим випадком дерев рішень. В інших випадках відповідей і, відповідно, гілок дерева, що виходять з його внутрішнього вузла, може бути більше двох.

Розглянемо більш складний приклад. База даних, на основі якої має здійснюватися прогнозування, містить такі ретроспективні дані про клієнтів банку, які є її атрибутами: вік, наявність нерухомості, освіта, середньомісячний дохід, повернення клієнтом вчасно кредиту. Завдання полягає в тому, щоб на підставі перерахованих вище даних (крім останнього атрибуту) визначити, чи варто видавати кредит новому клієнтові.

Таке завдання вирішується в два етапи: побудова класифікаційної моделі і її використання.

На етапі побудови моделі, власне, і будується дерево класифікації або створюється набір будь-яких правил. На етапі використання моделі побудоване дерево, або шлях від його кореня до однієї з вершин, що є набором правил для конкретного клієнта,

використовується для відповіді на поставлене запитання: «Чи видавати кредит?»

Правилом є логічна конструкція, надана у вигляді «якщо: то:».

На рис. 3.3 наведено приклад дерева класифікації, за допомогою якого вирішується задача «Видавати кредит клієнту?» Вона є типовою задачею класифікації, і за допомогою дерев рішень отримують досить хороші варіанти її рішення.

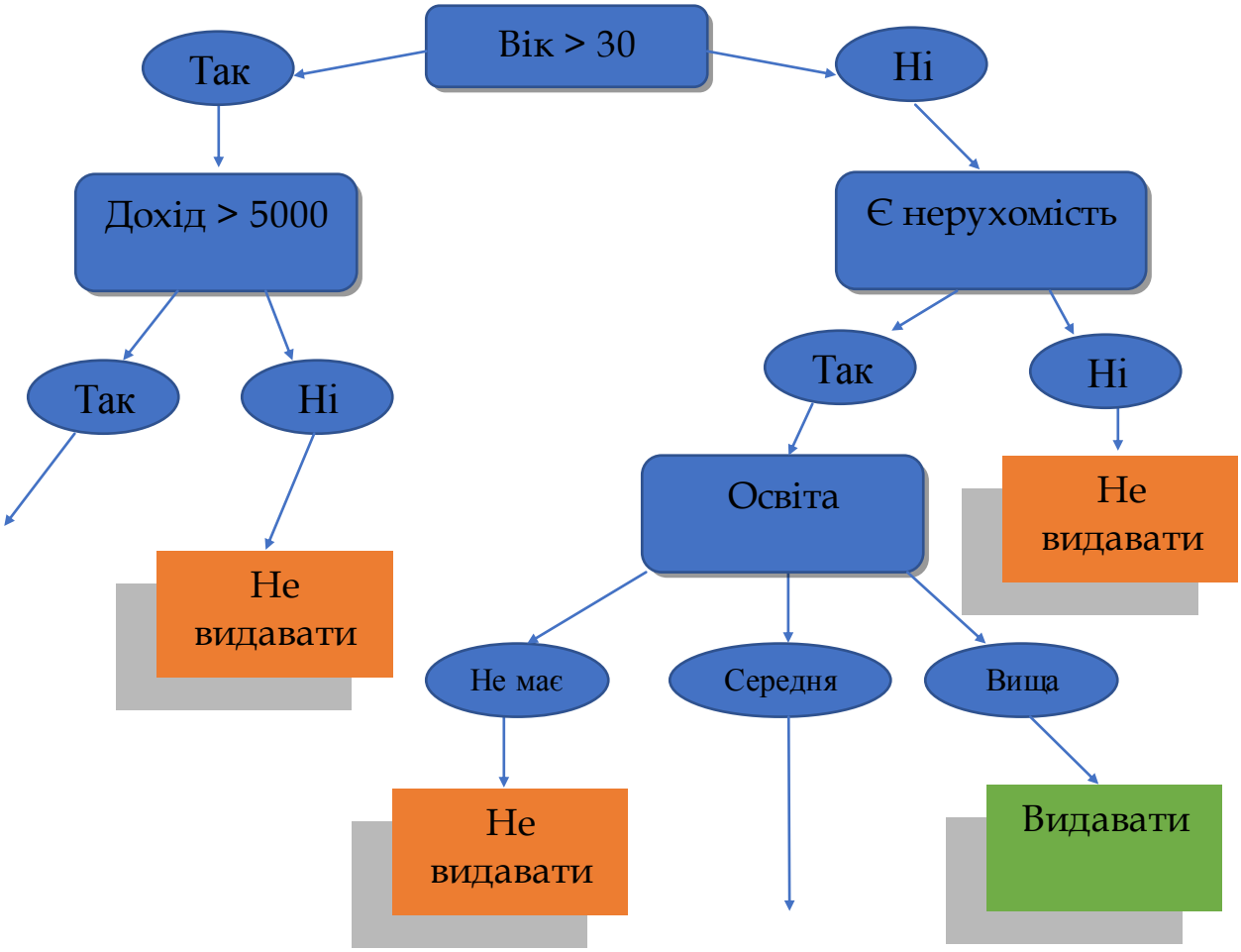


Рис. 3.3. Дерево класифікації

Як бачимо, внутрішні вузли дерева (вік, наявність нерухомості, дохід і освіта) є атрибутами описаної вище бази даних. Ці атрибути називають атрибутами розщеплення. Кінцеві вузли дерева, або листи, іменуються мітками класу, які є значеннями залежної категоріальної змінної «видавати» або «не видавати» кредит.

Кожна гілка дерева, що йде від внутрішнього вузла, зазначена предикатом розщеплення. Останній може ставитися лише до одного атрибуту розщеплення цього вузла. Характерна особливість предикатів розщеплення: кожен запис використовує унікальний шлях

від кореня дерева тільки до одного вузла-рішення. Об'єднана інформація про атрибути розщеплення і предиката розщеплення у вузлі називається критерієм розщеплення.

Для цієї задачі (як і для будь-якої іншої) може бути побудована множина дерев рішень будь-якої якості з різною прогнозуючою точністю.

Якість побудованого дерева рішення залежить від правильного вибору критерію розщеплення. Над розробкою і вдосконаленням критеріїв працюють багато дослідників.

Метод дерев рішень часто називають «наївним» підходом. Але завдяки ряду переваг цей метод є одним з найбільш популярних для вирішення задач класифікації.

Переваги дерев рішень.

1. Інтуїтивність дерев рішень. Класифікаційна модель, представлена у вигляді дерева рішень, є інтуїтивною і спрощує розуміння розв'язуваної задачі. Результат роботи алгоритмів конструювання дерев рішень на відміну, наприклад, від нейронних мереж, що являють собою «чорні ящики», легко інтерпретується користувачем. Ця властивість дерев рішень не тільки важлива під час віднесення до певного класу нового об'єкта, а й корисна у процесі інтерпретації моделі класифікації в цілому. Дерево рішень дозволяє зрозуміти і пояснити, чому конкретний об'єкт належить до певного класу.

2. Дерева рішень дають можливість отримувати правила з бази даних природною мовою.

3. Дерева рішень дозволяють створювати класифікаційні моделі в тих сферах, де аналітику досить складно формалізувати знання.

4. Алгоритм конструювання дерева рішень не вимагає від користувача вибору вхідних атрибутів (незалежних змінних). На вхід алгоритму можна подавати всі існуючі атрибути, алгоритм сам вибере найбільш значущі серед них, і тільки вони будуть використані для побудови дерева. Порівняно, наприклад, з нейронними мережами це значно полегшує користувачеві роботу, оскільки в нейронних мережах вибір кількості вхідних атрибутів істотно впливає на час навчання.

5. Точність моделей, створених за допомогою дерев рішень, порівняна з іншими методами побудови класифікаційних моделей (статистичні методи, нейронні мережі).

6. Розроблено ряд масштабованих алгоритмів, які можуть бути використані для побудови дерев рішення на надвеликих базах даних; масштабованість тут означає, що із зростанням кількості прикладів або записів бази даних час, що витрачається на навчання, тобто побудова дерев рішень, зростає лінійно.

7. Швидкий процес навчання. На побудову класифікаційних моделей за допомогою алгоритмів конструювання дерев рішень потрібно значно менше часу, ніж, наприклад, на навчання нейронних мереж.

8. Більшість алгоритмів конструювання дерев рішень мають можливість спеціальної обробки пропущених значень.

9. Багатокласичні статистичні методи, за допомогою яких вирішуються завдання класифікації, можуть працювати тільки з числовими даними, тоді як дерева рішень працюють і з числовими, і з категоріальним типами даних.

10. Багато статистичних методів є параметричними, і користувач повинен заздалегідь володіти певною інформацією, наприклад знати вид моделі, мати гіпотезу про вид залежності між змінними, припускати, який вид розподілу мають дані. Дерева рішень, на відміну від таких методів, будують непараметричні моделі. Таким чином, дерева рішень здатні вирішувати такі завдання Data Mining, в яких відсутня апріорна інформація про вид залежності між досліджуваними даними.

12.2. Загальні характеристики алгоритмів конструювання дерев рішень

Алгоритми конструювання дерев рішень складаються з етапів «побудова» або «створення» дерева і «скорочення» дерева. У ході створення дерева вирішуються питання вибору критерію розщеплення і зупинки навчання (якщо це передбачено алгоритмом). Під час етапу скорочення дерева вирішується питання відсікання деяких його гілок.

1. Критерій розщеплення. Процес створення дерева відбувається зверху вниз. У процесі алгоритм повинен знайти такий критерій розщеплення, іноді також званий критерієм розбиття, щоб розбити множину на підмножини, які б асоціювалися з цим вузлом перевірки. Кожен вузол перевірки повинен бути позначений певним атрибутом. Існує правило вибору атрибута: він повинен розбивати вихідну множину даних таким чином, щоб об'єкти підмножин, одержуваних унаслідок цього розбиття, були представниками одного класу або ж максимально наближені до такого розбиття. Останнє означає, що кількість об'єктів з інших класів, так званих «домішків», у кожному класі має прямувати до мінімуму.

Існують різні критерії розщеплення. Найбільш відомі – міра ентропії та індекс Gini.

У деяких методах для вибору атрибута розщеплення використовується так звана міра інформативності підпросторів атрибутів, яка ґрунтується на ентропійному підході і відома під назвою «міра інформаційного виграшу» або міра ентропії.

Інший критерій розщеплення реалізований в алгоритмі CART і називається індексом Gini. За допомогою цього індексу атрибут вибирається на підставі відстаней між розподілами класів.

Якщо дано множину T, що включає приклади з n-класів, то індекс Gini визначається за формулою:

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2,$$

де T – поточний вузол, p_j – імовірність класу j у вузлі T, n – кількість класів.

Зауважимо, що чим більше окремих випадків описано в дереві рішень, тим менша кількість об'єктів потрапляє в кожен окремий випадок. Такі дерева називають «гіллястими» або «кущистими», вони складаються з невиправдано значної кількості вузлів і гілок, вихідна множина розбивається на значну кількість підмножин, що складаються з досить малої кількості об'єктів. У результаті здатність «переповнення» таких дерев до узагальнення зменшується, і побудовані моделі не можуть давати правильні відповіді.

У процесі побудови дерева, щоб його розміри не стали надмірно великими, використовують спеціальні процедури, які дозволяють створювати оптимальні дерева, так звані дерева «відповідних розмірів».

Дерево має бути досить складним, щоб враховувати інформацію з досліджуваного набору даних, але водночас воно має бути досить простим на скільки це є можливим. Іншими словами, дерево повинно використовувати інформацію, що покращує якість моделі, й ігнорувати ту інформацію, яка її не покращує.

Тут існує дві можливі стратегії. Перша полягає в нарощуванні дерева до певного розміру відповідно до параметрів, заданих користувачем. Визначення цих параметрів може ґрунтуватися на досвіді та інтуїції аналітика, а також на деяких «діагностичних повідомленнях» системи, яка конструює дерево рішень.

Друга стратегія полягає у використанні набору процедур, що визначають «відповідний розмір» дерева.

Процедури, які використовують для запобігання створенню надмірно великих дерев, включають: скорочення дерева шляхом відсікання гілок; використання правил зупинки навчання.

Варто зазначити, що не всі алгоритми під час конструювання дерева працюють за однією схемою. Деякі алгоритми включають два окремих послідовних етапи: побудова дерева і його скорочення; інші чергують ці етапи в процесі своєї роботи для запобігання нарощування внутрішніх вузлів.

2. Зупинка побудови дерева. Правило зупинки повинно визначити, чи є цей вузол внутрішнім вузлом, при цьому він буде розбиватися далі, або ж він є кінцевим вузлом, тобто вузлом-рішенням.

Зупинка – такий момент у процесі побудови дерева, коли потрібно припинити подальші розгалуження.

Один з варіантів правил зупинки – «рання зупинка», вона визначає доцільність розбиття вузла. Перевага використання такого варіанта – зменшення часу на навчання моделі. Однак тут виникає ризик зниження точності класифікації. Тому рекомендується замість зупинки використовувати відсікання.

Другий варіант зупинки навчання – обмеження глибини дерева. У цьому випадку побудова закінчується, якщо досягнута задана глибина.

Ще один варіант зупинки – завдання мінімальної кількості прикладів, які будуть міститися в кінцевих вузлах дерева. За цим варіантом розгалуження тривають до того моменту, доки всі кінцеві вузли дерева не будуть чистими або міститимуть не більше ніж задану кількість об'єктів.

3. Скорочення дерева або відсікання гілок. Рішенням проблеми занадто гіллястого дерева є його скорочення шляхом відсікання деяких гілок.

Якість класифікаційної моделі, побудованої за допомогою дерева рішень, характеризується двома основними ознаками: точністю розпізнавання і помилкою.

Точність розпізнавання розраховується як відношення об'єктів, правильно класифікованих у процесі навчання, до загальної кількості об'єктів набору даних, які брали участь у навчанні.

Помилка розраховується як відношення об'єктів, неправильно класифікованих у процесі навчання, до загальної кількості об'єктів набору даних, які брали участь у навчанні.

Відсікання гілок або заміну деяких гілок піддерева варто проводити там, де ця процедура не приводить до зростання помилки. Процес проходить знизу доверху, тобто є висхідним. Це більш популярна процедура, ніж використання правил зупинки. Дерева, отримані після відсікання деяких гілок, називають усіченими.

Якщо таке дерево все ще не є інтуїтивним і складне для розуміння, то використовують вилучення правил, які об'єднують у набори для опису класів. Кожен шлях від кореня дерева до його вершини або листа дає одне правило. Умовами правила є перевірки на внутрішніх вузлах дерева.

12.3. Алгоритми побудови дерев рішень

На сьогодні існує велика кількість алгоритмів, за допомогою яких будують дерева рішень: CART, C4.5, CHAID, CN2, NewId, ITrule та інші.

1. Алгоритм CART (Classification and Regression Tree), як видно з назви, вирішує завдання класифікації і регресії. Атрибути набору даних можуть мати як дискретне, так і числове значення.

Алгоритм CART призначений для побудови бінарного дерева рішень.

Інші особливості алгоритму CART:

– функція оцінки якості розбиття;

– механізм відсікання дерева;

– алгоритм обробки пропущених значень;

– побудова дерев регресії.

Кожен вузол бінарного дерева під час розбиття має тільки двох нащадків, які називаються дочірніми гілками. Подальший поділ гілки залежить від того, чи багато вихідних даних описується. На кожному кроці побудови дерева правило формується у вузлі, ділить задану множину прикладів на дві частини. Права його частина – це та частина множини, в якій правило виконується; ліва – та, для якої правило не виконується.

Функція оцінки якості розбиття, яка використовується для вибору оптимального правила, – індекс Gini, був описаний вище. Зазначимо, що ця оціночна функція заснована на ідеї зменшення невизначеності у вузлі. Припустимо, є вузол, і він розбитий на два класи. Максимальна невизначеність у вузлі буде досягнута у разі розбиття його на дві підмножини по 50 прикладів, а максимальна визначеність – під час розбиття на 100 і 0 прикладів.

Правила розбиття. Нагадаємо, що алгоритм CART працює з числовими і категоріальним атрибутами. У кожному вузлі розбиття може йти тільки за одним атрибутом. Якщо атрибут є числовим, то у внутрішньому вузлі формується правило виду $x_i <= c$. Значення c у більшості випадків обирається як середнє арифметичне двох сусідніх упорядкованих значень змінної x_i навчального набору даних. Якщо ж атрибут належить до категоріального

типу, то у внутрішньому вузлі формується правило x_i $V(x_i)$, де $V(x_i)$ – деяка непорожня підмножина множини значень змінної x_i у навчальному наборі даних.

Механізм відсікання. Цим механізмом алгоритм CART принципово відрізняється від інших алгоритмів конструювання дерев рішень. У розглянутому алгоритмі відсікання це – певний компроміс між отриманням дерева «відповідного розміру» й отриманням найбільш точної оцінки класифікації. Метод полягає в отриманні послідовності дерев, що зменшуються, але дерева розглядаються не повним обсягом, а лише «кращі представники».

Перехресна перевірка є найбільш складною та одночасно оригінальною частиною алгоритму CART. Вона являє собою шлях вибору остаточного дерева за умови, що набір даних має невеликий об'єм або ж записи набору даних настільки специфічні, що розділити набір на навчальну і тестову вибірки не є можливим.

2. Алгоритм C4.5 будує дерево рішень з необмеженою кількістю гілок у вузлі. Цей алгоритм може працювати тільки з дискретним залежним атрибутом і тому може вирішувати тільки задачі класифікації. C4.5 вважається одним з найвідоміших і широко використовуваних алгоритмів побудови дерев класифікації (рішень).

Для роботи алгоритму C4.5 варто дотримуватися таких вимог:

– кожен запис набору даних має бути асоційований з одним з визначених класів, тобто один з атрибутів набору даних повинен бути міткою класу;

– класи мають бути дискретними. Кожен приклад повинен однозначно ставитися до одного з класів;

– кількість класів має бути значно менша кількості записів у досліджуваному наборі даних.

Алгоритм C4.5 повільно працює на надвеликих і зашумлених множинах даних.

3. Одними з найбільш популярних алгоритмів побудови дерев рішень є алгоритм ID3 (C4.5 – його модифікація).

Алгоритм ID3 починає роботу з усіма навчальними прикладами в кореневому вузлі дерева. Для поділу множини прикладів кореневого вузла вибирається один з атрибутів, і для кожного

значення, прийнятого цим атрибутом, будується гілка і створюється дочірній вузол. Потім усі приклади розподіляються за дочірніми вузлами відповідно до значення атрибута.

Алгоритми побудови дерев рішень відрізняються такими характеристиками:

– вид розщеплення – бінарне, множинне;

– критерії розщеплення – ентропія, Gini та інші;

– можливість обробки пропущених значень;

– процедура скорочення гілок або відсікання;

– можливості вилучення правил з дерев.

Жоден алгоритм побудови дерева не можна апріорі вважати найкращим або досконалим, підтвердження доцільності використання конкретного алгоритму має бути перевірено і підтверджено експериментом.

Якість роботи методу дерев рішень залежить як від вибору алгоритму, так і від набору досліджуваних даних. Незважаючи на всі переваги цього методу, варто пам'ятати, що для того щоб побудувати якісну модель, необхідно розуміти природу взаємозв'язку між залежними і незалежними змінними і підготувати достатній набір даних.

Термін «дерево рішень» використовується не тільки в Data Mining, а й у суміжних сферах аналізу й обробки даних, штучний інтелект. Наприклад, у завданнях алгоритмізації дерево рішень – це спосіб надання процесу прийняття рішення, що має вигляд відповідей на серію питань, що утворюють деревоподібну структуру, кожна кінцева вершина якої являє собою елементарне рішення. У менеджменті під деревом рішень розуміють графічне зображення альтернативних дій і їхніх наслідків. Однак між усіма цими тлумаченнями терміна багато спільного.

Запитання до параграфа 12

1. Яка сутність методу дерев рішень?
2. Що таке ієрархічна структура?
3. Які поняття зустрічаються в теорії дерев рішень?
4. Що таке вузол?
5. Що таке лист у дереві рішень?

6. Що таке атрибут?

7. Що таке бінарні дерева?

8. Що називають правилом?

9. Які переваги дерев рішень?

10. Які загальні характеристики алгоритмів конструювання дерев рішень?

11. Коли відбувається зупинка побудови дерева рішень?

12. Що таке скорочення дерева або відсікання гілок?

13. Які існують алгоритми побудови дерев рішень?

14. Які особливості алгоритму CART?

15. Яких для роботи алгоритму C4.5 варто дотримуватися вимог?

16. Яка сутність алгоритму ID3?

§ 13. Методи класифікації і прогнозування. Метод опорних векторів. Метод «найближчого сусіда». Байєсівська класифікація

13.1. Метод опорних векторів

Метод опорних векторів (Support Vector Machine – SVM) належить до групи граничних методів. Він визначає класи за допомогою кордонів сфер.

