Тема 2. Поняття даних. Типи та формати зберігання даних. Бази даних. СУБД

План

1. Дані, набір даних та їх атрибути.
2. Формати зберігання даних.
3. Якісний аналіз даних із використанням Data Mining (DM).
4. Системи управління базами даних.

Мета вивчення теми: засвоїти поняття «дані» та особливості різних типів даних; вивчити етапи якісного процесу аналізу даних; засвоїти сутність систем управління базами даних.

Перелік ключових слів та понять із теми

Data Mining, дані, атрибути даних, змінна, шкала, зберігання даних, системи управління базами даних

Теоретичні відомості з теми

1. Дані, набір даних та їх атрибути

У широкому розумінні **дані – це факти, текст, графіки, картинки, звуки, аналогові або цифрові відео-сегменти**. Дані можуть бути отримані в результаті вимірювань, експериментів, арифметичних і логічних операцій. Дані повинні бути представлені у формі, придатній для зберігання, передачі й обробки. Іншими словами, дані – це необроблений матеріал, що надається постачальниками даних і використовується споживачами для формування інформації на основі даних.

У таблиці 2.1 представлена двовимірна таблиця, що представляє собою набір даних.

Таблиця 2.1 − Двовимірна таблиця «об’єкт-атрибут»

	Атрибути				
Об’єкти	Код клієнта	Вік	Сімейний статус	Прибуток	Клас
	1	19	неодр.	1234	1
	2	23	одр.	1222	1
	3	34	одр.	2700	1
	4	24	неодр.	2343	1
	5	26	одр.	1765	2
	6	32	розл.	2652	1
	7	19	неодр.	1200	2
	8	22	неодр.	1765	2
	9	40	одр.	1998	1
	10	43	розл.	4332	1

По горизонталі таблиці розташовуються атрибути об’єкта або його ознаки. По вертикалі таблиці – об’єкти. Об’єкт описується як набір атрибутів. Об’єкт також відомий як запис, випадок, приклад, рядок таблиці тощо.

Атрибут – властивість, що характеризує об’єкт. Наприклад: колір очей людини, температура води. Атрибут також називають змінною, полем таблиці, виміром, характеристикою.

Змінна (variable) – властивість або характеристика, загальна для всіх досліджуваних об’єктів, прояв якої може змінюватися від об’єкта до об’єкта.

Значення (value) змінної є проявом ознаки.

При аналізі даних, як правило, немає можливості розглянути всю сукупність об’єктів, що нас цікавить. Вивчення дуже великих обсягів даних є дорогим процесом, що вимагає великих затрат часу, а також неминуче призводить до помилок, пов’язаних із людським фактором.

Цілком достатньо розглянути деяку частину всієї сукупності, тобто вибірку, і отримати цікаву для нас інформацію на її підставі.

Однак розмір вибірки повинен залежати від різноманітності об’єктів, представлених у генеральній сукупності. У вибірці повинні бути представлені різні комбінації та елементи генеральної сукупності.

Генеральна сукупність (population) – вся сукупність досліджуваних об’єктів, що цікавить дослідника.

Вибірка (sample) – частина генеральної сукупності, певним способом відібрана з метою дослідження та отримання висновків про властивості та характеристики генеральної сукупності.

Параметри – числові характеристики генеральної сукупності.

Статистики – числові характеристики вибірки. Часто дослідження ґрунтуються на гіпотезах. Гіпотези перевіряються за допомогою даних.

Гіпотеза – припущення щодо параметрів сукупності об’єктів, яке має бути перевірено на її частині. Це частково обґрунтована закономірність знань, що служить або для зв’язку між різними емпіричними фактами, або для пояснення факту групи фактів.

Приклад гіпотези: між показниками тривалості життя та якістю харчування є зв’язок. У цьому випадку метою дослідження може бути пояснення змін конкретної змінної, в даному випадку – тривалості життя. Припустимо, існує гіпотеза, що залежна змінна (тривалість життя) змінюється залежно від деяких причин (якість харчування, спосіб життя, місце проживання тощо), які й є незалежними змінними.

