МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

СУМСЬКИЙ НАЦІОНАЛЬНИЙ АГРАРНИЙ УНІВЕРСИТЕТ

Кафедра кібернетики та інформатики

Курсова робота

З дисципліни «Технологія бізнес аналітики»

На тему: «Застосування методів Data mining для вирішення практичних завдань на базі ГІС платформи»

Виконав студент 1ст. курсу

Групи ІСТ2001ст.

Кримов Артем Миколайович

Перевірили: (підпис) _____ (ПІБ) Винник О.Б.

(підпис) _____ (ПІБ) Гасько Н.І.

(підпис) _____ (ПІБ) Андрієвська С.В.

Суми 2021

ЗМІСТ

ВСТУП

Data Mining буквально в перекладі з англійської означає " дані, факти, відомості, інформація», і «видобуток корисних копалин». Область Data Mining почалася з семінару, проведеного Григорієм П'ятецьким-Шапіро в 1989 році. Спочатку, завдання ставилося наступним чином: є досить велика база даних, передбачається, що в базі даних знаходяться якісь «приховані знання», необхідно розробити методи виявлення знань, прихованих у великих обсягах вихідних «сирих» даних. Тоді ж був запропонований термін Data Mining як основна назва, що використовується для позначення сукупності методів виявлення в даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень в різних сферах людської діяльності. Одне з найважливіших призначень методів Data Mining полягає в наочному поданні результатів обчислень, що дозволяє використовувати інструментарій Data Mining людьми, які не мають спеціальної математичної підготовки. У той же час, застосування статистичних методів аналізу даних вимагає хорошого володіння теорією ймовірностей і математичною статистикою.

Отже, Data Mining - це процес підтримки прийняття рішень, заснований на пошуку в даних прихованих закономірностей (шаблонів інформації). Суть і мета технології Data Mining можна охарактеризувати як технологію, яка призначена для пошуку у великих обсягах даних неочевидних, об'єктивних і корисних на практиці закономірностей.

Метою даної роботи є розкриття основних властивостей можливостей технології "видобутку знань", а також розгляд можливостей застосування технології Data Mining та просторового аналізу і візуалізації електронних таблиць за допомогою GIS. XL.

РОЗДІЛ 1. Теоретичний аналіз моделей та методів інтелектуального аналізу даних

1.1 Основні поняття Data Mining

Data Mining - це процес підтримки прийняття рішень, заснований на пошуку в даних прихованих закономірностей (шаблонів інформації) [3].

Технологію Data Mining досить точно визначає Григорій Пиатецкий-Шапіро (Gregory Piatetsky-Shapiro) - один із засновників цього напрямку:

Data Mining - це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень у різних сферах людської діяльності. Суть і мета технології Data Mining можна охарактеризувати так: це технологія, яка призначена для пошуку у великих обсягах даних неочевидних, об'єктивних і корисних на практиці закономірностей.

Неочевидних - це означає, що знайдені закономірності не виявляються стандартними методами обробки інформації або експертним шляхом.

Об'єктивні - це означає, що виявлені закономірності будуть повністю відповідати дійсності, на відміну від експертної думки, яке завжди є суб'єктивним.

Практично корисних - це означає, що висновки мають конкретне значення, яким можна знайти практичне застосування.

Знання - сукупність відомостей, яка утворює цілісний опис, відповідне деякому рівню обізнаності про описуваному питанні, предмет, проблеми і т. д.

Використання знань (knowledge deployment) означає дійсне застосування знайдених знань для досягнення конкретних переваг. Якщо наводити ще кілька визначень поняття Data Mining. Data Mining - це процес виділення з даних неявній і неструктурованої інформації і представлення її у вигляді, придатному для використання.

Data Mining - це процес виділення, дослідження і моделювання великих обсягів даних для виявлення невідомих до цього структур (patterns) з метою досягнення переваг у бізнесі (визначення SAS Institute).

Data Mining - це процес, мета якого - виявити нові значущі кореляції, зразки та тенденції в результаті просівання великого обсягу збережених даних з використанням методик розпізнавання зразків плюс застосування статистичних і математичних методів (визначення Gartner Group).

В основу технології Data Mining покладена концепція шаблонів (patterns), які являють собою закономірності, властиві підвиборок даних, які можуть бути виражені у формі, зрозумілій людині. "Mining" по-англійськи означає "видобуток корисних копалин", а пошук закономірностей у величезній кількості даних дійсно схожа цього процесу.

Мета пошуку закономірностей - подання даних у вигляді, що відбиває шукані процеси. Побудова моделей прогнозування також є метою пошуку закономірностей.

Data Mining як частина ринку інформаційних технологій

Агентство Gartner Group, що займається аналізом ринків інформаційних технологій, в 1980-х роках ввів термін "Business Intelligence " (BI), ділової інтелект або бізнес-інтелект. Цей термін запропонований для опису різних концепцій і методів, які покращують бізнес рішення шляхом використання систем підтримки прийняття рішень. У 1996 році агентство уточнило визначення даного терміна.