За допомогою цього методу вирішуються завдання бінарної класифікації.

В основі методу полягає поняття площин рішень. Площина рішення розділяє об'єкти з різною класовою приналежністю.

На рис. 3.4 наведено приклад, в якому беруть участь об'єкти двох типів. Лінія задає кордон, праворуч від якої – всі об'єкти типу «круг», а ліворуч – типу «ромб». Новий об'єкт, що потрапляє праворуч, класифікується як об'єкт класу «ромб» або як об'єкт класу «круг», якщо він розташувався ліворуч від прямої. У цьому випадку кожен об'єкт характеризується двома вимірами.

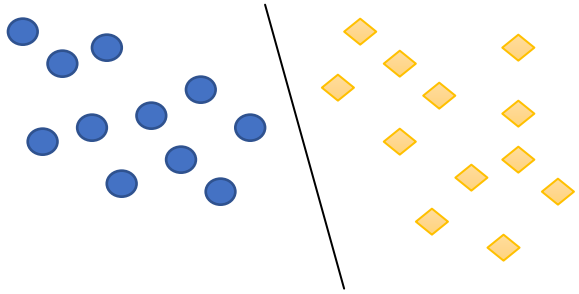


Рис. 3.4. Поділ класів прямою лінією

Мета методу опорних векторів – знайти площину (лінію), що розділяє дві множини об'єктів; така площина показана на рис. 3.5. На цьому рисунку множина зразків поділена на два класи: одні належать класу A, інші – класу B.

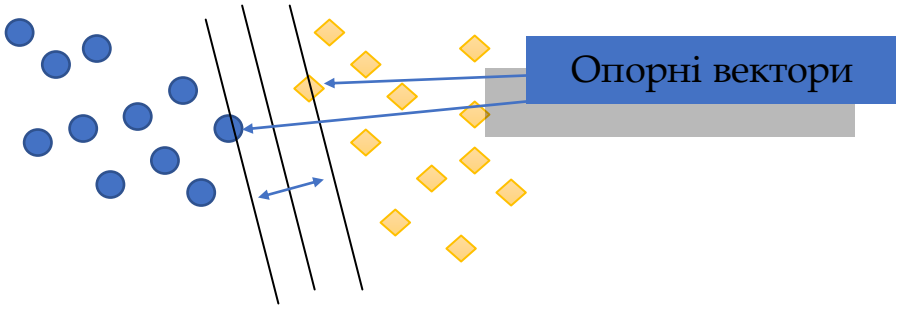


Рис. 3.5. До визначення опорних векторів

Метод шукає зразки, що містяться на кордонах між двома класами, тобто опорні вектори.

Опорними векторами називаються об'єкти множини, що перебувають на кордонах сфер. Класифікація вважається хорошою, якщо сфера між кордонами порожня.

Лінійний SVM. Рішення завдання бінарної класифікації за допомогою методу опорних векторів полягає в пошуку деякої лінійної функції, яка правильно поділяє набір даних на два класи.

Розглянемо задачу класифікації, де кількість класів дорівнює двом. Завдання можна сформулювати як пошук функції $f(x)$, що приймає значення менше нуля для векторів одного класу і більше нуля – для векторів іншого класу. Як вихідні дані для вирішення поставленого завдання, тобто пошуку функції $f(x)$, дано навчальний набір векторів простору, для яких відома їхня належність до одного з класів. Сімейство класифікаційних функцій можна описати через функцію $f(x)$. Гіперплощина визначена вектором a й значенням $в$, тобто $f(x) = ax + в$. Рішення цієї задачі проілюстровано на рис. 3.6.

У результаті рішення задачі, тобто побудови SVM-моделі, знайдена функція, що приймає значення менше нуля для векторів одного класу і більше нуля – для векторів іншого класу. Для кожного нового об'єкта негативне або позитивне значення визначає належність об'єкта до одного з класів.

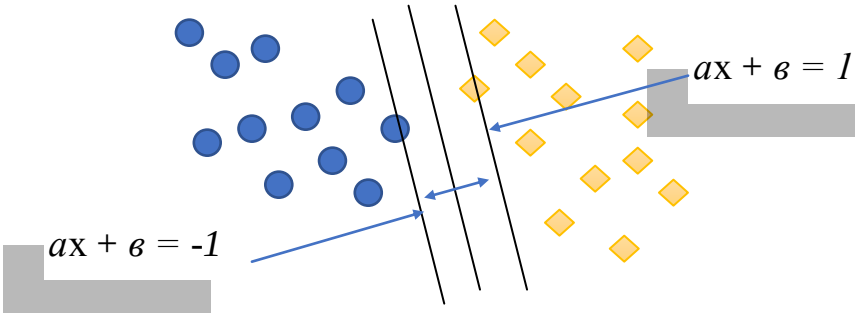


Рис. 3.6. Лінійний SVM

Найкращою функцією класифікації є функція, для якої очікуваний ризик мінімальний. Поняття очікуваного ризику в цьому випадку означає очікуваний рівень помилки класифікації.

Безпосередньо оцінити очікуваний рівень помилки побудованої моделі неможливо, це можна зробити за допомогою поняття емпіричного ризику. Однак варто враховувати, що мінімізація останнього не завжди приводить до мінімізації очікуваного ризику. Цю обставину потрібно пам'ятати під час роботи з відносно невеликими наборами тренувальних даних.

Емпіричний ризик – рівень помилки класифікації на навчальній виборці.

Таким чином, у результаті рішення задачі методом опорних векторів для лінійно подільних даних отримуємо функцію класифікації, яка мінімізує верхню оцінку очікуваного ризику.

Однією з проблем, пов'язаних з вирішенням завдань класифікації цим методом, є та обставина, що не завжди можна легко знайти лінійний кордон між двома класами.

У таких випадках один з варіантів – збільшення розмірності, тобто перенесення даних з площини у тривимірний простір, де можливо побудувати таку площину, яка ідеально розділить множини зразків на два класи. Опорними векторами в цьому випадку слугуватимуть об'єкти з обох класів, які є екстремальними.

Таким чином, за допомогою додавання так званого оператора ядра і додаткових розмірностей перебувають кордони між класами у вигляді гіперплощин.

Проте варто пам'ятати: складність побудови SVM-моделі полягає в тому, що чим вище розмірність простору, тим складніше з ним працювати. Один з варіантів роботи з даними високої розмірності – це попереднє застосування будь-якого методу зниження розмірності даних для виявлення найбільш істотних компонент, а потім використати метод опорних векторів.

Як і будь-який інший метод, метод SVM має свої сильні і слабкі сторони, які потрібно враховувати під час вибору цього методу.

Недолік методу полягає в тому, що для класифікації використовується не вся множина елементів, а лише їхня невелика частина, яка міститься на кордонах.

Перевага методу полягає в тому, що для класифікації методом опорних векторів, на відміну від більшості інших методів, достатньо невеликого набору даних. У процесі правильної роботи моделі, побудованої на навчальній множині, цілком можливе застосування цього методу на реальних даних.

Метод опорних векторів дозволяє:

– отримати функцію класифікації з мінімальною верхньою оцінкою очікуваного ризику (рівня помилки класифікації);

– використовувати лінійний класифікатор для роботи з нелінійно подільними даними, поєднуючи простоту з ефективністю.

13.2. Метод «найближчого сусіда»

Метод «найближчого сусіда» належить до класу методів, робота яких ґрунтується на зберіганні даних у пам'яті для порівняння з новими елементами. Під час появи нового запису для прогнозування перебувають відхилення між цим записом і подібними наборами даних, і найбільш подібний (де найменше відхилення, ближній сусід) ідентифікується. Наприклад, у разі розгляду нового клієнта банку його атрибути порівнюються з усіма існуючими клієнтами цього банку (дохід, вік тощо). Множина «найближчих сусідів» потенційного клієнта банку обирається на підставі найближчого значення доходу, віку та ін.

У разі такого підходу використовується термін «k-найближчий сусід». Термін означає, що вибирається *k* «верхніх» (найближчих) сусідів для їхнього розгляду як множина «найближчих сусідів». Оскільки не завжди зручно зберігати всі дані, іноді зберігається тільки множина «типових» випадків. У такому випадку цей метод називають міркуванням за аналогією (Case Based Reasoning, CBR), міркуванням на основі аналогічних випадків, міркуванням за прецедентами.

Прецедент – це опис ситуації в поєднанні з детальним зазначенням дій, що вживаються в цій ситуації.

Підхід, заснований на прецедентах, умовно можна поділити на такі етапи:

– збір докладної інформації про поставлену задачу;

– зіставлення цієї інформації з деталями прецедентів, що зберігаються в базі, для виявлення аналогічних випадків;

– вибір прецеденту найбільш близького до поточної проблеми з бази прецедентів;

– адаптація обраного рішення до поточної проблеми, якщо це необхідно;

– перевірка коректності кожного знову отриманого рішення;

– занесення детальної інформації про новий прецедент у базу прецедентів.

Отже, висновок, заснований на прецедентах, являє собою такий метод аналізу даних, який робить висновки щодо цієї ситуації за результатами пошуку аналогій, що зберігаються в базі прецедентів.

Цей метод за своєю сутністю належить до категорії «навчання без учителя», тобто є «самонавчальною» технологією, завдяки чому робочі характеристики кожної бази прецедентів з плином часу і накопиченням прикладів поліпшуються. Розробка баз прецедентів за конкретною предметною сферою відбувається природною для людини мовою, отже, може бути виконана найбільш досвідченими співробітниками компанії – експертами або аналітиками, які працюють у цій галузі.

Однак це не означає, що CBR-системи самостійно можуть приймати рішення. Останнє завжди залишається за людиною, цей метод лише пропонує можливі варіанти рішення і вказує на більш «розумний» з її точки зору.

Переваги методу:

– простота використання отриманих результатів;

– рішення неунікальні для конкретної ситуації, можливе їхнє використання для інших випадків;

– метою пошуку є не гарантовано правильне рішення, а найкраще з можливих.

Недоліки методу «найближчого сусіда»:

– цей метод не створює будь-яких моделей або правил, просто узагальнюють попередній досвід: у виборі рішення вони ґрунтуються на всьому масиві доступних історичних даних, тому неможливо зазначити, на якій підставі будуються відповіді;

– існує складність вибору критерію «близькості» (метрики). Від цього критерію основним чином залежить обсяг множини записів, які потрібно зберігати в пам'яті для досягнення задовільної класифікації або прогнозу. Також існує висока залежність результатів класифікації від обраної метрики;

– під час використання методу виникає необхідність повного перебору навчальної вибірки у процесі розпізнавання, наслідок цього – обчислювальна трудомісткість;

– типові завдання цього методу – це завдання невеликої розмірності за кількістю класів і змінних.

За допомогою цього методу вирішуються завдання класифікації і регресії.

Рішення завдання класифікації нових об'єктів. Це завдання схематично зображено на рис. 3.7. Приклади (відомі екземпляри) відмічені знаком «+» або «–», що визначає приналежність до відповідного класу («+» або «–»), а новий об'єкт, який потрібно класифікувати, позначений маленьким кружечком. Нові об'єкти також називають точками запиту.

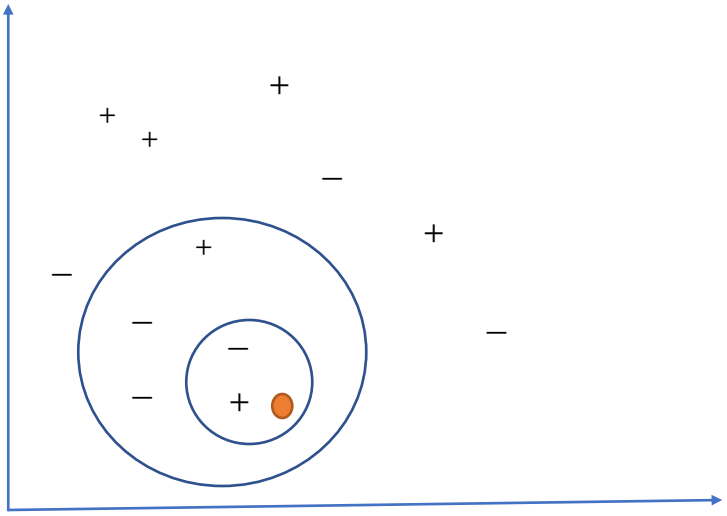


Рис. 3.7. Класифікація об'єктів множини за різним значенням параметра *k*

Наша мета полягає в оцінці (класифікації) відгуку точок запиту з використанням спеціально обраної кількості їхніх найближчих сусідів. Іншими словами, хочемо дізнатися, до якого класу варто віднести точку запиту – як знак «+» або як знак «–».

Для початку розглянемо результат роботи методу *к*-найближчих сусідів з використанням одного найближчого сусіда. У цьому випадку відгук точки запиту буде класифікований як знак плюс, оскільки найближча сусідня точка має знак плюс.

Тепер збільшимо кількість використовуваних найближчих сусідів до двох. Цього разу метод не зможе класифікувати відгук точки запиту, оскільки друга найближча точка має знак мінус і обидва знаки рівноцінні.

Далі збільшимо кількість використовуваних найближчих сусідів до 5. Таким чином, буде визначена околиця точки запиту. Оскільки у сфері утримується 2 точки зі знаком «+» і 3 точки зі знаком «–», алгоритм *к*-найближчих сусідів присвоїть знак «–» відгуку точки запиту.

Незалежні і залежні змінні набору даних можуть бути як безперервними, так і категоріальним. Для безперервних залежних змінних завдання розглядається як задача прогнозування, для дискретних змінних – як задача класифікації.

Передбачення в задачі прогнозування здійснюється шляхом усереднення виходів *k*-найближчих сусідів, а рішення задачі класифікації засноване на принципі «за більшістю голосів».

Критичним моментом у використанні методу *k*-найближчих сусідів є вибір параметра *k*. Він один з найбільш важливих факторів, що визначають якість прогнозної або класифікаційної моделі.

Якщо вибрано досить маленьке значення параметра *k*, виникає ймовірність великого розкиду значень прогнозу. Якщо вибране значення занадто велике, це може призвести до значного зміщення моделі. Таким чином, повинно бути вибрано оптимальне значення параметра *k*. Тобто це значення має бути настільки великим, щоб звести до мінімуму ймовірність неправильної класифікації, і одночасно досить малим, щоб *k* сусідів були розташовані досить близько до точки запиту.

Таким чином, *k* розглядають як параметр, для якого повинен бути знайдений компроміс між величиною розмаху (розкиду) моделі і її зміщенням.

Один з варіантів оцінки параметра *k*-проведення крос-перевірки. Крос-перевірка – відомий метод отримання оцінок невідомих параметрів моделі. Основна ідея методу – поділ вибірки даних на *v* «складок». *V* «складки» – випадковим чином виділені ізольовані підвибірки.

З фіксованим значеням *k* будується модель *k*-найближчих сусідів для отримання прогнозів на *v*-му сегменті (інші сегменти при цьому використовуються як приклади) й оцінюється помилка класифікації. Для регресійних задач найбільш часто як оцінка помилки виступає сума квадратів, а для класифікаційних задач зручніше розглядати точність (відсоток коректно класифікованих спостережень).

Далі процес послідовно повторюється для всіх можливих варіантів вибору *v*. Після вичерпання *v* «складок» (циклів) обчислені помилки усереднюються і використовуються як міра стійкості моделі. Вищеописані дії повторюються для різних *k*, і значення, яке відповідає найменшій помилці, приймається як оптимальне (оптимальне в сенсі методу крос-перевірки).

Варто враховувати, що крос-перевірка – процедура, яка вимагає значних обчислень, і необхідно надати час для роботи алгоритму, особливо якщо обсяг вибірки досить великий.

Другий варіант вибору значення параметра *k*-самостійно задати його значення. Однак цей спосіб потрібно використовувати, якщо є обґрунтовані припущення щодо можливого значення параметра, наприклад попередні дослідження подібних наборів даних.

Метод *k*-найближчих сусідів показує досить непогані результати в найрізноманітніших задачах.

Прикладом реального використання описаного вище методу є програмне забезпечення центру технічної підтримки компанії Dell, розроблене компанією Inference. Ця система допомагає співробітникам центру відповідати на більшу кількість запитів, одразу пропонуючи відповіді на поширені питання, та дозволяє отримувати доступ до бази під час розмови по телефону з користувачем. Співробітники центру технічної підтримки, завдяки реалізації цього методу, можуть відповідати одночасно на значну кількість дзвінків. Програмне забезпечення CBR нині розгорнуто в мережі Intranet компанії Dell.

13.3. Байєсівська класифікація

Спочатку байєсівську класифікацію використовували для формалізації знань експертів в експертних системах, нині байєсівська класифікація також застосовується як один з методів Data Mining.

Так звана наївна класифікація, або наївно-байєсівський підхід, є найбільш простим варіантом методу, що використовує байєсівські мережі. При цьому підході вирішуються завдання класифікації, результатом роботи методу є так звані «прозорі» моделі.

«Наївна» класифікація – досить прозорий і зрозумілий метод класифікації. «Наївною» вона називається тому, що виходить з припущення про взаємну незалежність ознак.

Властивості наївної класифікації:

– використання всіх змінних і визначення всіх залежностей між ними;

– наявність двох припущень щодо змінних: усі змінні є однаково важливими; усі змінні є статистично незалежними, тобто значення однієї змінної нічого не свідчить про значення іншої.

Більшість інших методів класифікації припускають, що перед початком класифікації ймовірність того, що об'єкт належить певному класу, однакова, але це не завжди правильно.

Припустимо, відомо, що певний відсоток даних належить конкретному класу. Виникає питання, чи можемо використовувати цю інформацію під час побудови моделі класифікації? Існує множина реальних прикладів використання цих апріорних знань, які допомагають класифікувати об'єкти. Типовий приклад з медичної практики. Якщо доктор відправляє результати аналізів пацієнта на додаткове дослідження, він відносить пацієнта до якогось певного класу. Яким чином можна застосувати цю інформацію? Можемо використовувати її як додаткові дані у процесі побудови класифікаційної моделі.

Зазначають такі переваги байєсівських мереж як метод Data Mining:

– у моделі визначаються залежності між усіма змінними, це дозволяє легко обробляти ситуації, в яких значення деяких змінних невідомі;

– байєсівські мережі досить просто інтерпретуються і дозволяють на етапі прогностичного моделювання легко проводити аналіз за сценарієм «що, якщо»;

– байєсівський метод дозволяє природним чином поєднувати закономірності, виведені з даних, і, наприклад, експертні знання, отримані в явному вигляді;

– використання байєсівських мереж дозволяє уникнути проблеми перенавчання, тобто надмірного ускладнення моделі, що є слабкою стороною багатьох методів (наприклад, дерев рішень і нейронних мереж).

Наївно-байєсівський підхід має такі недоліки:

– множити умовні ймовірності коректно тільки тоді, коли всі вхідні змінні дійсно статистично незалежні; хоча часто цей метод показує досить хороші результати у разі недотримання умови статистичної незалежності, але теоретично така ситуація повинна оброблятися більш складними методами, заснованими на навчанні байєсівських мереж;

– неможлива безпосередня обробка безперервних змінних, потрібне їхнє перетворення, щоб атрибути були дискретними; однак такі перетворення іноді можуть призводити до втрати значущих закономірностей;

– на результат класифікації впливають тільки індивідуальні значення вхідних змінних, комбінований вплив пар або трійок значень різних атрибутів тут не враховується. Це могло б поліпшити якість класифікаційної моделі з точки зору її прогнозуючої точності, однак збільшило б кількість перевірених варіантів.

Запитання до параграфа 13

1. Яка сутність методу опорних векторів?
2. Яка мета методу опорних векторів?
3. Яка сутність методу «найближчого сусіда»?
4. Яка сутність підходу, заснованого на прецедентах?
5. Які переваги методу «найближчого сусіда»?
6. Які недоліки методу «найближчого сусіда»?
7. У чому сутність методу *k*-найближчих сусідів?
8. Що таке байєсівська класифікація?
9. Які властивості наївної класифікації?
10. Які зазначають переваги байєсівських мереж як методу Data Mining?

§ 14. Нейронні мережі

14.1. Загальна характеристика

Ідея нейронних мереж народилася в рамках теорії штучного інтелекту, в результаті спроб імітувати здатність біологічних нервових систем навчатися і виправляти помилки.

Нейронні мережі – це моделі біологічних нейронних мереж мозку, в яких нейрони імітуються відносно простими, часто однотипними елементами (штучними нейронами).

Нейронна мережа може бути представлена спрямованим графом із зваженими зв’язками, в якому штучні нейрони є вершинами, а синаптичні зв’язки – дугами. Нейронні мережі широко використовуються для вирішення різноманітних завдань. Серед

сфер застосування нейронних мереж – автоматизація процесів розпізнавання образів, прогнозування, адаптивне управління, створення експертних систем, організація асоціативної пам'яті, обробка аналогових і цифрових сигналів, синтез та ідентифікація електронних ланцюгів і систем. За допомогою нейронних мереж можна, наприклад, передбачати обсяги продажів виробів, показники біржового ринку, виконувати розпізнавання сигналів. Моделі нейронних мереж можуть бути програмного й апаратного виконання.