Однак змінна першопочатково не є залежною або незалежною, вона стає такою після формулювання конкретної гіпотези. Залежна змінна в одній гіпотезі може бути незалежною в іншій.

Вимірювання – процес присвоєння чисел характеристикам досліджуваних об’єктів згідно певного правила.

У процесі підготовки даних вимірюється не сам об’єкт, а його характеристики.

Шкала – правило, відповідно до якого об’єктам присвоюються числа.

Багато інструментів Data Mining при імпорті даних з інших джерел пропонують вибрати тип шкали для кожної змінної та/або вибрати тип даних для вхідних і вихідних змінних (символьні, числові, дискретні та безперервні).

Користувачеві такого інструменту необхідно володіти цими поняттями.

Змінні можуть бути числовими даними або символьними.

Числові дані, своєю чергою, можуть бути дискретними і неперервними.

Дискретні дані є значеннями ознаки, загальне число яких скінченне або нескінченне, але може бути підраховане за допомогою натуральних чисел від одного до нескінченності.

Прикладом дискретних даних є тривалість маршруту тролейбуса (кількість варіантів тривалості скінченне): 10, 15, 25 хв.

Неперервні дані – дані, значення яких можуть набувати якого завгодно значення в деякому інтервалі. Вимірювання неперервних даних передбачає велику точність.

Приклад неперервних даних: температура, висота, вага, довжина тощо.

Шкали. Існує п’ять типів шкал вимірювань: номінальна, порядкова, інтервальна, відносна і дихотомічна.

Номінальна шкала (nominal scale) – шкала, яка містить тільки категорії; дані в ній не можуть упорядковуватися, з ними не можуть бути зроблені ніякі арифметичні дії.

Номінальна шкала складається з назв, категорій, імен для класифікації і сортування об’єктів або спостережень за деякою ознакою.

Приклад такої шкали: професії, місто проживання, сімейний стан.

Для цієї шкали застосовні тільки такі операції: дорівнює (=), не дорівнює (≠).

Порядкова шкала (ordinal scale) – шкала, в якій числа присвоюють об’єктам для позначення відносної позиції об’єктів, але не величини відмінностей між ними.

Шкала вимірювань дає можливість ранжувати значення змінних. Вимірювання ж у порядковій шкалі містять інформацію лише про порядок проходження величин, але не дозволяють сказати наскільки одна величина більше іншої, або наскільки вона менше іншої.

Приклад такої шкали: місце (1-ше, 2-ге, 3-є), яке команда отримала на змаганнях, номер студента в рейтингу успішності (1-й, 23-й), при цьому невідомо, наскільки один студент успішніше іншого, відомий лише його номер у рейтингу.

Для цієї шкали застосовуються тільки такі операції: дорівнює (=), не дорівнює (≠), більше (>), менше (<).

Інтервальна шкала (interval scale) **–** шкала, різниці між значеннями якої можуть бути обчислені, проте їх відношення не мають сенсу.

Ця шкала дозволяє знаходити різницю між двома величинами, має властивості номінальної та порядкової шкал, а також дозволяє визначити кількісну зміну ознаки.

Приклад такої шкали: температура води в морі вранці – 19 градусів, ввечері – 24, тобто вечірня на 5 градусів вище, але не можна сказати, що вона в 1,26 разів вище.

Номінальна і порядкова шкали є дискретними, а інтервальна шкала – неперервною. Вона дозволяє здійснювати точні вимірювання ознаки і виробляти арифметичні операції додавання, віднімання, множення, ділення.

Для цієї шкали застосовуються тільки такі операції: дорівнює (=), не дорівнює (≠), більше (>), менше (<), операції додавання (+) і віднімання (-).

Відносна шкала (ratio scale) – шкала, в якій є певна точка відліку і можливі відносини між значеннями шкали.

Приклад такої шкали: вага новонародженої дитини (4 кг і 3 кг). Перша в 1,33 рази важче.

Ціна на картоплю в супермаркеті вище в 1,2 рази, ніж ціна на ринку.

Відносні та інтервальні шкали є числовими.