Business Intelligence - програмні засоби, що функціонують у рамках підприємства і забезпечують функції доступу і аналізу інформації, яка знаходиться у сховищі даних, а також забезпечують прийняття правильних і обґрунтованих управлінських рішень. Поняття BI об'єднує в собі різні засоби і технології аналізу і обробки даних масштабу підприємства. На основі цих коштів створюються BI-системи, мета яких - підвищити якість інформації для прийняття управлінських рішень. BI-системи також відомі під назвою Систем Підтримки Прийняття Рішень (СППР, DSS, Decision Support System). Ці

системи перетворюють дані в інформацію, на основі якої можна приймати рішення, тобто підтримує прийняття рішень.

Gartner Group визначає склад ринку систем Business Intelligence як набір програмних продуктів наступних класів:

- засоби побудови сховищ даних (data warehousing, ХД);
- системи оперативної аналітичної обробки (OLAP);
- інформаційно-аналітичні системи (Enterprise Information Systems EIS);
- засоби інтелектуального аналізу даних (data mining);
- інструменти для виконання запитів і побудови звітів (query and reporting tools).

Класифікація Gartner базується на методі функціональних завдань, де програмні продукти кожного класу виконують певний набір функцій або операцій з використанням спеціальних технологій.

Наведемо кілька коротких цитат [4] найбільш впливових членів бізнес-спільнот, які є експертами в цій відносно новій технології. Керівництво по придбанню продуктів Data Mining (Enterprise Data Mining Buying Guide) компанії Aberdeen Group: "Data Mining - технологія видобутку корисної інформації з баз даних. Однак у зв'язку з істотними відмінностями між інструментами, досвідом і фінансовим станом постачальників продуктів, підприємствам необхідно ретельно оцінювати передбачуваних розробників Data Mining і партнерів. Щоб максимально використовувати потужність масштабованих інструментів Data Mining комерційного рівня, підприємству необхідно вибрати, очистити і перетворити дані, іноді інтегрувати інформацію, отриману із зовнішніх джерел, і встановити спеціальне середовище для роботи Data Mining алгоритмів.

Результати Data Mining великою мірою залежать від рівня підготовки даних, а не від "чудесних можливостей" якогось алгоритму або набору алгоритмів. Близько 75 % роботи над Data Mining полягає в зборі даних, який відбувається ще до того, як запускаються самі інструменти. Неграмотно

застосувавши деякі інструменти, підприємство може безглуздо розтратити свій потенціал, а іноді і мільйони доларів".

Думка Херба Эдельштайна (Herb Edelstein), відомого у світі експерта в області Data Mining, Сховищ даних і CRM: "Недавнє дослідження компанії Two Crows показало, що Data Mining знаходиться все ще на ранній стадії розвитку. Багато організації цікавляться цією технологією, але лише деякі активно впроваджують такі проекти. Вдалося з'ясувати ще один важливий момент: процес реалізації Data Mining на практиці виявляється більш складним, ніж очікується.

IT-команди захопилися міфом про те, що засоби Data Mining прості у використанні. Передбачається, що досить запустити такий інструмент на терабайтної базі даних, і моментально з'явиться корисна інформація. Насправді, успішний Data Mining-проект вимагає розуміння суті діяльності, знання даних і інструментів, а також процесу аналізу даних".

Перш ніж використовувати технологію Data Mining, необхідно ретельно проаналізувати її проблеми, обмеження та критичні питання, з нею пов'язані, а також зрозуміти, чого ця технологія не може. Data Mining не може замінити аналітика. Технологія не може дати відповіді на ті питання, які не були задані. Вона не може замінити аналітика, а лише дає йому потужний інструмент для полегшення і покращення його роботи. Складність розробки та експлуатації програми Data Mining. Оскільки дана технологія є мультидисциплінарною областю, для розробки програми, що включає Data Mining, необхідно задіяти фахівців з різних областей, а також забезпечити їх якісне взаємодія.

Кваліфікація користувача

Різні інструменти Data Mining мають різну ступінь "доброзичливості" інтерфейсу і вимагають певної кваліфікації користувача. Тому програмне забезпечення повинно відповідати рівню підготовки користувача. Використання Data Mining повинно бути нерозривно пов'язане з підвищенням кваліфікації користувача. Однак фахівців з Data Mining, які б добре розбиралися в бізнесі, поки ще мало.

Витяг корисних відомостей неможливо без гарного розуміння суті даних Необхідний ретельний вибір моделі та інтерпретація залежностей або шаблонів, які виявлені. Тому робота з такими засобами вимагає тісної співпраці між експертом в предметній області і фахівцем з інструментів Data Mining. Побудовані моделі повинні бути грамотно інтегровані в бізнес-процеси для можливості оцінки та оновлення моделей. Останнім часом системи Data Mining поставляються як частина технології сховищ даних.

Складність підготовки даних

Успішний аналіз вимагає якісної попередньої обробки даних. За твердженням аналітиків і користувачів баз даних, процес попередньої обробки може зайняти до 80% відсотків всього Data Mining-процесу. Таким чином, щоб технологія працювала на себе, потрібно багато зусиль і часу, які йдуть на попередній аналіз даних, вибір моделі і її коригування.