Якщо зазначити простою мовою, шарувата нейронна мережа являє собою сукупність нейронів, які становлять шари. У кожному шарі нейрони між собою ніяк не пов'язані, але пов'язані з нейронами попереднього і наступного шарів. Інформація надходить з першого на другий шар, з другого – на третій тощо.

Серед завдань Data Mining, що вирішуються за допомогою нейронних мереж, виділяють такі:

– класифікація (навчання з учителем). Приклади завдань класифікації: розпізнавання тексту, розпізнавання мови, ідентифікація особистості тощо;

– прогнозування. Для нейронної мережі завдання прогнозування може бути поставлене таким чином: знайти найкраще наближення функції, заданої кінцевим набором вхідних значень (навчальних прикладів). Наприклад, нейронні мережі дозволяють вирішувати завдання відновлення пропущених значень;

– кластеризація (навчання без учителя). Прикладом завдання кластеризації може бути задача стиснення інформації шляхом зменшення розмірності даних.

14.2. Елементи нейронних мереж

Штучний нейрон (формальний нейрон) – елемент штучних нейронних мереж, що моделює деякі функції біологічного нейрона.

Основна функція штучного нейрона – формувати вихідний сигнал залежно від сигналів, що надходять на його входи.

У найпоширенішій конфігурації вхідні сигнали обробляються адаптивним суматором, потім вихідний сигнал суматора надходить у нелінійний перетворювач, де перетворюється функцією активації, і результат подається на вихід (у точку розгалуження).

Загальний вигляд штучного нейрона наведено на рис. 3.8.

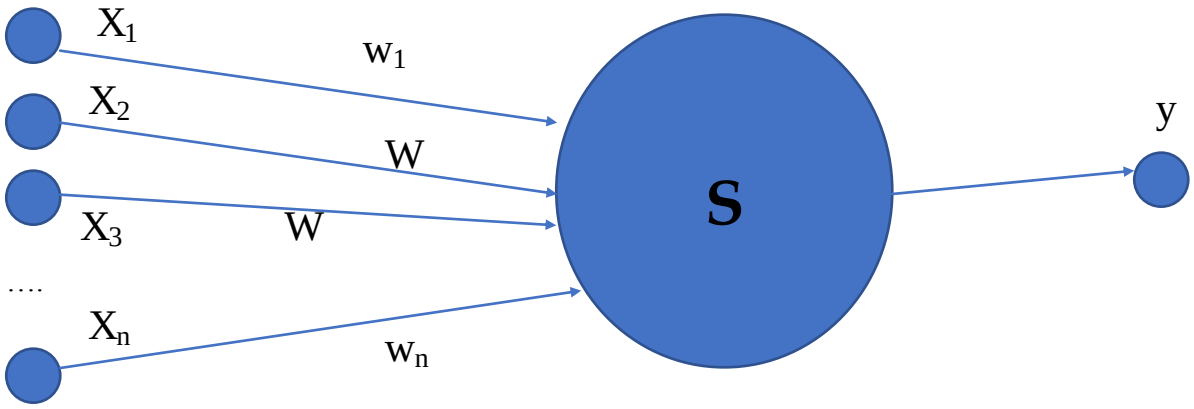


Рис. 3.8. Вигляд штучного нейрона

Нейрон характеризується поточним станом і володіє групою синапсів – односпрямованих вхідних зв'язків, з'єднаних з виходами інших нейронів. Нейрон має аксон – вихідний зв'язок цього нейрона, з якої сигнал (збудження або гальмування) надходить на синапси наступних нейронів. Кожен синапс характеризується величиною синаптичного зв'язку (її вагою w_i). Поточний стан нейрона визначається як зважена сума його входів:

$$S = \sum_{i=1}^{n} x_i \cdot w_i .$$

Вихід нейрона є функцією його стану: y = *f(s)*.

Активаційна функція, яку також називають характеристичної функцією, – це нелінійна функція, що обчислює вихідний сигнал формального нейрона.

Часто використовувані активаційні функції:

1. Жорстка порогова функція.
2. Лінійний поріг.
3. Сигмоїдальна функція.

Вибір активаційної функції визначається специфікою поставленого завдання або обмеженнями, що накладаються деякими алгоритмами навчання.

Нелінійний перетворювач – це елемент штучного нейрона, що перетворює поточний стан нейрона (вихідний сигнал адаптивного суматора) у вихідний сигнал нейрона за деяким нелінійним законом.

Точка розгалуження (вихід) – це елемент формального нейрона, який посилає його вихідний сигнал за кількома адресами та має один вхід і декілька виходів.

На вхід точки розгалуження зазвичай подається вихідний сигнал нелінійного перетворювача, який потім надсилається на входи інших нейронів.

14.3. Архітектура нейронних мереж

Нейронні мережі можуть бути синхронні й асинхронні. У синхронних нейронних мережах кожного моменту часу свій стан змінює лише один нейрон. В асинхронних – стан змінюється одразу у групи нейронів, як правило, у всього шару.

Можна виділити дві базові архітектури – шаруваті та повнозв'язні мережі. Ключовим у шаруватих мережах є поняття шару.

Шар – один або декілька нейронів, на входи яких подається один і той самий загальний сигнал.

Шаруваті нейронні мережі – нейронні мережі, в яких нейрони розбиті на окремі групи (верстви) так, що обробка інформації здійснюється пошарово.

У шаруватих мережах нейрони *i*-го шару отримують вхідні сигнали, перетворюють їх і через точки розгалуження передають нейронам (*i* + *1*) шару. І так до *k*-го шару, який видає вихідні сигнали для інтерпретатора і користувача. Кількість нейронів у кожному шарі, не пов'язана з кількістю нейронів в інших шарах, може бути довільною.

У рамках одного шару дані обробляються паралельно, а в масштабах усієї мережі обробка ведеться послідовно – від шару до шару. До шаруватих нейронних мереж належать, наприклад, багатошарові персептрони, мережі радіальних базисних функцій, когнітрон, некогнітрон, мережі асоціативної пам'яті.

Однак сигнал не завжди подається на всі нейрони прошарку. У когнітроні, наприклад, кожен нейрон поточного шару отримує сигнали тільки від близьких йому нейронів попереднього шару.

Шаруваті мережі, в свою чергу, можуть бути одношаровими і багатошаровими.

Одношарова мережа – мережа, що складається з одного шару. Багатошарова мережа – мережа, що має декілька шарів.

У багатошаровій мережі перший шар називається вхідним, наступні – внутрішніми або прихованими, останній шар – вихідним. Таким чином, проміжні шари – це всі верстви багатошарової нейронної мережі, крім вхідного і вихідного. Вхідний шар мережі реалізує зв'язок з вхідними даними, вихідний – з вихідними. Таким чином, нейрони можуть бути вхідними, вихідними і прихованими.

Вхідний шар організований з вхідних нейронів, які отримують дані і поширюють їх на входи нейронів прихованого шару мережі.

Прихований нейрон – це нейрон, що міститься в прихованому шарі нейронної мережі.

Вихідні нейрони, з яких організовано вихідний шар мережі, видає результати роботи нейронної мережі.

У повнозв'язних мережах кожен нейрон передає свій вихідний сигнал іншим нейронам, включаючи самого себе. Вихідними сигналами мережі можуть бути всі або деякі вихідні сигнали нейронів після кількох тактів функціонування мережі. Усі вхідні сигнали подаються всім нейронам.

Перед використанням нейронної мережі її необхідно навчити. Процес навчання нейронної мережі полягає в підстроюванні її внутрішніх параметрів під конкретну задачу. Алгоритм роботи нейронної мережі є ітеративним, його кроки називають епохами або циклами.

Епоха – одна ітерація в процесі навчання, що включає надання всіх прикладів з навчальної множини і, можливо, перевірку якості навчання на контрольній множині. Процес навчання здійснюється на навчальній вибірці.

Навчальна вибірка включає вхідні значення і відповідні їм вихідні значення набору даних. У ході навчання нейронна мережа знаходить якісь залежності вихідних полів від вхідних.

Таким чином, перед нами постає питання – які вхідні поля (ознаки) нам необхідно використовувати. Спочатку вибір здійснюється евристичний, далі кількість входів може бути змінено.

Складність може викликати питання про кількість спостережень у наборі даних. Існують правила, що описують зв'язок між необхідною кількістю спостережень і розміром мережі, проте їхня правильність не доведена. Кількість необхідних спостережень залежить від складності розв'язуваної задачі. Під час збільшення кількості ознак кількість спостережень зростає нелінійно, ця проблема носить назву «прокляття розмірності». У разі недостатньої кількості даних рекомендується використовувати лінійну модель.

Аналітик повинен визначити кількість шарів у мережі і кількість нейронів у кожному шарі.

Далі необхідно призначити такі значення ваг і зміщень, які зможуть мінімізувати помилку рішення. Ваги і зміщення автоматично налаштовуються таким чином, щоб мінімізувати різницю між бажаним і отриманим на виході сигналами, яка називається помилкою навчання. Помилка навчання для побудованої нейронної мережі обчислюється шляхом порівняння вихідних і цільових (бажаних) значень. З отриманих різниць формується функція помилок.

Функція помилок – це цільова функція, що вимагає мінімізації в процесі керованого навчання нейронної мережі.

За допомогою функції помилок можна оцінити якість роботи нейронної мережі під час навчання (наприклад, сума квадратів помилок). Від якості навчання нейронної мережі залежить її здатність вирішувати поставлені перед нею завдання.

14.4. Моделі нейронних мереж

Розглянемо найбільш прості моделі нейронних мереж: одношаровий і багатошаровий персептрони.

Найпростіша модель нейронної мережі – одношаровий персептрон.

Одношаровий персептрон (персептрон Розенблатта) – одношарова нейронна мережа, всі нейрони якої мають жорстку порогову функцію активації.

Одношаровий персептрон має простий алгоритм навчання і здатний вирішувати лише найпростіші завдання. Ця модель викликала до себе великий інтерес на початку 1960-х років і стала поштовхом до розвитку штучних нейронних мереж. Класичний приклад такої нейронної мережі показаний на рис. 3.9.

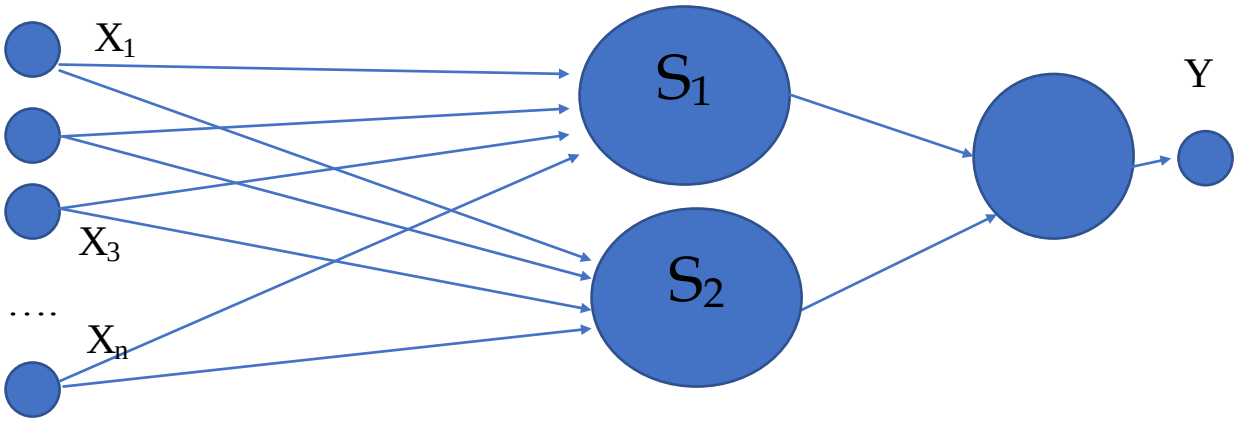


Рис. 3.9. Одношаровий персептрон

Мережа, зображена на рисунку, має *n* входів, на які надходять сигнали, що йдуть по синапсах на 2 нейрона. Ці два нейрона утворюють єдиний шар цієї мережі і видають два вихідних сигнали.

Багатошаровий персептрон (MLP) – нейронна мережа прямого поширення сигналу, в якій вхідний сигнал перетвориться у вихідний, проходячи послідовно через декілька шарів.

Перший з таких шарів називають вхідним, останній – вихідним. Ці шари містять так звані вироджені нейрони й іноді в кількості шарів не враховуються. Крім вхідного і вихідного шарів, у

багатошаровому персептроні є один або декілька проміжних шарів, які називають прихованими. У цій моделі персептрона повинен бути хоча б один прихований шар. Присутність декількох таких шарів виправдано лише в разі використання нелінійних функцій активації.

Метод зворотного поширення помилки – алгоритм навчання багатошарових персептронів, заснований на обчисленні градієнта функції помилок. У процесі навчання ваги нейронів кожного шару нейромережі коригуються з урахуванням сигналів, що надійшли з попереднього шару, і нев'язки кожного шару, яка обчислюється рекурсивно у зворотному напрямку від останнього шару до першого.

Запитання до параграфа 14

1. Які завдання Data Mining вирішуються за допомогою нейронних мереж?
2. Які є елементи нейронних мереж?
3. Що таке штучний нейрон?
4. Яка основна функція штучного нейрона?
5. Який вигляд штучного нейрона?
6. Яка архітектура нейронних мереж?
7. Що таке шаруваті нейронні мережі?
8. Що таке навчальна вибірка, яке її призначення?
9. Які є моделі нейронних мереж?
10. Що таке одношаровий персептрон?
11. Що таке багатошаровий персептрон?

§ 15. Кластеризація

15.1. Загальна характеристика задачі кластеризації

Задача кластеризації подібна до задачі класифікації і є її логічним продовженням, але її відмінність у тому, що класи досліджуваного набору даних заздалегідь не визначені. Синонімами терміна «кластеризація» є «автоматична класифікація», «навчання без вчителя» і «таксономія».

Кластеризація призначена для розбиття сукупності об'єктів на однорідні групи (кластери або класи). Якщо ці вибірки уявити як точки в просторі ознак, то задача кластеризації зводиться до визначення «згущень точок».

Мета кластеризації – пошук існуючих структур.

Кластеризація є описовою процедурою, вона не робить ніяких статистичних висновків, але дає можливість провести розвідувальний аналіз і вивчити «структуру даних». Саме поняття «кластер» визначено неоднозначно: в кожному дослідженні свої «кластери». Перекладається поняття «кластер» (cluster) як «скупчення», «гроно». Кластер можна охарактеризувати як групу об'єктів, що мають спільні властивості.

Характеристиками кластера можна назвати дві ознаки:

– внутрішня однорідність;

– зовнішня ізольованість.

Питання, що задається аналітиками під час вирішення багатьох завдань, полягає в тому, як організувати дані в наочні структури.

Найбільше застосування кластеризація спочатку отримала в таких науках, як біологія, антропологія, психологія. Для вирішення економічних завдань кластеризація тривалий час мало використовувалася через специфіку економічних даних і явищ.

У таблиці 3.10 наведено порівняння деяких параметрів задач класифікації та кластеризації.

Таблиця 3.10

Порівняння класифікації і кластеризації

Характеристика	Класифікація	Кластеризація
Контрольованість навчання	Контрольоване навчання	Неконтрольоване навчання
Стратегія	Навчання з учителем	Навчання без учителя
Наявність позначки класу	Навчальна множина супроводжується міткою, яка вказує клас, до якого належить спостереження	Мітки класу навчальної множини невідомі
Підстава для класифікації	Нові дані Класифікуються на підставі навчальної множини	Дано множину даних з метою встановлення існування класів або кластерів даних

Кластери можуть бути такими, що не перетинаються, або ексклюзивними, і такими, що перетинаються, чіткими множинами або нечіткими. Схематичне зображення непересічних і пересічних кластерів подано на рис. 3.10.

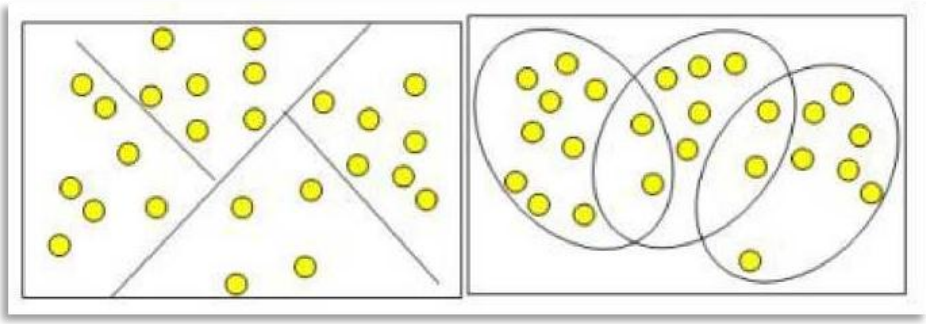


Рис. 3.10. Чіткі та нечіткі кластери

Варто зазначити, що в результаті застосування різних методів кластерного аналізу можуть бути отримані кластери різної форми. Наприклад, можливі кластери типу «ланцюга», коли кластери представлені довгими «ланцюгами», кластери подовженої форми тощо, а деякі методи можуть створювати кластери довільної форми.

Різні методи можуть прагнути створювати кластери певних розмірів (наприклад, малих або великих) чи припускати в наборі даних наявність кластерів різного розміру. Деякі методи кластерного аналізу особливо чутливі до шумів або викидів, інші – менш.

У результаті застосування різних методів кластеризації можуть бути отримані неоднакові результати, це нормально і є особливістю роботи певного алгоритму. Такі особливості варто враховувати під час вибору методу кластеризації.

У процесі виконання кластеризації важливо визначити, скільки в результаті цієї процедури повинно бути побудовано кластерів.

Апріорі передбачається, що кластеризація повинна виявити природні локальні згущення об'єктів. Тому кількість кластерів є параметром, який часто суттєво ускладнює алгоритм, якщо передбачається, що він невідомий і суттєво впливає на якість результату.

Алгоритм кластеризації зазвичай будується як певний спосіб перебору кількості кластерів і визначення його оптимального значення в процесі перебору та включає 5 основних кроків:

1. Відбір вибірки для кластеризації.

2. Визначення ознак, за якими будуть оцінюватися об'єкти у вибірці.

3. Обчислення значень певної міри подібності між об'єктами.

4. Застосування кластерного аналізу для створення груп подібних об'єктів.

5. Перевірка достовірності результатів кластерного розв'язку.

На сьогодні розроблено більше сотні різних алгоритмів кластеризації.

Наведемо коротку характеристику підходів до кластеризації.

1. Алгоритми, засновані на поділі даних, у тому числі ітеративні:

– поділ об'єктів на *k*-кластерів;

– ітеративний перерозподіл об'єктів для поліпшення кластеризації.

2. Ієрархічні алгоритми:

– агломерація: кожен об'єкт спочатку є кластером, кластери, з'єднуючись один з одним, формують більший кластер тощо.

3. Методи, що базуються на концентрації об'єктів:

– засновані на можливості з'єднання об'єктів;

– ігнорують шуми, знаходження кластерів довільної форми.

4. Грід-методи:

– квантування об’єктів у грід-структури.

5. Модельні методи:

– використання моделі для знаходження кластерів, які найбільше відповідають даним.

Оцінка якості кластеризації. Оцінка якості кластеризації може бути проведена на основі таких процедур:

– ручна перевірка;

– встановлення контрольних точок і перевірка на отриманих кластерах;

– визначення стабільності кластеризації шляхом додавання в модель нових змінних;

– створення і порівняння кластерів з використанням різних методів. Різні методи кластеризації можуть створювати різні кластери, і це нормальне явище. Однак створення подібних кластерів різними методами вказує на правильність кластеризації.

Процес кластеризації залежить від обраного методу і майже завжди є ітеративним. Він може включати множину експериментів з вибору різноманітних параметрів. Проте експерименти не повинні бути самоціллю, адже кінцевою метою кластеризації є отримання змістовних відомостей про структуру досліджуваних даних. Отримані результати вимагають подальшої інтерпретації, дослідження і вивчення властивостей і характеристик об’єктів для можливості точного опису сформованих кластерів.

Кластерний аналіз застосовується в різних сферах. Він корисний, коли потрібно класифікувати велику кількість інформації. Так, у медицині використовується кластеризація захворювань, лікування захворювань або їх симптомів, а також таксономія пацієнтів, препаратів тощо. В археології встановлюються таксономії кам’яних споруд і стародавніх об’єктів тощо. У маркетингу це може бути задача сегментації конкурентів і споживачів. У менеджменті прикладом завдання кластеризації буде розбиття персоналу на різні групи, класифікація споживачів і постачальників, виявлення подібних виробничих ситуацій, під час яких виникає шлюб. У медицині – класифікація симптомів. У соціології завдання кластеризації – розбиття респондентів на однорідні групи.