Для цієї шкали можуть бути застосовані тільки такі операції: дорівнює (=), не дорівнює (≠), більше (>), менше (<), операції додавання (+) і віднімання (-), множення (*) і ділення (/).

Дихотомічна шкала (dichotomous scale) – шкала, яка містить тільки дві категорії.

Приклад такої шкали: стать (чоловіча і жіноча).

Приклад використання різних шкал для вимірювань властивостей різних об’єктів, у даному випадку характеристик людей, наведено в таблиці 2.2.

Таблиця 2.2 − Множина вимірювань властивостей різних об’єктів

Номер об’єкту	**Професія (номінальна шкала)**	**Середній бал (інтервальна шкала)**	**Освіта (порядкова шкала)**
1	Слюсар	22	середня
2	Вчений	55	вища
3	Вчитель	47	вища

Приклад використання різних шкал для вимірювань властивостей однієї системи, у даному випадку температурних умов, наведено в таблиці 2.3.

Таблиця 2.3 − Множина вимірювань властивостей однієї системи

Дата змінення	**Хмарність (номінальна шкала)**	**Температура о 7 годині (інтервальна шкала)**	**Сила вітру (порядкова шкала)**
3 жовтня	Хмарно	22ºС	Сильний вітер
4 жовтня	Напівхмарно	17ºС	Слабий вітер
5 жовтня	Ясно	23ºС	Дуже сильний вітер

Типи наборів даних. Найбільш часто зустрічаються дані, що складаються із записів (record data).

Приклади таких наборів даних: табличні дані, матричні дані, документальні дані, транзакційні або операційні.

Табличні дані – дані, що складаються із записів, кожен з яких складається з фіксованого набору атрибутів.

Транзакційні дані представляють собою особливий тип даних, де кожен запис, що є транзакцією, включає набір значень.

Приклад транзакційної бази даних, що містить перелік покупок клієнтів магазину, наведено в таблиці 2.4.

Таблиця 2.4 − Приклад транзакційних даних

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Графічні дані. Приклади графічних даних: молекулярні структури; графи (рис. 2.1); карти.

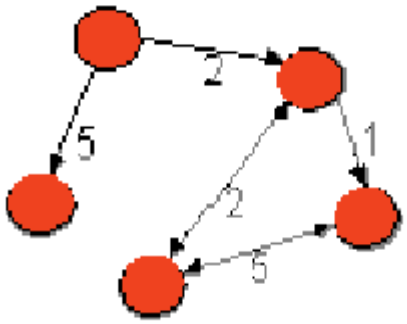

Рисунок 2.1 – Приклад графу

За допомогою карт, наприклад, можна відстежити зміни об’єктів у часі та просторі, визначити характер їх розподілу на площині або в просторі.

Перевагою графічного представлення даних є простота їх сприйняття у порівнянні, наприклад, з табличними даними.

Приклад карти, що є картою Кохонена (моделлю нейронних мереж), представлений на рис. 2.2.

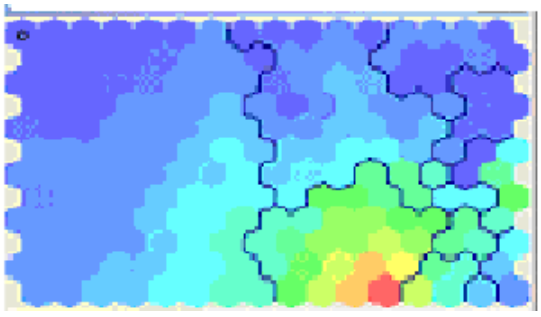

Рисунок 2.2 – Приклад даних типу «Карта Кохонена»

Хімічні дані представляють собою особливий тип даних. Приклад таких даних: молекула бензолу C_6H_6 (рис. 2.3).