У маркетингових дослідженнях кластерний аналіз застосовується досить широко – як у теоретичних дослідженнях, так і практикуючими маркетологами. При цьому вирішуються питання про групи клієнтів, продуктів тощо. Так, одним з найбільш важливих завдань у процесі застосування кластерного аналізу в маркетингових дослідженнях є аналіз поведінки споживача, а саме групування споживачів в однорідні класи для отримання максимально повного уявлення про поведінку клієнта з кожної групи і про фактори, що впливають на його поведінку.

Термін «кластерний аналіз» включає в себе більше 100 різних алгоритмів. На відміну від завдань класифікації, кластерний аналіз не вимагає апріорних припущень про набір даних, що не накладає обмеження на подання досліджуваних об'єктів, дозволяє аналізувати показники різних типів даних. При цьому необхідно пам'ятати, що змінні повинні вимірюватися в порівнянних шкалах.

Кластерний аналіз дозволяє скорочувати розмірність даних, робити її наочною.

Кластерний аналіз може застосовуватися до сукупностей часових рядів, тут можуть виділятися періоди подібності деяких показників і визначатися групи часових рядів з подібною динамікою.

У більшості методів існує по дві і більше назви. Це істотно ускладнює роботу під час використання кластерного аналізу.

Задачі кластерного аналізу можна об'єднати в такі групи:

1. Розробка типології або класифікації.
2. Дослідження корисних концептуальних схем групування об'єктів.
3. Подання гіпотез на основі дослідження даних.
4. Перевірка гіпотез або досліджень для визначення, чи дійсно типи (групи), виділені певним способом, присутні в наявних даних.

Як правило, у процесі практичного використання кластерного аналізу одночасно вирішується кілька із зазначених завдань.

Розглянемо приклад процедури кластерного аналізу. Припустимо, що маємо набір даних А, який складається з 14-ти прикладів, у яких є по дві ознаки X і Y. Дані наведені в таблиці 3.11.

Таблиця 3.11

Набір даних А

№ прикладу	Ознака X	Ознака Y
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47

Дані в табличній формі не носять інформативний характер. Уявімо змінні X і Y у вигляді діаграми розсіювання, зображеної на рис. 3.11.

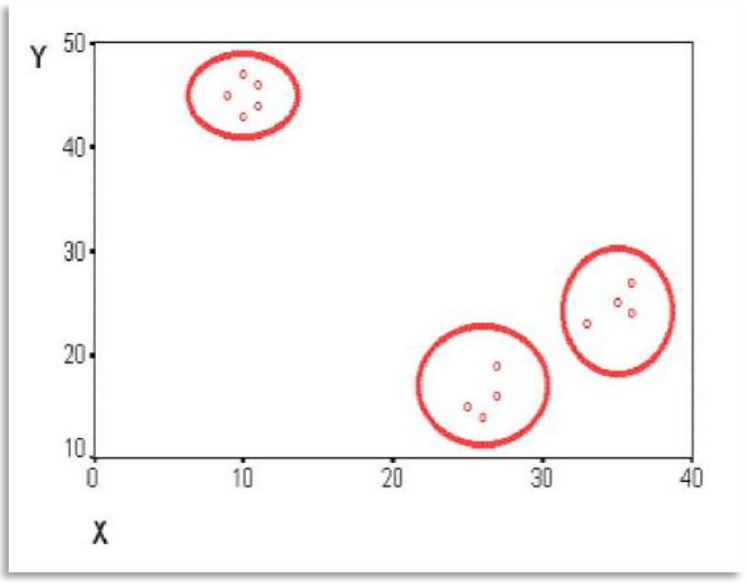


Рис. 3.11. Діаграма розсіювання

На рисунку видно кілька груп «подібних» прикладів. Приклади (об’єкти), які за значеннями X і Y «подібні» один на одного, належать до однієї групи (кластера); об’єкти з різних кластерів не подібні один на одного.

Критерієм для визначення подібності та відмінності кластерів є відстань між точками на діаграмі розсіювання. Цю подібність можна «виміряти», вона дорівнює відстані між точками на графіку. Способів визначення міри відстані між кластерами, яку називають ще мірою близькості, існує декілька. Найбільш поширений – обчислення евклідової відстані між двома точками на площині, коли відомі їх координати X і Y:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Кластер має такі математичні характеристики: центр, радіус, середньоквадратичне відхилення, розмір кластера.

Центр кластера – це середнє геометричне місце точок у просторі змінних.

Радіус кластера – максимальна відстань точок від центру кластера.

Кластери можуть перекриватися. У цьому випадку неможливо за допомогою математичних процедур однозначно віднести об’єкт до одного з двох кластерів. Такі об’єкти називають спірними.

Спірний об’єкт – це об’єкт, який за подібністю може бути віднесений до кількох кластерів.

Розмір кластера може бути визначений або за радіусом кластера, або за середньоквадратичним відхиленням об’єктів для цього кластера. Об’єкт належить до кластера, якщо відстань від об’єкта до центра кластера менша, ніж радіус кластера. Якщо ця умова виконується для двох і більше кластерів, об’єкт є спірним. Неоднозначність цього завдання може бути усунена експертом або аналітиком.

Кластерний аналіз спирається на два припущення. Перше припущення – ознаки об’єкта в принципі допускають бажане розбиття сукупності об’єктів на кластери. Друге – правильність вибору масштабу або одиниць вимірювання ознак.

Вибір масштабу в кластерному аналізі має велике значення. Розглянемо приклад. Уявімо, що дані ознаки *x* у наборі даних A на два порядки більше даних ознаки *y*: значення змінної *x*

перебувають у діапазоні від 100 до 700, а значення змінної *y* – в діапазоні від 0 до 1. Тоді у разі розрахунку величини відстані між точками, що відбивають стан об'єктів у просторі їх властивостей, змінна має велике значення, тобто змінна *x* буде практично повністю домінувати над змінною з малими значеннями, тобто змінною *y*. Таким чином, через неоднорідність одиниць виміру ознак стає неможливо коректно поділити об'єкти на класи.

Ця проблема вирішується за допомогою попередньої стандартизації змінних. Стандартизація або нормалізація приводить значення всіх перетворених змінних до єдиного діапазону значень. Існують різні способи нормалізації вихідних даних.

Поряд зі стандартизацією змінних існує варіант додавання кожній з них певного коефіцієнта ваги, який би відображав значимість відповідної змінної. Як ваги, можуть виступати експертні оцінки, отримані в ході опитування експертів-фахівців предметної сфери. Отримані добутки нормованих змінних на відповідні ваги дозволяють отримувати відстані між точками в багатовимірному просторі з урахуванням різної ваги змінних.

У ході експериментів можливе порівняння результатів, отриманих з урахуванням експертних оцінок і без них, і вибір кращого з них.

Методи кластерного аналізу можна розділити на дві групи:

– ієрархічні (агломеративні – методи зв'язку (метод одиночного зв'язку, метод повного зв'язку, метод середнього зв'язку); центроїдні методи, дисперсійні методи (метод Варди);

– неієрархічні (послідовний пороговий метод, паралельний пороговий метод, метод оптимізуючого розподілу).

Кожна з груп включає сукупність підходів і алгоритмів.

Використовуючи різні методи кластерного аналізу, аналітик може отримати різні рішення для одних і тих самих даних. Це вважається нормальним явищем.

15.2. Методи кластерного аналізу. Ієрархічні методи

Ієрархічна кластеризація – це така кластеризація, за якої, починаючи з кластера, що складається з одного елемента, кластери ітеративно зливаються в кластери вищого рівня. Також можливо починати з єдиного великого макрокластера, який ітеративно розділяється на маленькі кластери. Таким чином, формується ієрархія кластерів. Для їх формування не потрібно задавати кількість кластерів, такий тип кластеризації є більш детермінованим та не потребує ітеративних уточнень. Отже, сутність ієрархічної кластеризації полягає в послідовному об'єднанні менших кластерів у великі або поділі великих кластерів на менші.

Ієрархічні методи кластерного аналізу характеризуються тим, що процес об'єднання об'єктів під час їх використання має ієрархічний характер і може бути поданий у вигляді дендрограми (деревоподібної діаграми).

Дендрограма – деревоподібна діаграма, що містить *n* рівнів, кожний з яких відповідає одному із кроків процесу послідовного укрупнення кластерів. При цьому на кожному кроці кількість кластерів змінюється в бік збільшення або зменшення. Робота більшості таких методів ґрунтується на обчисленні матриці подібності, яка містить міри відстаней – числа, що виражають «подібність» кожних двох об'єктів.

Ієрархічні агломеративні методи (Agglomerative Nesting, AGNES)

Ця група методів характеризується послідовним об'єднанням вихідних елементів і відповідним зменшенням кількості кластерів.

На початку роботи алгоритму всі об'єкти є окремими кластерами. На першому кроці найбільш подібні об'єкти об'єднуються в кластер. На наступних кроках об'єднання триває доти, доки всі об'єкти не складатимуть один кластер.

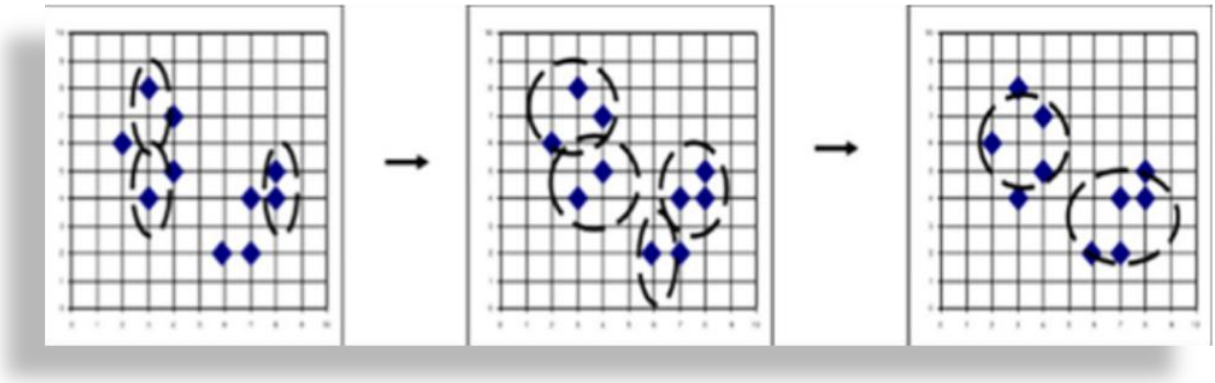
Агломеративні алгоритми – це такі, що кластеризують «знизу вверх». На початку алгоритму кожна точка розглядається як кластер, потім алгоритм намагається об'єднати найближчі сусідні точки в один більший кластер тощо, щоб зрештою об'єднати всі кластери в один великий кластер.

Агломеративні алгоритми також називають AGNES (AGglomerative NESting). Хід алгоритму виглядає таким чином:

– використовується метод одноканального зв'язку «найближчий сусід» та матриця відмінностей;

– вузли, що мають найменші відмінності, зливаються;

– усі вузли об'єднуються в один кластер.



Агломеративна кластеризація залежить від використання мір подібності кластерів:

– одноканальний зв'язок (найближчий сусід);

– повний зв'язок (діаметр);

– середній зв'язок (середнє за групою);

– центроїдний зв'язок (подібність центроїдів).

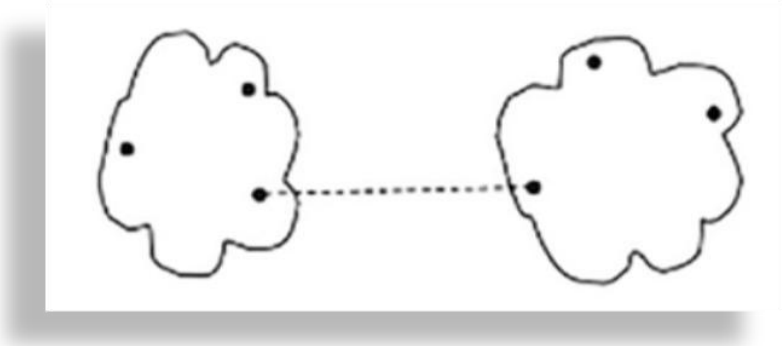
Одноканальний зв'язок (найближчий сусід):

– подібність двох кластерів – це подібність між їх найбільш подібними членами (найближчий сусід);

– приділяється увага найближчим точкам, ігнорується структура кластера;

– можливість будувати кластери неправильної форми;

– такий вид зв'язку чутливий до даних з шумами та значень, що «випадають» з множини.



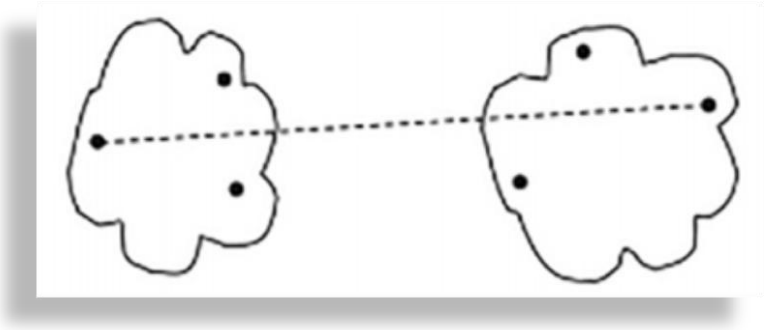
Повний зв’язок:

– подібність двох кластерів рахується як подібність їх найменш подібних членів;

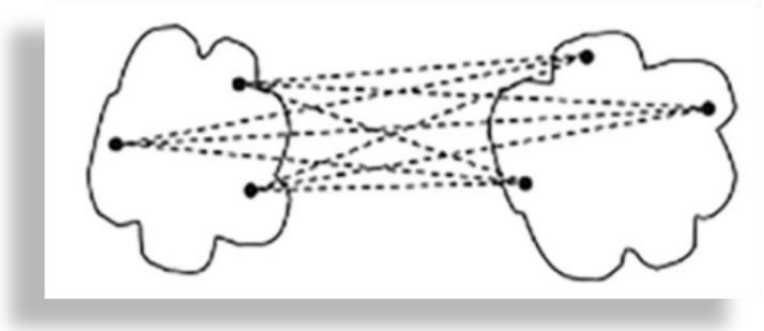
– два кластери, об’єднуючись, формують кластер із щонайменшим діаметром;

– на виході – кластери компактної форми;

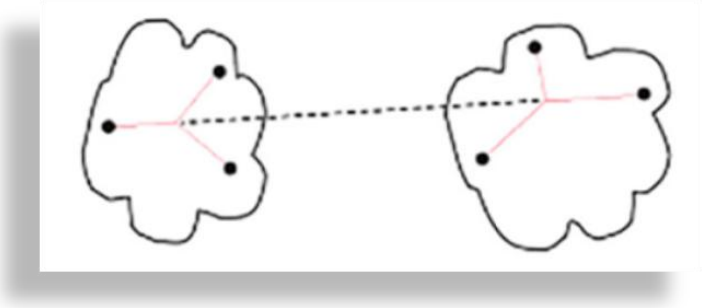
– чутливий до значень, що суттєво відрізняються.



Середній зв’язок – середня відстань між елементами в парі кластерів. Особливістю є затратне обчислення.

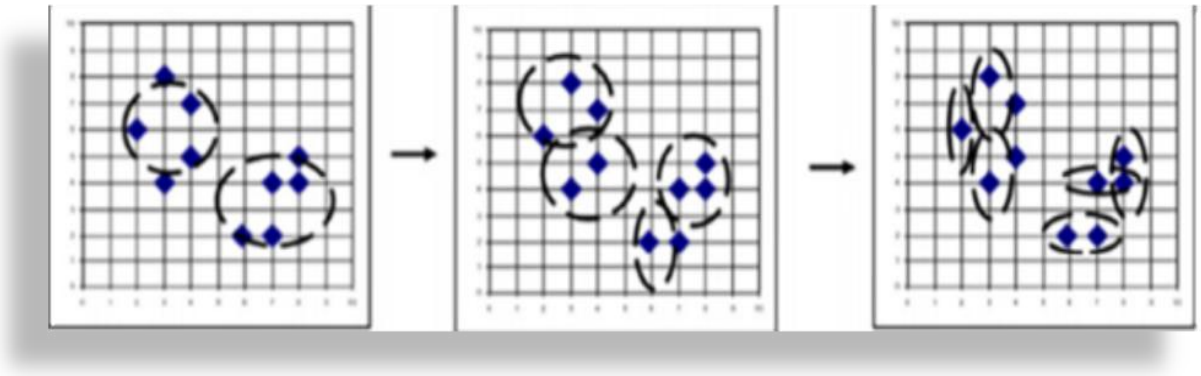


Центроїдний зв’язок – відстань між центроїдами двох кластерів. Цей алгоритм не потребує задання кількості кластерів та дозволяє об’єднувати в кластери дані без істотних змін через значення, що вибиваються, та шуми.



Ієрархічні дивізимні (подільні) методи.

Ці методи є логічною протилежністю агломеративним методам. На початку роботи алгоритму всі об'єкти належать одному кластеру, який на наступних кроках поділяються на менші кластери, в результаті утворюється послідовність розщеплених груп. Тобто DIANA (Divisive Analysis) – зворотний порядок дій від AGNES: у результаті кожний елемент являє собою кластер.



Ієрархічна кластеризація розділенням – це підхід «зверху вниз»:

– процес починається з кореня, розглядаючи всі точки множини як кластер;

– кластери вищого рівня рекурсивно розщеплюються для побудови діаграми;

– може розглядатися як глобальний підхід;

– може вважатися ефективнішим, але більш чутливим до шумів за AGNES.

Принцип роботи описаних вище груп методів у вигляді дендрограми показаний на рис. 3.12.

Програмна реалізація алгоритмів кластерного аналізу широко подана в різних інструментах Data Mining, які дозволяють вирішувати завдання досить великої розмірності.

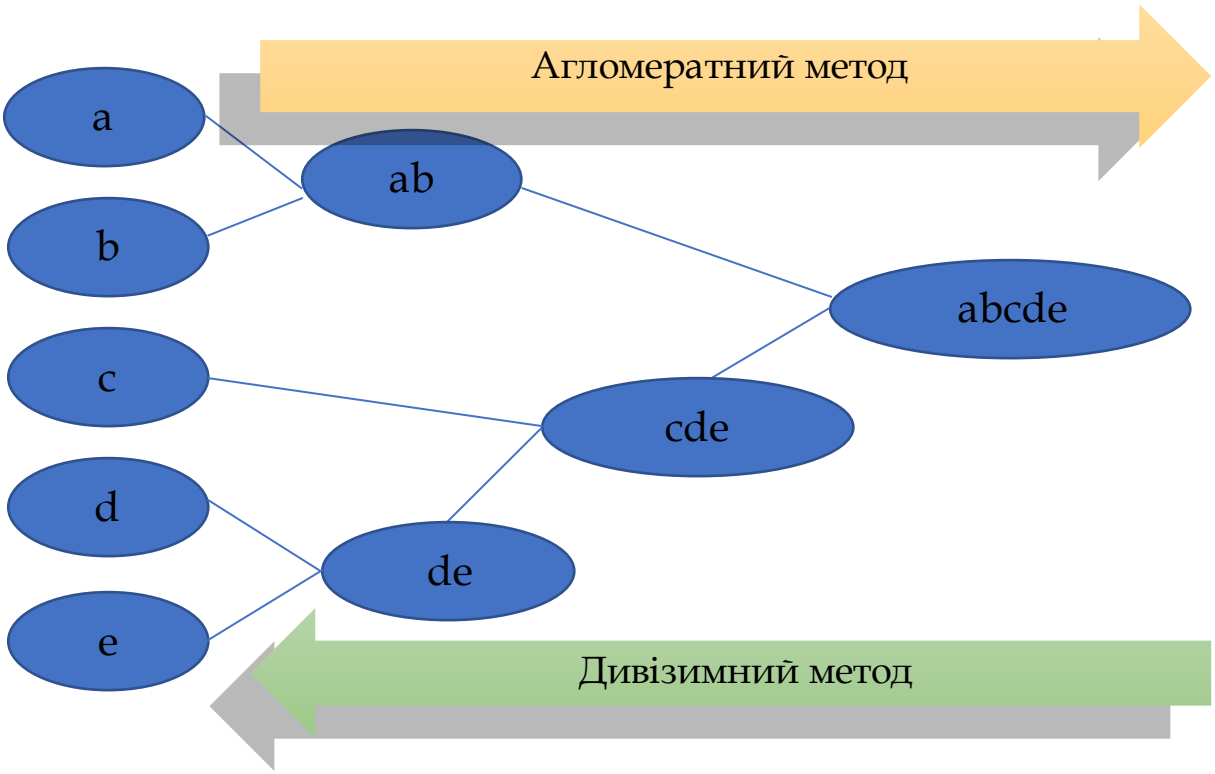


Рис. 3.12. Дендрограма агломеративних і дивізимних методів

Ієрархічні методи кластеризації розрізняються правилами побудови кластерів. Як правило, виступають критерії, які використовуються під час вирішення питання про «подібності» об'єктів у процесі їх об'єднання в групу (агломеративні методи) або поділу на групи (дивізимні методи).

Ієрархічні методи кластерного аналізу використовуються за невеликих обсягів наборів даних. Перевагою ієрархічних методів кластеризації є їх наочність.

Ієрархічні алгоритми пов'язані з побудовою дендрограм, які є результатом ієрархічного кластерного аналізу. Дендрограма описує близькість окремих точок і кластерів один до одного, у графічному вигляді являє собою послідовність об'єднання (поділу) кластерів.

Дендрограму також називають деревовидною схемою, деревом об'єднання кластерів, деревом ієрархічної структури.

Існує багато способів побудови дендограм. У дендограмах об'єкти можуть розташовуватися вертикально або горизонтально. Приклад вертикальної дендрограми наведено на рис. 3.13.

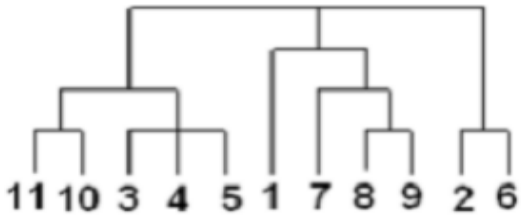


Рис. 3.13. Приклад дендрограми

Числа 11, 10, 3 тощо відповідають номерам об’єктів або спостережень вихідної вибірки. Бачимо, що на першому етапі кожне спостереження представляє один кластер (вертикальна лінія), на другому кроці спостерігаємо об’єднання таких спостережень: 11 і 10; 3, 4 і 5; 8 і 9; 2 і 6. На другому кроці триває об’єднання в кластери: спостереження 11, 10, 3, 4, 5 і 7, 8, 9. Цей процес триває доти, доки всі спостереження не об’єднаються в один кластер.

Міра подібності

Для обчислення відстані між об’єктами використовуються різні міри подібності, звані також метриками або функціями відстаней.

1. Квадрат евклідової відстані. Для додання великих ваг більш віддаленим один від одного об’єктів можемо скористатися квадратом евклідової відстані шляхом зведення в квадрат стандартної евклідової відстані.

2. Манхеттенська відстань (відстань міських кварталів), також яку ще називають «сіті-блок» відстанню.

$$d(q_i, q_j) = \sum_{k=1}^{m} \left| x_{i,k} - x_{j,k} \right|$$

Ця відстань розраховується як середнє різниць за координатами. Здебільшого ця міра відстані приводить до результатів, подібних до результатів під час розрахунків відстані Евкліда. Однак при цьому підході вплив окремих викидів менший, ніж у процесі використання евклідової відстані, оскільки тут координати не підносяться до квадрату.

3. Відстань Чебишева.

$$d(q_i, q_j) = \max_{1 \le k \le m} \left| x_{i,k} - x_{j,k} \right|$$

Цю відстань варто використовувати, коли необхідно визначити два об'єкти як «різні», за умови, що вони відрізняються за якимось одним вимірюванням.

4. Відсоток незгоди. Ця відстань обчислюється, якщо дані є категоріальними.

5. Евклідова відстань. Це найуживаніша міра відстані між об'єктами, яка являє собою геометричну відстань між об'єктами у багатомірному просторі.

$$d(q_i, q_j) = \sqrt{\sum_{k=1}^{m} (x_{i,k} - x_{j,k})^2}$$

Евклідова відстань обчислюється за вихідними нестандартизованими даним, а тому всі змінні повинні бути виміряні в одному масштабі (якщо це сантиметри, то всі змінні повинні бути виміряні в сантиметрах тощо).

6. Коефіцієнт кореляції Пірсона. Використовується у випадку, коли абсолютні значення та різниці між об'єктами несуттєві, а більш важливим є наявність зв'язку між ними. Ця міра чутлива лише до подібності профілів об'єктів. Приводить до результатів, близьких до факторного аналізу – кластери можуть наближатися до факторів.

7. Коефіцієнт кореляції Спірмена.

8. Відстань Махаланобіса. Обчислюється за формулою:

$$d(q_i, q_j) = (x_i - x_j) S^{-1} (x_i - x_j)^t$$

де – коваріаційна матриця розмірності $m \times m$, t – символ транспонування.

Методи об'єднання або зв'язку

Коли кожен об'єкт являє собою окремий кластер, відстані між цими об'єктами визначаються обраною мірою. Виникає таке питання: як визначити відстані між кластерами? Існують різні правила, які називають методами об'єднання або зв'язку для двох кластерів.

1. Метод ближнього сусіда або одиночний зв'язок. Відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) в різних кластерах. Цей метод дозволяє виділяти кластери як завгодно складної форми за умови, що різні частини таких кластерів з'єднані ланцюгами, близьких один до одного елементів. У результаті роботи цього методу кластери подаються довгими «ланцюгами» або «волокнистими» кластерами, «зчепленими разом» тільки окремими елементами, які випадково опинилися ближче інших один до одного.

2. Метод найбільш віддалених сусідів або повний зв'язок. Тут відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах («найбільш віддаленими сусідами»). Метод добре використовувати, коли об'єкти дійсно належать до різних класів. Якщо ж кластери мають дещо подовжену форму або їх природний тип є «ланцюговий», то цей метод не потрібно використовувати.

3. Метод Варда (Ward's). Як відстань між кластерами, береться приріст суми квадратів відстаней об'єктів до центрів кластерів, які утворюються в результаті їх об'єднання. На відміну від інших методів кластерного аналізу для оцінки відстаней між кластерами тут використовуються методи дисперсійного аналізу. На кожному кроці алгоритму об'єднуються такі два кластери, які приводять до мінімального збільшення цільової функції, тобто внутрішньогрупової суми квадратів. Цей метод направлений на об'єднання близько розташованих кластерів і «прагне» створювати кластери малого розміру.

4. Метод невиваженого попарного середнього (метод невиваженого попарного арифметичного). За відстань між двома кластерами береться середня відстань між усіма парами об'єктів у них. Цей метод варто використовувати, якщо об'єкти дійсно походять з різних «класів», у випадках присутності кластерів «ланцюгового» типу, у разі припущення нерівності розмірів кластерів.

5. Метод зваженого попарного середнього (метод зваженого попарного арифметичного середнього). Цей метод подібний на метод незваженого попарного середнього, різниця полягає лише в

тому, що тут як ваговий коефіцієнт використовується розмір кластера (кількість об'єктів, що містяться в кластері). Цей метод рекомендується використовувати за наявності припущення існування кластерів різних розмірів.

6. Незважений центроїдний метод (метод невиваженого попарного центроїдного). Як відстань між двома кластерами, в цьому методі береться відстань між їх центрами ваги.

7. Зважений центроїдний метод (метод зваженого попарного центроїдного усереднення). Цей метод подібний на попередній, різниця полягає в тому, що для розрахунку використовуються ваги. Цей метод переважно використовувати у випадках, якщо є припущення щодо істотних відмінностей у розмірах кластерів.

15.3. Методи кластерного аналізу. Ітеративні методи

За великої кількості спостережень ієрархічні методи кластерного аналізу не придатні. У таких випадках використовують неієрархічні методи, засновані на поділі, які являють собою ітеративні методи дроблення вихідної множини. У процесі поділу нові кластери формуються доти, доки не буде виконано правило зупинки.

Така неієрархічна кластеризація полягає в поділі набору даних на певну кількість окремих кластерів. Існує два підходи. Перший – полягає у визначенні меж кластерів як найбільш щільних ділянок у багатовимірному просторі вихідних даних, тобто визначення кластера там, де є велике «згущення точок». Другий – у мінімізації відмінності об'єктів.

Алгоритм k-means

Найбільш поширений серед неієрархічних методів алгоритм k-середніх, який також називають швидким кластерним аналізом. На відміну від ієрархічних методів, які не вимагають попередніх припущень щодо кількості кластерів, для використання цього методу необхідно мати гіпотезу про найбільш імовірну кількість кластерів.

Мета методу – розділити n спостережень на k кластерів так, щоб кожне спостереження належало до кластера з найближчим

до нього середнім значенням. Метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера.

Алгоритм k-середніх будує *k* кластерів, розташованих на якомога більших відстанях один від одного. Основний тип задач, які вирішує алгоритм k-середніх, – наявність припущень (гіпотез) щодо кількості кластерів, при цьому вони повинні бути різні настільки, наскільки це можливо. Вибір числа *k* може базуватися на результатах попередніх досліджень, теоретичних міркуваннях або інтуїції.

Загальна ідея алгоритму: задане фіксоване число *k* кластерів, спостереження розподіляються за кластерами так, що середні в кожному кластері (для всіх змінних) максимально відрізняються один від одного.

Маємо масив спостережень (об'єктів), кожен з яких має певні значення за деякими ознаками. Відповідно до цих значень об'єкт розташовується у багатовимірному просторі.

1. Дослідник визначає кількість кластерів, що необхідно утворити.

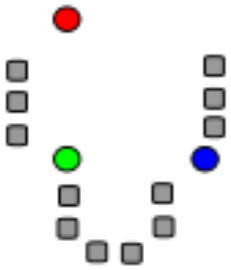
2. Випадковим чином обирається *k* спостережень, які на цьому кроці вважаються центрами кластерів.

3. Кожне спостереження «приписується» до одного з *n* кластерів – того, відстань до якого найкоротша.

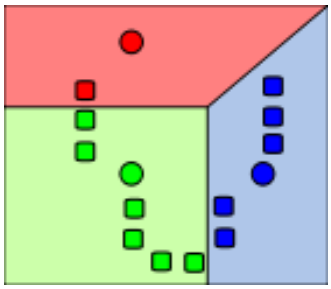
4. Розраховується новий центр кожного кластера як елемент, ознаки якого розраховуються як середнє арифметичне ознак об'єктів, що входять у цей кластер.

5. Відбувається така кількість ітерацій (повторюються кроки 3–4), доки кластерні центри стануть стійкими (тобто за кожної ітерації в кожному кластері опинятимуться одні й ті самі об'єкти), дисперсія всередині кластера буде мінімізована, а між кластерами – максимізована.

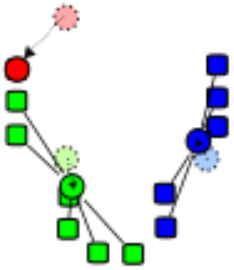
Вибір кількості кластерів відбувається на основі дослідницької гіпотези. Якщо її немає, то рекомендують створити 2 кластери, далі 3, 4, 5, порівнюючи отримані результати.



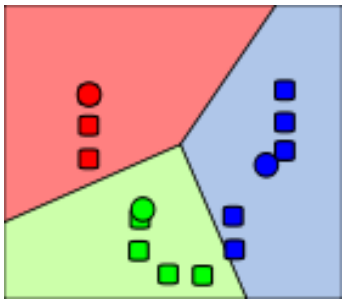
1. k початкових «середніх» (тут $k = 3$) випадково згенеровані у межах домени даних (кольорові).



2. Створено k кластерів, асоціюючи кожне спостереження з найближчим середнім. Розбиття відбувається згідно з діаграмою Вороного, утвореною середніми.



3. Центроїд кожного з k кластерів стає новим середнім.



4. Кроки 2 і 3 повторюються до досягнення збіжності.

На рис. 3.14 наведено приклад роботи алгоритму k-середніх для k = 2.

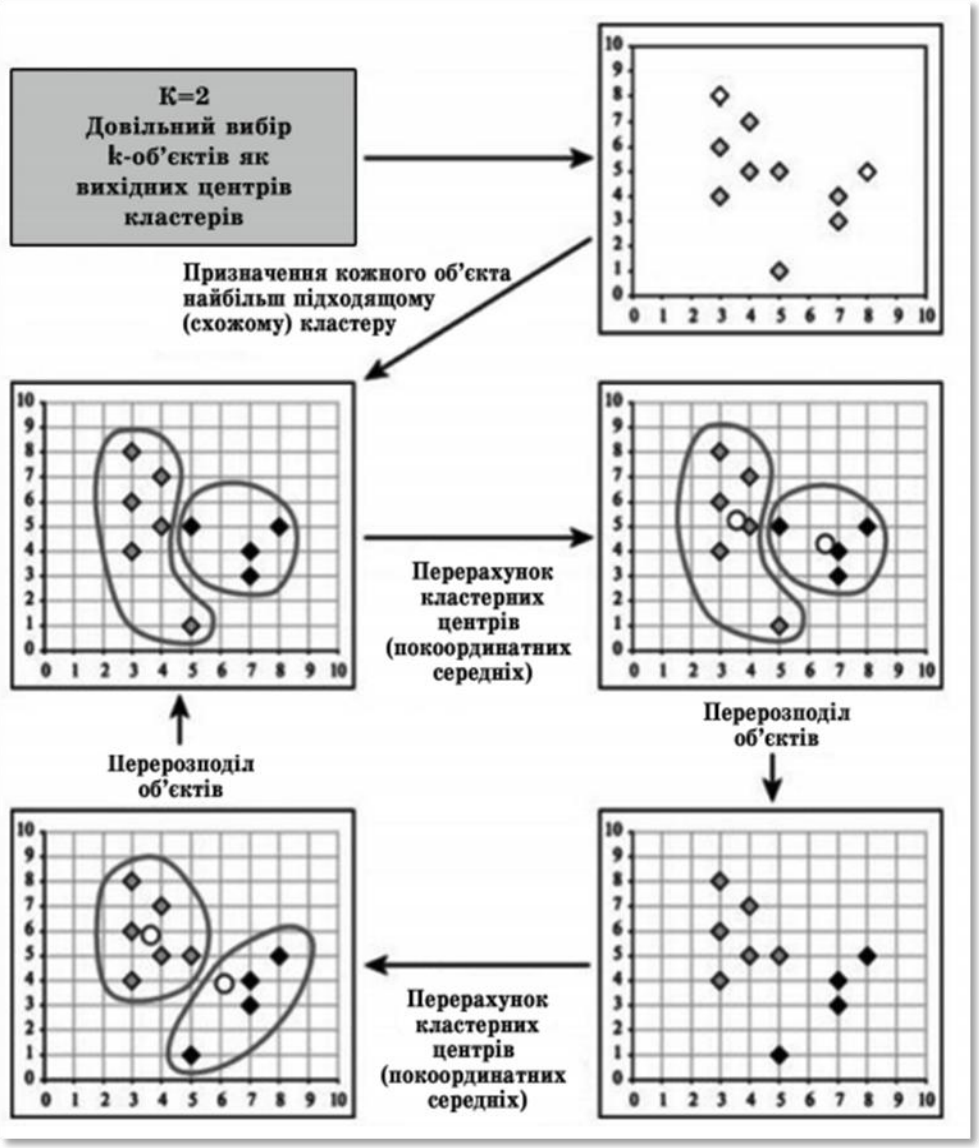


Рис. 3.14. Приклад роботи алгоритму k-середніх для k = 2

Перевірка якості кластеризації.

Після отримання результатів кластерного аналізу методом k-середніх варто перевірити правильність кластеризації (тобто оцінити, наскільки кластери відрізняються один від одного). Для цього розраховуються середні значення для кожного кластера. За якісної кластеризації повинні бути отримані середні для всіх вимірювань, які значно відрізняються.

Переваги алгоритму k-середніх:

– простота використання;

– швидкість використання;

– зрозумілість і прозорість алгоритму.

Недоліки алгоритму k-середніх:

– алгоритм занадто чутливий до викидів, які можуть спотворювати середнє. Можливим вирішенням цієї проблеми є використання модифікації алгоритму – алгоритм k-медіани;

– алгоритм може повільно працювати на великих базах даних. Можливим вирішенням цієї проблеми є використання вибірки даних.

Алгоритм PAM (partitioning around Medoids). PAM є модифікацією алгоритму k-середній, алгоритмом k-медіани (k-medoids). Алгоритм менш чутливий до шумів і викидів даних, ніж алгоритм k-means, оскільки медіана менше піддається впливам викидів. PAM ефективний для невеликих баз даних, але його не варто використовувати для великих наборів даних.

Попереднє скорочення розмірності. Розглянемо приклад. Є база даних клієнтів фірми, яких потрібно розбити на однорідні групи. Кожен клієнт описується за допомогою 25 змінних. Використання такої великої кількості змінних призводить до виділення кластерів нечіткої структури. У результаті аналітику досить складно інтерпретувати отримані кластери.

Більш зрозумілі і прозорі результати кластеризації можуть бути отримані, якщо замість множини вихідних змінних використовувати певні узагальнені змінні або критерії, які містять у стислому вигляді інформацію про зв'язки між змінними. Тобто виникає задача зниження розмірності даних. Вона може вирішуватися за допомогою різних методів; один з найбільш поширених – факторний аналіз.

Факторний аналіз. Факторний аналіз – це метод, застосовуваний для вивчення взаємозв'язків між значеннями змінних. Взагалі факторний аналіз переслідує дві мети:

– скорочення кількості змінних;

– класифікація змінних – визначення структури взаємозв'язків між змінними.

Відповідно, факторний аналіз може використовуватися для вирішення завдань скорочення розмірності даних або для вирішення задач класифікації.

Критерії або основні чинники, виділені в результаті факторного аналізу, містять у стислому вигляді інформацію про існуючі зв'язки між змінними. Ця інформація дозволяє отримати кращі результати кластеризації і краще пояснити семантику кластерів. Самим факторам може бути наданий певний сенс.

За допомогою факторного аналізу велика кількість змінних зводиться до меншої кількості незалежних величин, які називаються факторами. Фактор у «стиснутому» вигляді містить інформацію про декілька змінних. В один фактор об'єднуються змінні, які сильно корелюють між собою. У результаті факторного аналізу знаходяться такі комплексні фактори, які якомога повніше пояснюють зв'язок між розглянутими змінними. На першому кроці факторного аналізу здійснюється стандартизація значень змінних.

Факторний аналіз спирається на гіпотезу про те, що аналізовані змінні є непрямими проявами порівняно невеликої кількості будь-яких прихованих чинників.

Факторний аналіз – це сукупність методів, орієнтованих на виявлення та аналіз прихованих залежностей між спостережуваними змінними. Приховані залежності також називають латентними. Один з методів факторного аналізу – метод основних компонент – заснований на припущенні про незалежність факторів один від одного.

Рекомендовані етапи процесу кластерного аналізу. У загальному випадку всі етапи кластерного аналізу взаємопов'язані, і рішення, прийняті на одному з них, визначають дії на таких етапах:

1. Аналітику варто вирішити, чи використовувати всі спостереження, чи ж виключити деякі дані або вибірки з набору даних.

2. Вибір метрики і методу стандартизації вихідних даних.

3. Визначення кількості кластерів (для ітеративного кластерного аналізу).

4. Визначення методу кластеризації (правила об'єднання або зв'язку).

На думку багатьох фахівців, вибір методу кластеризації є вирішальним під час визначення форми та специфіки кластерів.

Аналіз результатів кластеризації. Цей етап має на увазі рішення таких питань:

– чи не є отримане розбиття на кластери випадковим;

– чи є розбиття надійним і стабільним на підвибірках даних;

– чи існує взаємозв'язок між результатами кластеризації і змінними, які не брали участі у процесі кластеризації;

– чи можна інтерпретувати отримані результати кластеризації.

Перевірка результатів кластеризації. Результати кластеризації також повинні бути перевірені формальними і неформальними методами. Формальні методи залежать від того методу, який використовувався для кластеризації. Неформальні включають такі процедури перевірки якості кластеризації:

– аналіз результатів кластеризації, отриманих на певних вибірках набору даних;

– крос-перевірка;

– проведення кластеризації під час зміни порядку спостережень у наборі даних;

– проведення кластеризації у разі видалення деяких спостережень;

– проведення кластеризації на невеликих вибірках.

Один з варіантів перевірки якості кластеризації – використання лише одного методу і порівняння отриманих результатів. Відсутність подібності результатів після застосування різних методів кластеризації не означає некоректність результатів, але присутність подібних груп вважається ознакою якісної кластеризації.

Складності і проблеми, які можуть виникнути у процесі застосування кластерного аналізу.

Як і будь-які інші методи, методи кластерного аналізу мають певні слабкі сторони, тобто деякі складності, проблеми та обмеження.

Під час проведення кластерного аналізу варто враховувати, що результати кластеризації залежать від критеріїв розбиття сукупності вихідних даних. У разі зниження розмірності даних можуть виникнути певні спотворення, за рахунок узагальнень можуть загубитися деякі індивідуальні характеристики об'єктів.

Існує ряд складностей, які потрібно продумати перед проведенням кластеризації.

1. Складність вибору характеристик, на основі яких проводиться кластеризація. Необдуманий вибір призводить до неадекватного розбиття на кластери і, як наслідок, – до неправильного рішення задачі.