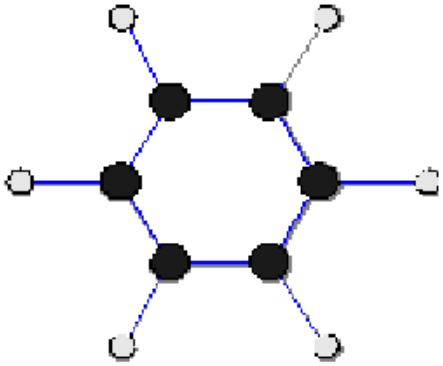

Рисунок 2.3 – Приклад хімічних даних

2. Формати зберігання даних

Одна з основних особливостей даних сучасного світу полягає в тому, що їх стає дуже багато.

Можливі чотири аспекти роботи з даними:

- визначення даних;
- обчислення;
- маніпулювання;
- обробка (збір, передача тощо).

При маніпулюванні даними використовується структура даних типу «файл». Файли можуть мати різні формати.

Більшість інструментів Data Mining дозволяють імпортувати дані з різних джерел, а також експортувати результуючі дані в різні формати.

Дані для експериментів зручно зберігати в якомусь одному форматі.

У деяких інструментах Data Mining ці процедури називаються імпорт / експорт даних, інші дозволяють напряму відкривати різні джерела даних і зберігати результати Data Mining в одному із запропонованих форматів.

Найбільш поширені формати зберігання даних представлені на рис. 2.4.

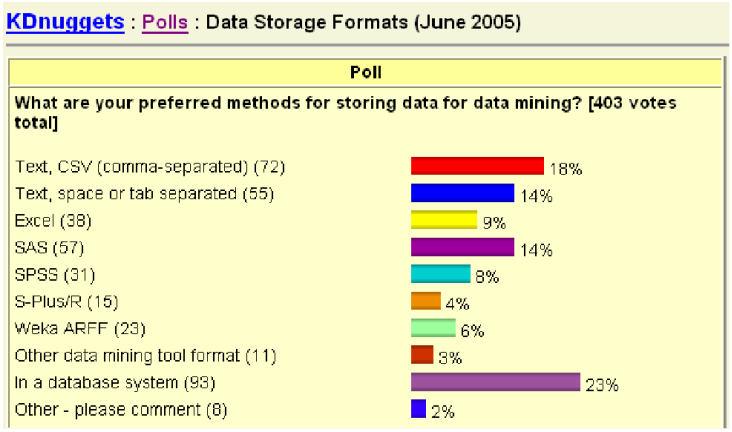

Рисунок 2.4 – Найбільш поширені формати зберігання даних

Найбільше число опитаних (23%) на он-лайн-ресурсі KDnuggets вважають за краще зберігати дані у форматі тієї бази даних, яку вони використовують. У форматі Text, CSV – 18%, по 14% опитаних зберігають дані у форматі Text, space or tab separated і SAS; у форматі Excel – 9%, SPSS – 8%, S-Plus/R – 4%, Weka ARFF – 6%, в інших форматах інструментів Data Mining – 2%.

Як бачимо з результатів опитування, найбільш поширеним форматом зберігання даних для Data Mining виступають бази даних.

3. Якісний аналіз даних з використанням DM.

Для якісного аналізу будь-яких даних слід дотримуватися загальної схеми використання DM:

1. Висування гіпотез.
2. Збір та систематизація даних.
3. Підбір адекватної моделі.
4. Тестування та інтерпретація отриманих даних.
5. Використання в реальних умовах.

Ця схема не залежить від предметної області та сфери діяльності. Вона є універсальною.

1) Висування гіпотез

Гіпотезою тут будемо вважати припущення про вплив певних факторів на процес, що досліджується.

Автоматизувати процес висування гіпотез є вкрай складно, тому цю задачу мають розв’язувати експерти – фахівці в предметній області.

Слід довіритися їх досвіду та здоровому глузду, максимально використати ці знання про предмет досліджень і зібрати як найбільше гіпотез/припущень.

Зазвичай, добрі результати надають тактики «круглого столу» або «мозкової атаки». На початку слід зібрати та систематизувати всі ідеї, а оцінювати їх пізніше. У результаті повинен бути складений перелік з описів всіх факторів досліджуваного об’єкту.

Наприклад, для задачі прогнозування попиту товару потрібно скласти перелік факторів, що впливатимуть на об’єкт і експертно оцінити суттєвість кожного з них (табл. 2.5).