2. Складність вибору методу кластеризації. Цей вибір вимагає хороші знання методів і передумов їх використання. Щоб перевірити ефективність конкретного методу в певній предметній сфері, доцільно застосувати таку процедуру: розглядають декілька апріорі різних між собою груп і випадково перемішують їх представників між собою. Далі проводиться кластеризація для відновлення вихідного розбиття на кластери. Частка збігів об'єктів у виявлених і вихідних групах є показником ефективності роботи методу.

3. Проблема вибору кількості кластерів. Якщо немає ніяких відомостей щодо можливої кількості кластерів, необхідно провести ряд експериментів і у результаті перебору різної кількості кластерів вибрати оптимальну їх кількість.

4. Проблема інтерпретації результатів кластеризації. Форма кластерів здебільшого визначається вибором методу об'єднання. Однак варто враховувати, що конкретні методи прагнуть створювати кластери певних форм, навіть якщо в досліджуваному наборі цих кластерів насправді немає.

15.4. Порівняльний аналіз ієрархічних і неієрархічних методів кластеризації

Перед проведенням кластеризації в аналітика може виникнути питання: якій групі методів кластерного аналізу віддати перевагу. Вибираючи між ієрархічними і неієрархічними методами, необхідно враховувати такі їх особливості.

1. Неієрархічні методи виявляють більш високу стійкість щодо шумів і викидів, некоректного вибору метрики, включення незначущих змінних у набір, який бере участь у кластеризації. Ціною, яку доводиться платити за ці переваги методу, є слово «апріорі». Аналітик повинен заздалегідь визначити кількість кластерів, кількість ітерацій або правило зупинки, а також деякі інші параметри кластеризації.

2. Якщо немає припущень щодо кількості кластерів, рекомендують використовувати ієрархічні алгоритми. Однак якщо обсяг вибірки не дозволяє це зробити, можливий шлях – проведення

ряду експериментів з різною кількістю кластерів, наприклад почати розбиття сукупності даних з двох груп і, поступово збільшуючи їх кількість, порівнювати результати. За рахунок такого «варіювання» результатів досягається досить велика гнучкість кластеризації.

3. Ієрархічні методи на відміну від неієрархічних будують повне дерево вкладених кластерів.

Складності ієрархічних методів кластеризації:

– обмеження обсягу набору даних; вибір міри близькості;

– негнучкість отриманих класифікацій.

Перевага цієї групи методів, порівняно з неієрархічними методами, – їх наочність і можливість отримати детальне уявлення про структуру даних.

Під час використання ієрархічних методів існує можливість досить легко ідентифікувати викиди в наборі даних і в результаті підвищити якість даних. Ця процедура полягає в основі двокрокового алгоритму кластеризації. Такий набір даних у подальшому може бути використаний для проведення ієрархічної кластеризації.

Існує ще один аспект. Це питання кластеризації всієї сукупності даних або ж її вибірки. Названий аспект істотний для обох розглянутих груп методів, однак він більш критичний для ієрархічних. Ієрархічні методи не можуть працювати з великими наборами даних, а використання деякої вибірки, тобто частини даних, могло б дозволити застосовувати ці методи.

Результати кластеризації можуть не мати достатнього статистичного обґрунтування. З іншого боку, у разі вирішення задач кластеризації допустима нестатистична інтерпретація отриманих результатів, а також досить велика різноманітність варіантів поняття кластера. Така нестатистична інтерпретація дає можливість аналітику отримати результати кластеризації достатньої якості, що під час використання інших методів часто буває важко.

15.5. Нові алгоритми і деякі модифікації алгоритмів кластерного аналізу

У зв'язку з появою надвеликих баз даних з'явилися нові вимоги, яким повинен задовольняти алгоритм кластеризації. Основна з них – це масштабованість алгоритму.

Зазначимо про інші властивості, яким повинен задовольняти алгоритм кластеризації:

– незалежність результатів від порядку вхідних даних;

– незалежність параметрів алгоритму від вхідних даних.

Останнім часом ведуться активні розробки нових алгоритмів кластеризації, здатних обробляти надвеликі бази даних. У них основна увага приділяється масштабованості. До таких алгоритмів належать узагальнене уявлення кластерів, а також вибірка і використання структур даних.

Розроблено алгоритми, в яких методи ієрархічної кластеризації інтегровані з іншими методами. До таких алгоритмів відносять: BIRCH, CURE, CHAMELEON, ROCK.

Алгоритм BIRCH. Завдяки узагальненим представленням кластерів швидкість кластеризації збільшується, алгоритм при цьому має велике масштабування.

У цьому алгоритмі реалізований двоетапний процес кластеризації.

У ході першого етапу формується попередній набір кластерів. На другому до виявлених кластерів застосовуються інші алгоритми кластеризації, придатні для роботи в оперативній пам'яті.

Якщо кожен елемент даних уявити собі як намистину, що лежить на поверхні столу, то кластери намистин можна «замінити» тенісними кульками і перейти до більш детального вивчення кластерів тенісних кульок. Кількість намистин може виявитися досить значна, проте діаметр тенісних кульок можна підібрати таким чином, щоб на другому етапі можна було, застосувавши традиційні алгоритми кластеризації, визначити дійсну складну форму кластерів.

Алгоритм WaveCluster. WaveCluster є алгоритм кластеризації на основі хвильових перетворень. На початку роботи алгоритму

дані узагальнюються шляхом накладення на простір даних багатовимірної решітки. На подальших кроках алгоритму аналізуються не окремі точки, а узагальнені характеристики точок, які потрапили в одну клітинку решітки. У результаті такого узагальнення необхідна інформація вміщується в оперативній пам'яті. Далі для визначення кластерів алгоритм застосовує хвильове перетворення до узагальнених даних.

Основні особливості WaveCluster:

1) складність реалізації;
2) алгоритм може виявляти кластери довільних форм;
3) алгоритм нечутливий до шумів;
4) алгоритм застосуємо тільки до даних низької розмірності.

Алгоритм CLARA (Clustering LARge Applications). Цей алгоритм був розроблений Kaufmann і Rousseeuw 1990 року для кластеризації даних у великих базах даних. Він будується в статистичних аналітичних пакетах, наприклад таких як S +.

Алгоритм CLARA витягує безліч зразків з бази даних. Кластеризація застосовується до кожного зі зразків, на виході алгоритму пропонується найкраща кластеризація.

Для великих баз даних цей алгоритм ефективніше, ніж алгоритм PAM. Ефективність алгоритму залежить від обраного зразка набору даних. Хороша кластеризація на обраному наборі може не дати хорошу кластеризацію на всій безлічі даних.

Алгоритми Clarans, CURE, DBScan. Алгоритм Clarans формулює завдання кластеризації як випадковий пошук у графі. У результаті роботи цього алгоритму сукупність вузлів графа являє собою розбиття множини даних на кількість кластерів, визначене користувачем. «Якість» отриманих кластерів визначається за допомогою критеріальної функції. Алгоритм Clarans сортує всі можливі розбиття множини даних у пошуках прийнятного рішення. Пошук рішення зупиняється в тому вузлі, де досягається мінімум серед визначеної кількості локальних мінімумів.

Серед нових масштабованих алгоритмів також можна відмітити алгоритм CURE – алгоритм ієрархічної кластеризації, і алгоритм DBScan, де поняття кластера формулюється з використанням концепції щільності.

Основним недоліком алгоритмів BIRCH, Clarans, CURE, DBScan є та обставина, що вони вимагають завдання деяких порогів щільності точок, а це не завжди прийнятно. Ці обмеження обумовлені тим, що описані алгоритми орієнтовані на надвеликі бази даних і не можуть користуватися значними обчислювальними ресурсами.

Над масштабованими методами нині активно працюють багато дослідників, основне завдання яких – подолати недоліки алгоритмів, існуючих на сьогодні.

Запитання до параграфа 15

1. Яка сутність загальної задачі кластеризації?
2. Яка мета кластеризації?
3. Які є характеристики кластерів?
4. Яка різниця між класифікацією і кластеризацією?
5. Що таке чіткі і нечіткі кластери?
6. Які кроки кластеризації?
7. Які є методи кластеризації?
8. Як здійснюється оцінка якості кластеризації?
9. Коли застосовують кластерний аналіз?
10. У які групи можна об'єднати задачі кластерного аналізу?
11. Які є методи визначення подібності кластерів?
12. Які математичні характеристики має кластер?
13. На які припущення спирається кластерний аналіз?
14. Для чого проводять стандартизацію змінних?
15. Яка сутність ієрархічних методів кластеризації?
16. Що таке дендограма?
17. Що таке агломеративний метод кластеризації?
18. Що таке дівізімний метод кластеризації?
19. Які є методи обчислення відстаней між об'єктами?
20. Що таке ітеративні методи?
21. Який алгоритм методу k-середніх?
22. Які переваги алгоритму k-середніх?
23. Яка мета факторного аналізу?
24. Що таке факторний аналіз?

§ 16. Ансамблі моделей

16.1. Введення в ансамблі моделей

У процесі створення алгоритмів машинного навчання розробники стикаються з такими проблемами, як обчислювальні витрати на реалізацію алгоритму, прозорість побудованих моделей для користувача, а також точність результатів. Більшість дослідників зосереджуються на підвищенні точності класифікації і передбаченні, тому продуктивність нових систем часто розглядається саме з цієї точки зору, оскільки точність відіграє вирішальну роль у всіх додатках машинного навчання і може бути легко оцінена, тоді як прозорість для користувача суб'єктивна.

В останні роки зростає інтерес до питання збільшення точності моделей Data Mining, заснованих на навчанні, за рахунок створення й агрегування набору класифікаторів. У результаті з'явилися нові підходи до аналізу на базі ансамблів класифікаторів, які застосовані до широкого кола систем машинного навчання і засновані на теоретичному дослідженні поведінки складових класифікаторів.

Приймаючи рішення, досвідчений керівник не тільки покладається на власні знання та інтуїцію, але і намагається залучити експертів у конкретних предметних сферах. Вважається, що висновки експертів, отримані на основі аналізу даних, пов'язаних з розв'язуваним завданням, дозволять зробити оптимальний вибір.

Наприклад, коли підприємство планує випуск нового продукту, зважуються всі «за» і «проти», залучаються фахівці у сферах економіки, маркетингу, виробництва, реклами тощо, і кожен з них висловлює свою думку. Вислухавши всіх, керівник приймає остаточне рішення. Однак висновки, зроблені різними експертами, можуть суперечити один одному. Тому неминуче постає питання про комбінацію експертних оцінок, щоб на їх основі прийняти правильне рішення.

Існують методи, які дозволяють ухвалити більш обґрунтоване рішення. Припустимо, що як оцінку певного рішення експерт вказує число балів: 100 балів відповідають позитивному вирішенню

(«випускати продукт»), 0 балів – негативному («не випускати продукт»). Вказуючи проміжні значення, експерти оцінюють імовірність успіху або провалу проєкту. Отримані оцінки підсумовуються, і якщо сума перевищить заданий поріг, то рішення буде позитивним, а якщо ні – негативним. За бажанням можна зробити усереднення оцінок, тобто розділити суму балів на кількість експертів.

Хороші результати часто дає зважування оцінок: залежно від важливості предметної сфери, професійного рівня експерта тощо. Експертній оцінці присвоюється певне вагове значення. Чим досвідченіший експерт і важливіша оцінка, тим більша вага.

Аналогічна ситуація складається під час використання моделей, заснованих на машинному навчанні, – дерев рішень, нейронних мереж тощо. Ці моделі фактично відіграють роль експертів і в явному або неявному вигляді надають інформацію, необхідну для обґрунтованого прийняття рішень. Можна обмежитися результатами, отриманими єдиною моделлю. Але так само як і експерти, моделі бувають слабкими і сильними. Якщо навченій моделі добре вдається розділити класи і вона допускає мало помилок класифікації, то така модель може розглядатися як сильна. Слабка модель, навпаки, не дозволяє надійно розділяти класи або давати точні прогнози, допускає в роботі велику кількість помилок.

Якщо в результаті навчання отримали слабку модель, то її необхідно вдосконалити, підбираючи тип класифікатора, алгоритми і параметри навчання. Але часто трапляються ситуації, коли всі можливості вдосконалення єдиної моделі вичерпані, а якість її роботи, як і раніше, незадовільна. Це може бути пов’язано зі складністю розв’язуваної задачі, низькою якістю навчальних даних та іншими факторами.

Неминуче виникає питання про посилення слабкої моделі і що необхідно зробити для підвищення її ефективності. Цілком логічним виходом із ситуації є спроба застосувати до невдалих результатів роботи першої моделі ще одну модель, завдання якої – класифікувати ті приклади, що залишилися нерозпізнаними. Якщо і після цього результати незадовільні, можна застосувати третю модель, і так далі доти, доки не буде отримано достатньо точне рішення.

Таким чином, для вирішення однієї задачі класифікації або регресії можемо застосувати кілька моделей, при цьому нас цікавить не результат роботи кожної окремої моделі, а результат, який дає весь набір моделей. Такі сукупності моделей називаються ансамблями моделей.

Набір моделей, що застосовуються спільно для вирішення єдиного завдання, називається ансамблем моделей.

Мета об'єднання моделей – поліпшити (посилити) рішення, яке дає окрема модель. При цьому передбачається, що єдина модель ніколи не зможе досягти тієї ефективності, яку забезпечить ансамбль.

Використання ансамблів замість окремої моделі здебільшого дозволяє підвищити якість рішень, однак такий підхід пов'язаний з рядом проблем, основними з яких є:

– збільшення тимчасових і обчислювальних витрат на навчання декількох моделей;

– складність інтерпретації результатів;

– неоднозначний вибір методів комбінування результатів, що видаються окремими моделями.

Перераховані проблеми аналогічні тим, що виникають під час роботи декількох людей-експертів.

В останнє десятиліття ансамблі моделей стали сферою досить активних досліджень у машинному навчанні, що призвело до розробки великої кількості різноманітних методів формування ансамблів.

Першим питанням у процесі формування ансамблю є вибір базової моделі. Ансамбль у цілому може розглядатися як складна модель, що складається з окремих (базових) моделей. Тут можливі два випадки.

1. Ансамбль складається з базових моделей одного типу, наприклад тільки з дерев рішень, нейронних мереж тощо.

2. Ансамбль складається з моделей різного типу – нейронних мереж, дерев рішень, регресійних моделей тощо.

Кожен підхід має свої переваги і недоліки. Використання моделей різних типів дає класифікаторам додаткову гнучкість. Але оскільки вихід однієї моделі застосовується для формування

навчальної множини іншої, можливо, будуть потрібні додаткові перетворення, щоб узгодити входи і виходи моделей.

Друге питання – це як використовувати навчальну множину під час побудови ансамблю. Тут також існують два підходи.

1. Перевиборка. З вихідної навчальної множини витягується кілька підвибірок, кожна з яких використовується для навчання однієї з моделей ансамблю. Якщо ансамбль будується на основі моделей різних типів, то для кожного типу буде свій алгоритм навчання.

2. Використання однієї навчальної множини для навчання всіх моделей ансамблю.

Третє питання стосується методу комбінування результатів, виданих окремими моделями, а саме що буде вважатися виходом ансамблю за певних станів виходів моделей. Зазвичай використовуються такі способи комбінування:

1. Голосування. Застосовується в задачах класифікації, тобто для категоріальної цільової змінної. Вибирається той клас, який був виданий простою більшістю моделей ансамблю.

2. Виважене голосування. В ансамблі одні моделі можуть працювати краще, а інші – гірше. Відповідно, до результатів одних моделей довіри більше, а до результатів інших – менше. Щоб врахувати рівень достовірності результатів, для моделей ансамблю можуть бути призначені ваги (бали).

3. Усереднення (зважене або незважене). Якщо за допомогою ансамблю вирішується завдання регресії, то виходи його моделей будуть числовими. Вихід усього ансамблю може визначатися як просте середнє значення виходів усіх моделей.

Дослідження ансамблів моделей у Data Mining стали проводитися порівняно недавно. Проте донині розроблено безліч різних методів і алгоритмів формування ансамблів. Серед них найбільшого поширення набули такі методи, як бегінг, бустінг і стекінг.

16.2. Бегінг

Поряд з бустінгом бегінг є одним з найсучасніших методів підвищення ефективності моделей. Бегінг формує набір класифікаторів, які комбінуються шляхом голосування або усереднення. Термін «бегінг» (bagging) походить від англійського словосполучення bootstrap aggregating, яке можна перекласти як «поліпшує об'єднання». В основі роботи беггінга полягає технологія, що отримала назву «збурення і комбінування» (perturb and combine).

«Збурення і комбінування». Моделі, які в процесі навчання адаптуються відповідно до навчальних множин, такі як дерева рішень і нейронні мережі, нестійкі – навіть невеликі зміни в навчальній множині (заміна або видалення одного прикладу) можуть призвести до істотних змін у стані моделі – у структурі дерева рішень або в розподілі ваг нейронної мережі. Іншими словами, вносячи навіть незначні зміни в навчальні дані, завжди будемо отримувати іншу модель. Але вихідна і змінена моделі функціонуватимуть приблизно однаково і з порівняльною точністю: незначні зміни в навчальних даних не приведуть до зміни основних закономірностей, які повинні бути виявлені моделлю.

Нестійкість дерев рішень багато в чому обумовлена конкуруючими вузлами, тобто вузлами, що працюють приблизно однаково. Тому навіть невелика зміна в даних може призвести до того, що процес навчання піде іншим вузлом і буде побудовано інше дерево рішень.

Нестійкість моделей, особливо дерев рішень, використовується для створення ансамблів моделей за допомогою технології «збурення і комбінування». Під збуренням розуміється внесення деяких змін, часто випадкового характеру, в навчальні дані і побудова декількох альтернативних моделей на змінених даних з подальшим комбінуванням результатів. Для збурення використовуються декілька прийомів.

1. Формування вибірок з навчальної множини. У цьому випадку шляхом семплінгу з вихідної навчальної множини формується кілька вибірок, і на кожній з них навчається окрема модель.

2. Вибірка з вибірок, формування всередині вибірок підгруп.

3. Додавання шуму.

4. Адаптивне зважування.

5. Випадковий вибір між конкуруючими вузлами (розбивками).

Додавання у процес навчання елемента випадковості часто називають рандомізацією.

Ідея бегінга проста. Спочатку на основі вихідної вибірки даних шляхом випадкового відбору формується кілька інших вибірок. Вони містять таку саму кількість прикладів, що і вихідна. Але оскільки відбір проводиться випадково, набір прикладів у цих вибірках буде різним: одні приклади можуть бути відібрані по кілька разів, а інші – жодного разу. Потім на основі кожної вибірки будується класифікатор і виходи всіх класифікаторів комбінуються (агрегуються) шляхом голосування або простого усереднення. Очікується, що отриманий результат буде набагато точнішим, ніж ті, які дають окремі моделі, побудовані на основі вихідного набору даних.

Схема бегінга застосовується і до регресійних моделей. У цьому випадку результат – середнє значення, розраховане за виходами всіх моделей ансамблю.

Бегінг, хоча і меншою мірою, ніж окремі моделі, схильний до перенавчання. Тому критерієм для зупинення процедури може слугувати зростання помилки на тестовій множині.

Результати бегінга часто дивують аналітиків: точність передбачення побудованих за його допомогою комбінованих класифікаторів виявляється значно вища, ніж точність окремих моделей. Причини цього неочевидні. Дійсно, якщо всі вибірки витягнуті з однієї навчальної множини, то можна очікувати, що побудовані на їх основі моделі будуть практично ідентичними. Однак це припущення помилково. Причина полягає в тій самій нестійкості моделей, особливо дерев рішень: навіть невеликі зміни в навчальних даних легко можуть призвести до вибору різних атрибутів у певному вузлі дерева, що викличе істотні зміни під час подальшого розгалуження. У результаті одні й ті ж тестові приклади будуть розпізнаватися різними деревами по-різному.

В основі ефективності методів комбінування моделей, зокрема бегінга, полягає ідея декомпозиції помилки ансамблю на

зміщення і дисперсію. Припустимо, що є нескінченна кількість незалежних навчальних множин однакового розміру, які використовуються для побудови нескінченної кількості класифікаторів. Навчальні приклади обробляються всіма класифікаторами, і їх загальний вихід визначається більшістю голосів. У цій ситуації помилки все ще мають місце, оскільки ідеальних навчальних процедур не існує. Значення помилки залежить від того, наскільки добре метод машинного навчання відповідає розв'язуваній задачі, а також від якості самих даних.

Припустимо, очікувана помилка оцінюється як середня помилка комбінованого класифікатора на нескінченній кількості незалежно обраних тестових прикладів. Значення помилки конкретного навчального алгоритму називається зміщенням (систематичною помилкою). Зсув є мірою стабільності помилки цього алгоритму навчання, яка не може бути виключена навіть застосуванням нескінченної кількості навчальних множин. Звичайно, на практиці така помилка не може бути обчислена точно, а тільки оцінюється приблизно.

Ще одним джерелом помилки під час навчання моделі є те, що навчальна множина використовується частково і, отже, не повністю представляє реальну сукупність спостережень. Очікуване значення цього компонента помилки з усіх можливих навчальних вибірок заданого розміру і всіх можливих тестових вибірок називається дисперсією методу навчання для цього завдання.

Таким чином, загальна очікувана помилка класифікатора складається із суми дисперсії і зміщення. У цьому і полягає сенс декомпозиції помилки на дисперсію і зміщення. Комбінування декількох класифікаторів дозволяє зменшити помилку за рахунок дисперсії. При цьому чим більше класифікаторів використовується, тим менша дисперсія.

Зсув визначається як середньоквадратична помилка, очікувана у разі усереднення за моделями, а дисперсія – це компонент очікуваної помилки окремої моделі, яка була побудована за допомогою окремої навчальної множини.

На практиці виникає проблема: де взяти велику кількість навчальних множин. Бегінг вирішує цю задачу, використовуючи єдину навчальну вибірку, а саме «тасує» її як колоду карт.

Різниця між бегінгом та ідеальною процедурою, описаною вище, полягає у способі формування навчальних множин. Замість отримання незалежних множин з предметної сфери бегінг просто «виробляє» нову вибірку з наявної. Такі вибірки відрізняються одна від одної, але не є незалежними, оскільки всі вони засновані на одній. Проте бегінг дозволяє створювати комбіновані моделі, які, як правило, працюють значно краще, ніж окрема модель.

16.3. Бустінг

Порівняно з бегінгом бустінг (boosting) дещо складніша процедура, але в багатьох випадках працює ефективніше. Як і бегінг, бустінг використовує нестійкість алгоритмів навчання і починає створення ансамблю на основі єдиної вихідної вибірки. Але якщо в бегінгу моделі будуються паралельно і незалежно одна від одної, то в бустінгу кожна нова модель будується на основі результатів раніше побудованих моделей, тобто моделі створюються послідовно. Бустінг створює нові моделі таким чином, щоб вони доповнювали раніше побудовані, виконували ту роботу, яку інші моделі зробити не змогли на попередніх кроках. І нарешті, остання відмінність бустінга від бегінга полягає в тому, що всім побудованим моделям залежно від їх точності присвоюються ваги (бегінг використовує зважене голосування або усереднення).

Сьогодні розроблено велику кількість різних модифікацій бустінга. Один з найбільш популярних алгоритмів – AdaBoost.M1, який призначений для вирішення задач класифікації.

Замість вилучення вибірок з вихідної множини даних бустінг як «збурюючий» фактор застосовує зважування прикладів. Вага кожного прикладу встановлюється відповідно до його впливу на навчання класифікатора. На кожній ітерації вектор ваг підлаштовується таким чином, щоб відображати ефективність цього класифікатора. У результаті вага неправильно класифікованих прикладів збільшується. Підсумковий класифікатор також агрегує навчені класифікатори шляхом голосування, але тепер голос класифікатора є функцією його точності.

Позначимо вагу прикладу x на ітерації t як w'_x, при цьому вага на першій ітерації задається як $w_x^1 = 1/N$ для кожного x (N – кількість прикладів). На кожній ітерації t = 1, 2 ... T класифікатор C створюється з цих прикладів відповідно до розподілу їх ваг. Помилка e^t класифікатора t також вимірюється щодо ваг і являє собою суму ваг прикладів, які були класифіковані неправильно. Коли e^t стає більше 0,5, ітерації припиняються, останній класифікатор видаляється і T змінюється на t – 1. Навпаки, якщо $e^t = 0$ (класифікатор правильно класифікував усі приклади), ітерації зупиняються і T = t. В інших випадках вектор ваг w^{t+1} для наступної ітерації генерується шляхом множення поточних ваг прикладів, які були правильно розпізнані класифікатором, на коефіцієнт $\beta^t = \frac{e^t}{(1-e^t)}$, а потім нормується так, щоб $\sum_x w_x^{t+1} = 1$. Тоді:

$$w^{t+1} = w^t \cdot \frac{e^t}{(1-e^t)}.$$

Підсумковий класифікатор виходить шляхом підсумовування голосів інших класифікаторів, де голос класифікатора визначається як $\log(1 / \beta^t)$ одиниць.

Було показано, що якщо помилка окремого класифікатора завжди менше 0,5, то значення помилки підсумкового класифікатора експоненціально прагне до 0 зі збільшенням кількості ітерацій t. Послідовність слабких класифікаторів може бути посилена до класифікатора C, який зазвичай виходить більш точним, ніж окремі класифікатори. Звичайно, при цьому не можна гарантувати високу узагальнюючу здатність C.

Процедура бустінга включає такі кроки:

1. Для всіх прикладів вихідної множини даних встановлюються рівні початкові ваги W_q.

2. На основі зваженого набору прикладів будується класифікатор C^t, обчислюється і запам'ятовується вихідна помилка цього класифікатора.

3. Розраховується корекція ваг прикладів навчальної множини і ваги коректуються.

4. Якщо помилка $e^t = 0$ або $e^t > 0,5$, то класифікатор видаляється і процедура бустінга зупиняється.

5. В іншому випадку здійснюється перехід на крок 2 і починається наступна ітерація.

Таким чином, параметрами, налаштованими на кожній ітерації, є ваги прикладів. При цьому чим більше разів приклад був неправильно розпізнаний попередніми моделями, тим вище його вага. Вагу можна розглядати як імовірність попадання прикладу на наступну ітерацію.

У побудові першої моделі братимуть участь усі приклади, в побудові другої – тільки ті, які були неправильно розпізнані першою, в побудові третьої – приклади, неправильно розпізнані двома попередніми моделями тощо. Моделі будуть доповнювати одна одну – працювати з тими прикладами, від яких «відмовилися» інші моделі. З теорії машинного навчання відомо, що саме важкі для розпізнавання приклади штовхають процес навчання вперед, змушуючи модель класифікувати їх, тоді як прості приклади розпізнаються швидко і для подальшого тренування не приносять користі.

Ваги прикладів збільшуються або зменшуються залежно від виходу кожного нового класифікатора. У результаті деякі важкі приклади можуть стати ще важчими, а легкі – ще легшими. Після кожної ітерації ваги відображають, наскільки часто певний приклад класифікувався неправильно.

Вихід ансамблю комбінується процедурою зваженого голосування. Під час визначення ваг класифікатор, який добре працює на навчальній вибірці, повинен отримати більшу вагу. Класифікатор, який працює гірше (помилка ближче до 0,5), отримує більш низьку вагу. Вага – це позитивне число в інтервалі від нуля до нескінченності. Крім того, формула пояснює, чому класифікатори з нульовою помилкою повинні бути видалені: при $e^t = 0$ вираз для ваг виявиться невизначеним через те, що аргумент логарифма дорівнює 1. У разі комбінування виходів ансамблю ваги всіх класифікаторів, які «голосують» за певний клас (тобто мають цей клас на виході), підсумовуються і вибирається клас, який набрав найбільшу суму голосів. Іншими словами, визначальну роль відіграють не стільки виходи моделей, скільки їх ваги.

Основним недоліком бустінга є те, що приклади з низькими вагами не потрапляють на наступну ітерацію, тому частина інформації, корисної для навчання, може бути втрачена. Крім того, у процесі бустінга з участі в навчанні поступово виключаються найпростіші приклади, що призводить до так званого виродження навчальних вибірок, на основі яких будуються останні моделі. У першу чергу, розпізнаються найбільш типові приклади, що добре відображають предметну сферу. Зразки, важкі для розпізнавання, часто є нетиповими, аномальними тощо. Навчання класифікаторів на пізніших ітераціях може майже повністю вестися на таких прикладах, що в кінцевому підсумку призводить до перенавчання. Враховуючи вище викладене, бустінг більш схильний до перенавчання, ніж бегінг.

Щоб уникнути перенавчання, кількість ітерацій має бути настільки малою, наскільки це можливо без втрати точності. Алгоритм зупиняється, коли на деякому C помилка наближається до 0.

16.4. Альтернативні методи побудови ансамблів

Застосування ансамблів моделей до вирішення різних завдань аналізу відкриває широкі можливості для підвищення ефективності моделей Data Mining. Тому в останні кілька років у цій сфері проводяться активні дослідження, в результаті чого з'явилася велика кількість методів.

Не всі методи працюють однаково добре в різних ситуаціях, а можливість вибору дозволяє аналітику отримати найкращі результати. Наприклад, бустінг і бегінг спочатку виникли як методи підвищення ефективності класифікаційних моделей, особливо дерев рішень, і не могли безпосередньо використовуватися для регресійних моделей. Тому для вирішення завдань регресії були розроблені варіанти бустінга, що отримали назву адитивної регресії, зокрема алгоритм LogitBoost. Поширені і такі методи комбінування моделей, як дерева вибору і стекінг.

Адитивна регресія. Одним з варіантів бустінга стала побудова так званих адитивних регресійних моделей. Поняття «адитивна регресія» належить до будь-яких методів передбачення, заснованих на комбінуванні (підсумовуванні) вкладів від декількох регресійних моделей.

В основі ідеї адитивної регресії полягає пряме ступінчасте адитивне моделювання (forward stagewise additive modeling), алгоритм якого містить такі кроки:

1. Спочатку будується звичайна регресійна модель, наприклад регресійне дерево рішень.

2. Обчислюються помилки, отримані на навчальній множині, як різниця між бажаним і спостережуваним значеннями. Ці помилки називаються залишками.

3. Проводиться мінімізація залишків за допомогою другої моделі (іншого дерева рішень), яка намагається передбачити наявні залишки. Для цього вихідні цільові значення замінюються відповідними залишками перед навчанням другої моделі.

4. Оскільки і друга модель не є ідеальною, вона також дасть залишки на деяких прикладах. Тому процес продовжиться: будується третя модель, яка навчається прогнозувати залишки від залишків та ін.

Якщо кожна модель мінімізує помилку передбачення, як здійснює звичайна лінійна регресійна модель, то алгоритм прямого ступінчатого моделювання мінімізує помилку ансамблю в цілому. Фактично під час використання звичайної лінійної регресії як базової моделі для адитивної регресії сума лінійних регресійних моделей також буде лінійною регресійною моделлю, і регресійний алгоритм сам мінімізує квадратичну помилку.

Для зручності створення ступінчастої адитивної регресії зазвичай починається з побудови моделі нульового рівня, яка просто прогнозує середнє значення виходу на навчальній множині, а кожна наступна модель підлаштовує залишки.

Адитивна регресія схильна до перенавчання, оскільки кожна модель, додана в ансамбль, робить його настройку все більш точною. Щоб визначити момент зупинки, використовується помилка на тестовій множині і вибирається те число ітерацій, яке її мінімізує.

Адитивна регресія може застосовуватися для вирішення завдань бінарної класифікації. Але оскільки для цього більше підходить логістична регресія, адитивна модель повинна бути відповідним чином модифікована. За допомогою логіт-перетворення проводиться оцінка ймовірності кожного класу, а потім завдання

регресії вирішується з використанням ансамблю моделей, наприклад регресійних дерев, так само як і у випадку адитивної регресії. На кожному кроці додається модель, яка максимізує ймовірність одного з класів.

Дерева вибору. До недоліків ансамблів моделей як складних складових структур належать труднощі інтерпретації результатів. Вирішити цю проблему можна, зробивши подання ансамблю більш компактним. Одним з варіантів є використання структур, званих деревами вибору (option tree). Вони відрізняються від звичайних дерев рішень тим, що містять два типи вузлів: рішень і вибору. У вузол рішень, як і в звичайному дереві рішень, входить тільки одна гілка, а у вузол вибору – дві. Таким чином, у вузол вибору надходять рішення з двох попередніх вузлів. Тоді вихід вузла вибору – це комбінація рішень, які в нього надійшли. Вибір може бути зроблений звичайним голосуванням.

Коли вузол вибору містить два входи і більше, результат залежатиме від того, чи є вони категоріальними або числовими. Якщо входи категоріальні, то використовується голосування, а якщо числові – отримані різними способами імовірнісні оцінки, зважене середнє або байєсовський підхід.

Дерева вибору можна створювати за допомогою модифікації звичайних дерев рішень шляхом вбудовування в них вузлів вибору.

Інша можливість побудувати дерево вибору – покроково додавати в нього вузли. При цьому використовується алгоритм бустінга. Отримане дерево називається знакозмінним деревом, вузли рішень – вузлами-розгалудження, а вузли вибору – вузлами передбачення. Вузли передбачення є листям, якщо з них більше не виходять вузли-розгалужувачі. Звичайні знакозмінні дерева застосовуються під час вирішення завдань бінарної класифікації, і з кожним вузлом передбачення асоційоване позитивне або негативне числове значення. Щоб передбачити клас певного прикладу, його пропускають через усі можливі гілки, після чого значення, отримані на виходах усіх вузлів передбачення, підсумовуються. Клас вибирається залежно від того, позитивною або негативною виявилася сума.

16.5. Стекінг

Стекінг (stacking) – один із способів створення складових моделей. Хоча метод був розроблений кілька років потому, він усе ще менш відомий, ніж бегінг і бустінг. Частково це пов'язано зі складністю теоретичного аналізу, а частково з тим, що загальна концепція використання цього методу досі відсутня – основна ідея може застосовуватися в найрізноманітніших варіантах.

На відміну від бегінга і бустінга стекінг зазвичай не використовується для комбінування моделей одного і того ж типу, наприклад набору дерев рішень. Замість цього він застосовується до моделей, побудованих за допомогою різних алгоритмів навчання. Припустимо, маємо три моделі – дерево рішень, байєсовський класифікатор і нейронну мережу, на основі яких потрібно побудувати класифікатор для певного набору даних. Звичайна процедура полягає в оцінці помилки кожного алгоритму і виборі кращого з них, який і буде використовуватися для класифікації нових даних.

Однією з технік комбінування виходів є голосування. Але воно має сенс тільки тоді, коли навчальні схеми працюють порівняно добре. Якщо два з трьох класифікаторів працюють погано, то очікувати хороших результатів від усього ансамблю не доводиться. Замість цього стекінг вводить концепцію метанавчання, яка замінює процедуру голосування. Проблема голосування полягає в тому, що незрозуміло, яким класифікаторам варто довіряти. Стекінг намагається навчити кожен класифікатор, використовуючи алгоритм метанавчання, який дозволяє виявити кращу комбінацію виходів базових моделей.

Базові моделі утворюють рівень 0. На вхід метамоделі, яка також називається моделлю рівня 1, подаються результати з виходів базових моделей. Приклад рівня 1 має стільки атрибутів, скільки існує моделей рівня 0, а самі значення атрибутів – виходи моделей нульового рівня. Потім на основі результатів, отриманих моделлю рівня 1, будується модель рівня 2 тощо, доки не буде виконана будь-яка з умов зупинки навчання.

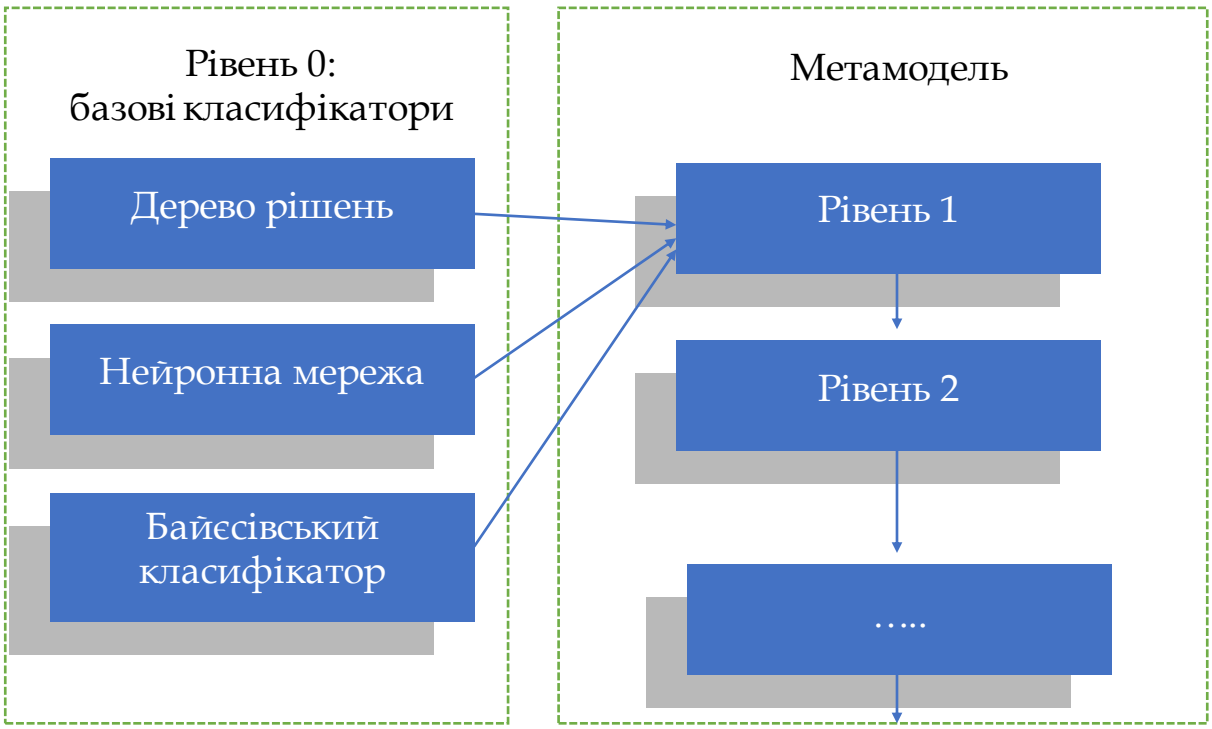


Рис. 3.15. Стекінг

Запитання до параграфа 16

1. Що таке ансамблі моделей?
2. Яка мета побудови ансамблів моделей?
3. Які задачі вирішує побудова ансамблів моделей?
4. Які групи питань необхідно вирішити перед побудовою ансамблю моделей?
5. Що таке бегінг?
6. Яка мета використання бегінга?
7. Яка ідея бегінга?
8. Що таке бустінг?
9. Яка мета використання бустінга?
10. Які особливості бустінга?
11. Які кроки включає процедура бустінга?
12. Які недоліки бустінга?
13. Які існують альтернативні методи побудови ансамблів?
14. Що таке стекінг?
15. Які особливості стекінга?
16. Яка загальна схема використання стекінга?

Тести до розділу 3

1. Кластеризація – це:

а) розбивка множини об’єктів на групи за певною ознакою;

б) розбивка множини об’єктів на групи за подібністю;

в) визначення властивостей певного кластера;

г) правильної відповіді немає.

2. Класифікація – це:

а) розбивка множини об’єктів на групи за певною ознакою;

б) розбивка множини об’єктів на групи за подібністю;

в) визначення властивостей певного кластера;

г) правильної відповіді немає.

3. Підтримкою називають:

а) відношення кількості транзакцій, в які входить набір F, до загальної кількості транзакцій;

б) впорядковану множину об’єктів;

в) встановлення відносин порядку між об’єктами множини;

г) сукупність певних товарів, яку купують разом у певний момент часу.

4. Послідовність – це:

а) впорядкована множина об’єктів за часом;

б) сукупність певних товарів, які купують разом у певний момент часу;

в) послідовність покупців, які купують один товар;

г) немає правильної відповіді.

5. Для розширення можливостей аналізу за допомогою пошуку асоціативних правил у досліджувані набори:

а) можна додавати додаткові об’єкти;

б) не можна додавати додаткові об’єкти;

в) правильної відповіді немає.

6. Асоціативні правила бувають:

а) корисні;

б) тривіальні;

в) незрозумілі;

г) усі відповіді правильні.

7. Підтримка показує:

а) який відсоток транзакцій підтримує це правило;

б) ймовірність того, що за наявності в транзакції набору X випливає наявність у ній набору Y;

в) чи корисне правило випадкового вгадування;

г) усі відповіді правильні.

8. Достовірність показує:

а) який відсоток транзакцій підтримує це правило;

б) ймовірність того, що за наявності в транзакції набору X випливає наявність у ній набору Y;

в) чи корисне правило випадкового вгадування;

г) усі відповіді правильні.

9. Поліпшення показує:

а) який відсоток транзакцій підтримує це правило;

б) ймовірність того, що за наявності в транзакції набору X випливає наявність у ній набору Y;

в) чи корисне правило випадкового вгадування;

г) усі відповіді правильні.

10. Алгоритм Аргіогі застосовується для:

а) виявлення частих наборів об’єктів;

б) кластеризації;

в) групування;

г) побудови регресії.

11. Дерево рішень – це:

а) спосіб представлення правил в ієрархічній, послідовній структурі;

б) один із методів класифікації;

в) метод, в основі якого полягає процес рекурсивного розбиття вихідної множини спостережень або об'єктів на підмножини;

г) усі відповіді правильні.