Таблиця 2.5 − Вплив кожного фактору (у %) на попит на товар

Сезон	100
День тижня	80
Обсяг продажів за попередні тижні	100
Обсяг продажів за аналогічний період минулого року	95
Рекламна компанія	60
Маркетингові заходи	40
Якість продукції	50
Бренд	25

Коливання ціни від середньоринкової	60
Наявність подібного товару в конкурентів	15

Згодом під час аналізу може з’ясуватися, що фактор, який експерти оцінили як важливий, буде мати незначний вплив на процес і навпаки.

2) Збір та систематизація даних

2.1. Збір даних

Для аналізу потрібно як найбільше даних, бо це надає можливість оцінити вплив максимальної кількості показників. Згодом простіше відхилити певну частину даних, аніж розпочинати новий збір.

Методи збору:

1. Отримання даних із внутрішніх джерел.

Це не складно, бо така інформація зазвичай зберігається в облікових системах у табличній формі, де існують різні механізми отримання звітів та експортування даних.

2. Отримання відомостей із непрямих даних.

Наприклад, потрібно оцінити реальний фінансовий стан мешканців певного регіону. Існує кілька категорій товару (зокрема, авто), що різняться за ціною – для незаможних, середнього класу, заможних. Якщо отримати звіт про продажі товару в цьому районі і проаналізувати пропорції, то дійдемо до висновку: чим більшим є відсоток продажів дорогого товару, тим заможнішими є мешканці.

3. Використання відкритих джерел.

До широкого загалу надаються статистичні збірники, звіти корпорацій, результати маркетингових досліджень, соціологічні опитування.

4. Влаштування власних маркетингових досліджень та подібних заходів по збору даних.

Це зазвичай є дорогим заходом, але доволі ефективним.

5. Наповнення даних згідно експертних оцінок співробітниками організації.

Слід оцінити вартість збору даних, що потрібні для аналізу. Одні дані беруться з публічних інформаційних джерел, інші мають бути оплачені, дані про діяльність конкурентів можуть бути доволі дорогими.

Вартість збору інформації різними методами суттєво різниться за ціною та витраченим часом, тому слід зважати на співвідношення теперішніх витрат із майбутніми результатами.

Від даних, які експерти вважають несуттєвими, певна річ можна відмовитися, але не від значущих даних, бо аналіз базуватиметься в цьому випадку на другорядних факторах і, відповідно, отримана модель буде надавати нестабільні та невірні результати.

4. Системи управління базами даних.

Не кожен блок інформації можна вважати базою даних. ***База даних*** – *це сукупність даних, яким властива структурованість і взаємопов'язаність, а також незалежність від прикладних програм.*

Пояснимо, що означають названі властивості бази даних. Щоб користувач легко міг знаходити потрібну інформацію, остання має бути організована певним чином. Це стосується не лише інформації в комп'ютері, а й будь-якої інформації про об'єкти реального світу. Скажімо, зручно знаходити потрібну книгу в бібліотеці, користуючись каталогом. Легко відшукати в газеті оголошення, що вас цікавлять. Така легкість пошуку можлива завдяки тому, що дані в каталозі або в газеті мають структуру, або, інакше, *структуровані*. Усі книги описані однаково: автор, назва, видавництво, рік видання тощо. Усі оголошення з продажу розміщені по рубриках і також мають визначену структуру: короткий опис товару, ціна, телефон.

Структура бази даних складніша, ніж структура простого каталогу або набору газетних оголошень. Це зумовлено насамперед властивістю *взаємопов'язаності* даних у базі. Пояснимо це на такому прикладі: скажімо, ви хотіли б, крім каталожних карток, що описують кожну книгу, мати картки з інформацією про кожного автора (рік народження, літературний жанр, хобі тощо). Якби такі картки були створені, це був би приклад взаємозалежності даних: відомості про окрему книгу пов'язані з інформацією про автора. Цей зв'язок здійснюється через визначений параметр – прізвище автора.

Нарешті, остання з названих властивостей баз даних – це їхня *незалежність від прикладних програм.* Бази даних складаються таким чином, щоб із ними можна було працювати в різних програмних середовищах і на різних комп'ютерних платформах.