12. Атрибутами в теорії дерев рішень називають:

а) ознаки, що описують об'єкти, які класифікуються;

б) кінцевий вузол дерева, вузол рішення;

в) залежна змінна, цільова змінна, вихідне поле;

г) шаблон, спостереження, запис.

13. Листом у теорії дерев рішень називають:

а) шаблон, спостереження, запис;

б) кінцевий вузол дерева, вузол рішення;

в) залежна змінна, цільова змінна, вихідне поле;

г) ознака, незалежна змінна, властивість, вхідне поле.

14. Міткою класу в теорії дерев рішень є:

а) шаблон, спостереження, запис;

б) кінцевий вузол дерева, вузол рішення;

в) залежна змінна, цільова змінна, вихідне поле;

г) ознака, незалежна змінна, властивість, вхідне поле.

15. Об'єктом у теорії дерев рішень є:

а) шаблон, спостереження, запис;

б) кінцевий вузол дерева, вузол рішення;

в) залежна змінна, цільова змінна, вихідне поле;

г) ознака, незалежна змінна, властивість, вхідне поле.

16. До варіантів зупинки побудови дерева рішень відносять:

а) «ранню зупинку»;

б) обмеження глибини дерева;

в) визначення мінімальної кількості прикладів, які будуть міститися в кінцевих вузлах дерева;

г) усі відповіді правильні.

17. Якість класифікаційної моделі, побудованої за допомогою дерева рішень, характеризується:

а) точністю розпізнавання;

б) помилкою;

в) усі відповіді правильні.

18. Алгоритми, за допомогою яких будують дерева рішень:

а) CART, C4.5;

б) CHAID, CN2;

в) NewId, ITrule;

г) усі відповіді правильні.

19. Алгоритми побудови дерев рішень відрізняються такими характеристиками:

а) вид розщеплення;

б) критерії розщеплення;

в) можливість обробки пропущених значень;

г) усі відповіді правильні.

20. Емпіричний ризик – це:

а) рівень помилки класифікації на навчальній виборці;

б) ймовірність настання несприятливої події;

в) неможливість отримати правильний результат;

г) немає правильної відповіді.

21. Метод опорних векторів дозволяє:

а) отримати функцію класифікації з мінімальною верхньою оцінкою очікуваного ризику (рівня помилки класифікації);

б) використовувати лінійний класифікатор для роботи з нелінійно подільними даними, поєднуючи простоту з ефективністю;

в) дві відповіді правильні;

г) немає правильних відповідей.

22. Прецедент – це:

а) опис ситуації в поєднанні з детальним зазначенням дій, що вживаються в цій ситуації;

б) адаптація обраного рішення до поточної проблеми, якщо це необхідно;

в) перевірка коректності кожного знову отриманого рішення;

г) занесення детальної інформації про новий прецедент у базу прецедентів.

23. Завдання Data Mining, що вирішується за допомогою нейронних мереж:

а) класифікація (навчання з учителем);

б) прогнозування;

в) кластеризація (навчання без вчителя);

г) усі відповідні правильні.

24. Кластеризація:

а) призначена для розбиття сукупності об’єктів на однорідні групи;

б) є описовою процедурою, вона не робить ніяких статистичних висновків;

в) це метод пошуку існуючих структур;

г) усі відповіді правильні.

25. Основна відмінність задачі класифікації від регресії полягає:

а) у тому, що у першому випадку залежна змінна є дискретною, категоріальною, в другому – неперервною;

б) у тому, що у першому випадку залежна змінна є неперервною, в другому – дискретною;

в) правильної відповіді немає;

г) у тому, що у першому випадку залежна змінна не є дискретною, категоріальною, в другому – неперервною.

26. Метою асоціативних правил є:

а) залежності, подані у вигляді правил;

б) знаходження об'єктів або подій;

в) знаходження залежностей між об'єктами або подіями;

г) усі відповіді правильні.

27. Аналіз зв'язків – це:

а) візуалізації, яка створюється графічним образом аналізованих даних;

б) задача знаходження залежностей у наборі даних;

в) опис конкретних груп об'єктів з аналізованого набору даних;

г) власне застосування методів Data Mining.

28. Основною відмінністю завдання сіквенціального аналізу є:

а) окремий випадок задачі класифікації;

б) дійсні числа;

в) усі відповіді правильні;

г) встановлення відносин порядку між досліджуваними наборами.

29. Кластеризація може застосовуватися практично в будь-якій сфері, де:

а) необхідно підходити до роботи із споживачами;

б) необхідні дослідження експериментальних або статистичних даних;

в) потрібно провести маркетингове дослідження;

г) правильної відповіді немає.

30. Які існують види моделей у Data Mining?

а) регресивні і підсумкові;

б) досліджувані й усні;

в) прогностичні й описові;

г) розрахункові й описові.

31. Яка мета технології Data Mining?

а) знаходження послідовних шаблонів;

б) знаходження тимчасових закономірностей між транзакціями;

в) виявлення та аналіз даних;

г) знаходження в даних таких моделей, які не можуть бути знайдені звичайними методами.

32. Концептуально сегментування засноване на передумові:

а) що всі споживачі різні;

б) що всі споживачі однакові;

в) що у всіх споживачів подібні потреби;

г) усі відповіді правильні.

33. Для того щоб визначити, чим відрізняються споживачі один від одного:

а) проводиться опитування споживачів;

б) проводиться сегментація споживачів;

в) проводиться оцінка споживачів;

г) правильної відповіді немає.

34. Основні етапи аналізу:

а) розуміння і формулювання завдання аналізу;

б) підготовка даних для автоматизованого аналізу;

в) перевірка побудованих моделей;

г) усі відповіді правильні.

СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

1. Berry M. J. A. Data Mining Techniques for Marketing, Sales and Customer Relationship Management / M. J. A. Berry, G. S. Linoff. – Wiley Publishing, Inc., 2004.

2. Berthold M. (Eds.) Intelligent Data Analysis / Berthold M., HandD. J. – Berlin ; Heidelberg : Springer-Verlag, 2007.

3. Chiu S. Data Mining and Market Intelligence for Optimal Marketing Returns / S. Chiu, D. Tavella. – Elsevier Inc., 2003.

4. Data Mining for Business Applications / Edited by L. Cao, Philip S. Yu, C. Zhang, H. Zhang. – Springer Science; Business Media, 2008.

5. Davenport T. Analytics at Work: Smarter Decisions, Better Results / T. Davenport, J. Harris, R. Morrison. – Harvard Business School Press, 2010. – 225 p.

6. Drucker Peter F. The Effective Executive Publisher / Peter F. Drucker. – New York : HarperBusiness, 1993.

7. Hamerly G. Learning the K in K-Means / G. Hamerly, C. Elkan // Neural Information Processing Systems. – MIT Press, 2003.

8. Han J. Data Mining: Concepts and Techniques / J. Han, M. Kamber. – Elsevier Inc., 2006.

9. Hosmer D. W. Applied Logistic Regression (Second Edition) / D. W. Hosmer, S. Lemeshow. – Wiley Publishing, Inc., 2000.

10. [Електронний ресурс]. – Режим доступу : https://www.iiba.org/standards-and-resources/babok/

11. [Електронний ресурс]. – Режим доступу : https://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=2657-12

12. ILO Database (База даних МОП) [Електронний ресурс]. – Режим доступу : http://www.ilo.org/public/english/suppor

13. Inmon W. H. Building the Data Warehouse / W. H. Inmon. – Wiley Publishing, Inc., 2005.

14. Larose D. T. Data Mining Methods and Models / D. T. Larose. – Wiley Publishing, Inc., 2006.

15. Larose D. T. Discovering Knowledge in Data: An Introduction in Data Mining / D. T. Larose. – Wiley Publishing, Inc., 2005.

16. Laursen G. Business Analytics for Managers. Taking Business Intelligence Beyond Reporting / G. Laursen, J. Thorlund. – New Jersy : John Wiley and Sons, 2010. – 273 p.

17. Vercellis C. Business Intelligence: Data Mining and Optimization for Decision Making / C. Vercellis. – Wiley Publishing, Inc., 2009.

18. Wang L. Data Mining with Computational Intelligence / L. Wang, X. Fu. – Berlin; Heidelberg : Springer-Verlag, 2005.

19. Азовцева И. Бизнес-аналитика в банковской сфере / И. Азовцева // Банковские технологии. – 2012. – С. 58–60.

20. Аналитические технологии управления: Business Intelligence // Банковские технологии. – 2011. – № 4. – С. 10–11.

21. Аналітична і статистична інформація по розвиненим ринкам і тим, що розвиваються [Електронний ресурс]. – Режим доступу : www.damodaran.com

22. Управленческий учет / Э. А. Аткинсон, Р. Д. Банкер, Р. С. Каплан, М. С. Янг ; пер. с англ. А. Д. Рахубовского, Д. А. Рахубовской ; под ред. А. Д. Рахубовского. – 3-е изд. – М. : Вильямс, 2003. – 878 с.

23. Бариленко В. И. Подготовка бизнес-аналитиков / В. И. Бариленко // Экономический анализ: теория и практика. – 2011. – № 33. – С. 42–47.

24. Бизнес-анализ для современной компании // Банковские технологии. – 2011. – № 5. – С. 22.

25. Бизнес-аналитика в банке: текущее состояние и перспективы развития / Т. Паламарчук и др. ; беседовал А. Новиков // Аналитический банковский журнал. – 2013. – № 4. – С. 74–80, 82, 84.

26. Бизнес-аналитика на пике спроса // Банковские технологии. – 2012. – № 7. – С. 16.

27. Бычкова Н. Бизнес-аналитика для банков: как повысить качество принимаемых решений. Подходы и опыт РДТЕХ / Н. Бычкова // Банковские технологии. – 2010. – № 6. – С. 6–7.

28. Видяев С. Веб-аналитика: как сделать свой бизнес лучше / С. Видяев // Рекламодатель: теория и практика. – 2010. – № 5. – С. 19–22.

29. Голов С. Ф. Управлінський облік / С. Ф. Голов. – Х. : Фактор, 2009. – 784 с.

30. Державна служба статистики України / вебсайт [Електронний ресурс]. – Режим доступу : www.ukrstat.gov.ua

31. Дэвенпорт Т. Аналитика как конкурентное преимущество. Новая наука побеждать / Т. Дэвенпорт, Д. Харрис. – СПб. : Best Business Books, 2010. – 264 с.

32. Лозовая Е. Задача BI – помочь бизнесу воспользоваться накопленной информацией / Е. Лозовая // Банковские технологии. – 2010. – № 3. – С. 44–46.

33. Електронний каталог Національної парламентської бібліотеки України : політемат. база даних містить відом. про вітчизн. та зарубіж. кн., брош., що надходять у фонд НПБ України. – Київ : Нац. парлам. б-ка України, 2018 [Електронний ресурс]. – Режим доступу : catalogue.nplu.org. – Назва з екрана.

34. Еремяшев А. BI-системы: нажми на кнопку – получишь результат / А. Еремяшев // Консультант. – 2010. – № 7. – С. 66–71.

35. Вопросы экономики : журнал / вебсайт [Електронний ресурс]. – Режим доступу : http://www.vopreco.ru/

36. Економіка України : журнал / вебсайт [Електронний ресурс]. – Режим доступу : http://www.economukraine.com.ua/

37. Компьютерные исследования и моделирование : журнал / вебсайт [Електронний ресурс]. – Режим доступу : http://crm.ics.org.ru/

38. Математические модели и информационные технологи в организаци производства : журнал / вебсайт [Електронний ресурс]. – Режим доступу : http://www.istu.ru/period-izd/1159-matematicheskie-modeli-i-informatsionnye-tekhnologii-v-organizatsii-proizvodstva

39. Регіональна економіка : журнал / вебсайт [Електронний ресурс]. – Режим доступу : http://ird.gov.ua/pe/

40. Иванова В. В. О международном стандарте бизнес-анализа / В. В. Иванова, Т. А. Лезина // Вестник Санкт-Петербургского университета. Сер. 5. Экономика. – 2011. – Вып. 4. – С. 107–115.

41. Корнипаев И. Требования для программного обеспечения: рекомендации по сбору и документированию / И. Корнипаев. – М. : Издательство «Книга по требованию», 2013. – 118 с.

42. Исаев Д. В. Аналитические информационные системы / Д. В. Исаев. – М. : ВШЭ, 2008. – 60 с.

43. Казакова Н. А. Финансовая и управленческая бизнес-аналитика в программах высшего образования нового поколения / Н. А. Казакова, А. Ю. Казаков // Экономический анализ: теория и практика. – 2012. – № 9. – С. 30–35.

44. Кизим М. О. Збалансована система показників : монографія / М. О. Кизим, А. А. Пилипенко, В. А. Зінченко. – Харків : ВД «Інжек», 2007. – 192 с.

45. Когнитивная бизнес-аналитика : учеб. для студентов вузов / под ред. Н. М. Абдикеев. – Москва : ИНФРА-М, 2011.

46. Колеченков Р. Классический BI открывает огромные резервы для развития банков / Р. Колеченков // Банковские технологии. – 2013. – № 10. – С. 32–33.

47. Комп'ютерне моделювання та аналіз бізнес-процесів ДПС України [Текст] : монографія / В. А. Одинець та ін. ; за ред. канд. екон. наук В. А. Одинця ; Держ. податк. служба України, Нац. ун-т держ. податк. служби України. – Ірпінь : Вид-во Нац. ун-ту ДПС України, 2012. – 100 с.

48. Конотопов П. Ю. Аналитика: методология, технология и организация информационно-аналитической работы / П. Ю. Конотопов, Ю. В. Курносов. – М. : РУСАКИ, 2004. – 512 с.

49. Кравець Р. Б. Інформаційні технології організації бізнесу : навчальний посібник / Р. Б. Кравець, Ю. О. Сєров, О. В. Марковець. – Львів : Видавництво Львівської політехніки, 2013. – 228 с.

50. Кулаичев А. П. Методы и средства комплексного анализа данных : учеб. пособие / А. П. Кулаичев. – 4-е изд., перераб. и доп. – Москва : Форум: Инфра-М, 2010. – 512 с.

51. ЛІГА БізнесІнформ українська мережа ділової інформації [Електронний ресурс]. – Режим доступу : www.liga.net

52. Ліннiченко А. Н. Збір та аналіз інформації. Стислий огляд методів, що використовуються у бізнес-розвідці та промисловому

шпигунстві [Текст] : навч. посіб. для студ. вищ. навч. закл. ; за матер. легаліз. у відкр. друці та л-рі / А. Н. Ліннiченко. – Тернопіль : Навчальна книга-Богдан, 2012. – 396 с.

53. Мельников В. П. Информационное обеспечение систем управления : учеб. для студентов вузов / В. П. Мельников. – Москва : Академия, 2010.

54. Міністерство доходів і зборів України / вебсайт [Електронний ресурс]. – Режим доступу : http://sta.gov.ua/

55. Міністерство фінансів України / вебсайт [Електронний ресурс]. – Режим доступу : http://www.minfin.gov.ua/

56. Моор П. К. Базы данных : учеб. пособие / П. К. Моор, А. П. Моор ; Тюм. гос. ун-т. – Тюмень : Изд-во ТюмГУ, 2010. – 288 с.

57. Мхитарян С. В. Бизнес-аналитика в менеджменте : практикум [Электронный ресурс] / С. В. Мхитарян. – М. : Евразийский открытый институт, 2011. – 72 с. – Режим доступа : http://bibHodub.m/index.php?page=book&id=90808 (30.09.2014).

58. Національна бібліотека України імені В. І. Вернадського / вебсайт. – Київ : НБУВ, 2018 [Електронний ресурс]. – Режим доступу : www.nbuv.gov.ua – Назва з екрана.

59. Новые инструменты для бизнес-аналитики от IBM // Банковские технологии. – 2009. – № 11. – С. 58–59.

60. Основы бизнес-анализа : учебное пособие / под ред. В. И. Бариленко. – М. : КНОРУС, 2013.

61. Паклин Н. Б. Бизнес-аналитика : учеб. пособие / Н. Б. Паклин, В. И. Орешков. – 2-е изд., доп. и перераб. – Санкт-Петербург : Питер, 2013. – 704 с.

62. Плотніченко І. Б. Теоретико-методологічні аспекти управлінського обліку на промислових підприємствах / І. Б. Плотніченко, О. В. Акіншина // Вісник Національного університету «Львівська політехніка». – 2011. – № 704. – С. 57–54.

63. Поляков К. Л. Основы бизнес-аналитики. – Лекция 8 : Составляющая внутренних бизнес-процессов [Электронный ресурс] / К. Л. Поляков. – М. : Национальный Открытый Университет «ИНТУИТ», 2014. – Ч. 1 : Презентация. – 27 с. – Режим доступа : http://bibliodub.m/mdex.php?page=book&id=239506 (09.09.2014).

64. Проненко В. Бизнес-аналитика: повышение эффективности и успешности компании / В. Проненко // ИТМ. Информационные технологии для менеджмента. – 2011. – № 1–2. – С. 35.

65. Сайт Державної комісії з цінних паперів та фондового ринку [Електронний ресурс]. – Режим доступу : http://www.ssmsc.gov.ua

66. Сайт Національного банку України [Електронний ресурс]. – Режим доступу : http://www.bank.gov.ua/

67. Сафронов А. Аналитика в реальном времени / А. Сафронов // Банковское обозрение. – 2012. – № 7. – С. 112–114.

68. Свиридова А. Внедрение систем бизнес-анализа. Можно ли повысить отдачу? / А. Свиридова, А. Дробышевская // Банковские технологии. – 2013. – № 10. – С. 38–41.

69. Статистична база даних по російській економіці (Державний Університет – Вища Школа Економіки) [Електронний ресурс]. – Режим доступу : http://stat.hse.ru/

70. Сухарев П. Блеск и нищета сводных таблиц / П. Сухарев // КомпьютерПресс. – 2013. – № 1, ч. 12. – С. 85–88.

71. Телио Ю. Банку сейчас особенно нужны системы бизнес-аналитики / Ю. Телио // Банковские технологии. – 2012. – № 6. – С. 56–57.

72. Інформаційні технології та моделювання бізнес-процесів / О. М. Томашевський, Г. Г. Цегелик, М. Б. Вітер, В. І. Дубук. – К. : Видавництво «Центр учбової літератури», 2012. – 296 с.

73. Туманов В. Е. Проектирование хранилищ данных для систем бизнес-аналитики : учебное пособие [Электронный ресурс] / В. Е. Туманов. – М. : Интернет-Университет Информационных Технологий, 2010. – 616 с. : ил., табл., схем. – (Основы информационных технологий). – Режим доступу : http://bibliodub.m/mdex.php?page= book&id=233492 (09.09.2014).

74. Український інститут інтелектуальної власності / вебсайт. – К. : УІІВ, 2018 [Електронний ресурс]. – Режим доступу : http://www.uipv.org – Назва з екрана.

75. Чумачкова Е. Yes I No I Ask Tomorrow, или инструменты для принятия бизнес-решений / Е. Чумачкова, М. Феноменов // Мобильные телекоммуникации (Mobile Communications International. Russian Edition). – 2012. – № 8. – С. 12–13.

76. Шайкан А. В. Бухгалтерський облік у прийнятті управлінських стратегічних рішень : монографія / А. В. Шайкан. – К. : КНЕУ, 2009. – 303 с.

77. Шапцев В. А. Информационные технологии анализа и проектирования бизнес-процессов : учеб. пособие / В. А. Шапцев. – Тюмень : Изд-во ТюмГУ, 2011.

78. Щербаков А. Интернет-аналитика. Поиск и оценка информации в web-ресурсах : практическое пособие [Электронный ресурс] / А. Щербаков. – М. : Книжный мир, 2012. – 78 с. – Режим доступа : http://biblioclub.ru/index.php?page=book&id=89693

79. Экономическая экспертная группа (ЭЭГ) [Электронный ресурс]. – Режим доступа : http://www.eeg.ru/

Навчальне видання

СЕРІЯ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ»

Лаговський Володимир Вікторович,
Краєвський Володимир Миколайович

Бізнес-аналітика та моделювання

Навчальний посібник

Відповідальний за випуск	*Д. Ф. Салахова*
Редактор	*М. М. Грабарчук*
Форматування та комп'ютерна верстка	*Д. П. Завальніцька*

Здано до друку 21.02.2020. Формат 60×84/16
Папір офсетний № 1. Гарнітура «Times New Roman».
Друк. арк. 25,5.
Тираж 300 примірників. Замовлення № 825.

Підготовлено до друку Видавничо-поліграфічним центром
Університету ДФС України
08205, вул. Університетська, 31, м. Ірпінь, Київська область, Україна

Свідоцтво про внесення суб'єкта видавничої справи
до державного реєстру видавців, виготовлювачів і
розповсюджувачів видавничої продукції
Серія ДК № 5104 від 20.05.2016