Щоб оперувати даними, які становлять базу, необхідна окрема програма – система управління базами даних. *Керівна програма, призначена для збереження, пошуку й обробки даних у базі, називається* ***системою управління базами даних*** *(скорочено СУБД).*

Система управління базами даних — це прикладна програма, реалізована на електронній обчислювальній машині чи обчислювальному комплексі. За допомогою її можна:

1) створювати структуру бази даних, вводити інформацію та зберігати її на зовнішніх носіях;
2) виконувати певне коло операцій із даними;
3) одержувати результати та зберігати їх на зовнішніх носіях або передавати на віддалені термінали;
4) виводити інформацію на термінал у зручній для користувача формі або на друкувальні пристрої;
5) давати можливість працювати з базами даних багатьом користувачам.

У цьому визначенні відсутній людський фактор – персонал, який відповідає за дані (адміністратор бази даних), але для розуміння роботи СУБД буде достатньо попереднього визначення.

Сучасні СУБД – це програмні додатки, які дозволяють виконувати різноманітні завдання. Усі існуючі системи задовольняють, як правило, таким вимогам:

✓**можливості маніпулювання даними** (введення, вибір, вставка, відновлення, видалення тощо). Основні операції з даними виконуються під керуванням СУБД. Важливими показниками є продуктивність СУБД, витрати на збереження і використання даних, простота звернення до бази даних тощо;

✓**можливість пошуку і формування запитів.** За допомогою запитів користувач може оперативно одержувати різну інформацію, що зберігається в базі даних.

✓**забезпечення цілісності (узгодженості) даних.** Під час використання даних багатьма користувачами важливо забезпечити коректність операцій, щоб запобігти порушенню узгодженості даних. Порушення узгодженості даних може призвести до їх невідновної втрати;

✓**забезпечення захисту і таємності.** Крім захисту від некоректних дій користувачів, важливо забезпечити захист даних від несанкціонованого доступу і від апаратних збоїв. Проникнення в базу осіб, які не мають на це права, може спричинити руйнацію даних. Таємність бази даних дозволяє визначити коло осіб, що мають доступ до інформації, і порядок доступу.

Сьогодні існує багато СУБД, що відрізняються архітектурою, внутрішньою мовою програмування, операційною системою, якою вони керуються, а також іншими характеристиками. Найпопулярнішими СУБД, що встановлюються в невеликих організаціях і орієнтовані на роботу з кінцевими користувачами, є Access, FoxPro, Paradox. До складніших систем належать розподілені СУБД, що призначені для роботи з великими базами даних, розподіленими на кількох серверах (сервери можуть міститися в різних регіонах). Потужними СУБД такого типу є Oracle, Sybase, Informix.

Вимоги до СУБД

СУБД разом із БД іноді називають банком даних. У банках даних повинні бути передбачені засоби, що забезпечують захист певних областей даних від несанкціонованого доступу.

Банк даних повинен відповідати таким вимогам:

1. Мати можливість оновлення, поповнення та розширення БД.
2. Забезпечити високу надійність зберігання інформації.
3. Видавати повну та вірогідну інформацію на запити.
4. Мати засоби, що забезпечують захист БД від несанкціонованого доступу.

Основні функції СУБД

До основних функцій СУБД належать такі:

1) опис БД (вказати назви полів, їх довжину, тип та інше);

2) введення в БД підготовлених даних;

3) перевірка правильності введення даних (контроль за типом);

4) редагування даних (вилучення, заміна, коректування, вставка, доповнення);

5) обробка запитів від користувачів (пошук певної інформації);

6) забезпечення одночасної роботи декількох користувачів з однією БД;

7) захист даних.

Питання для самоконтролю

1. Дайте визначення поняттю вибірка (sample).
2. Дайте визначення поняттю гіпотеза.
3. Наведіть приклади застосування гіпотез.
4. Які види шкал ви знаєте? Чим вони відрізняються?
5. Які типи даних ви знаєте?
6. З яких етапів складається загальна схема використання DM?