Великий відсоток неправдивих, недостовірних або безглуздих результатів.

З допомогою Data Mining можна відшукувати дійсно дуже цінну інформацію, яка незабаром дасть великі дивіденди у вигляді фінансової та конкурентної вигоди.

Однак Data Mining досить часто робить безліч помилкових і не мають сенсу відкриттів. Багато фахівців стверджують, що Data Mining - кошти можуть видавати величезна кількість статистично недостовірних результатів. Щоб цього уникнути, необхідна перевірка адекватності отриманих моделей на тестових даних.

Висока вартість

Якісна Data Mining - програма може коштувати досить дорого для компанії. Варіантом є придбання вже готового рішення з попередньою перевіркою його використання, наприклад на демо-версії з невеликою вибіркою даних.

Наявність достатньої кількості репрезентативних даних

Засоби Data Mining, на відміну від статистичних, теоретично не вимагають наявності строго певної кількості ретроспективних даних. Ця особливість може стати причиною виявлення недостовірних, помилкових моделей і, як результат, прийняття на їх основі неправильних рішень. Необхідно здійснювати контроль статистичної значимості виявлених знань.

1.2 Відмінності Data Mining від інших методів аналізу даних

Традиційні методи аналізу даних (статистичні методи) і OLAP в основному орієнтовані на перевірку заздалегідь сформульованих гіпотез (verification-driven data mining) і на "грубий" розвідувальний аналіз, що становить основу оперативної аналітичної обробки даних (OnLine Analytical Processing, OLAP), в той час як одне з основних положень Data Mining - пошук неочевидних закономірностей. Інструменти Data Mining можуть знаходити такі закономірності самостійно і також самостійно будувати гіпотези про взаємозв'язки. Оскільки саме формулювання гіпотези щодо залежностей є найбільш складним завданням, перевага Data Mining порівняно з іншими методами аналізу є очевидним.

Більшість статистичних методів для виявлення взаємозв'язків в даних використовують концепцію усереднення за вибіркою, що приводить до операцій над неіснуючими величинами, тоді як Data Mining оперує реальними значеннями.

OLAP більше підходить для розуміння ретроспективних даних, Data Mining спирається на ретроспективні дані для отримання відповідей на питання про майбутнє.

Перспективи технології Data Mining

Потенціал Data Mining дає "зелене світло" для розширення меж застосування технології. Щодо перспектив Data Mining можливі наступні напрямки розвитку:

- виділення типів предметних областей з відповідними їм евристиками, формалізація яких полегшить вирішення відповідних задач Data Mining, що належать до цих галузей;
- створення формальних мов і логічних засобів, за допомогою яких будуть формалізовані міркування і автоматизація яких стане інструментом вирішення задач Data Mining в конкретних предметних областях;
- створення методів Data Mining, здатних не тільки витягати з даних закономірності, але і формувати якісь теорії, які спираються на емпіричні дані;
- подолання істотного відставання можливостей інструментальних засобів Data Mining від теоретичних досягнень в цій області.

Якщо розглядати майбутнє Data Mining в короткостроковій перспективі, то очевидно, що розвиток цієї технології найбільш направлено до областей, пов'язаних з бізнесом. У короткостроковій перспективі продукти Data Mining можуть стати такими ж звичайними і необхідними, як електронна пошта, і, наприклад, використовуватися користувачами для пошуку найнижчих цін на певний товар або найбільш дешевих квитків. У довгостроковій перспективі майбутнє Data Mining є дійсно захоплюючим - це може бути пошук інтелектуальними агентами нових видів лікування різних захворювань, так і нового розуміння природи всесвіту. Однак Data Mining таїть у собі потенційну небезпеку - адже все більша кількість інформації стає доступним через всесвітню мережу, в тому числі і відомості приватного характеру, і все більше знань можливо добути з неї.

Не так давно найбільший онлайновий магазин "Amazon" опинився в центрі скандалу з приводу отриманого патенту "Методи і системи допомоги користувачам при купівлі товарів", який являє собою не що інше як черговий продукт Data Mining, призначений для збору персональних даних про відвідувачів магазину. Нова методика дозволяє прогнозувати майбутні запити на підставі фактів покупок, а також робити висновки про їх призначення. Мета даної методики - те, про що говорилося вище - одержання як можна більшої кількості інформації про клієнтів, у тому числі і приватного характеру (стать,

вік, уподобання тощо). Таким чином, збираються дані про приватного життя покупців магазину, а також членів їх сімей, включаючи дітей. Останнім заборонено законодавством багатьох країн - збір інформації про неповнолітніх можливий там тільки з дозволу батьків.

Дослідження відзначають, що існують як успішні рішення, які використовують Data Mining, так і невдалий досвід застосування цієї технології [5]. Області, де застосування технології Data Mining, швидше за все, будуть успішними, мають такі особливості:

- вимагають рішень, заснованих на знаннях;
- мають змінюється навколишнє середовище;
- дешеві, достатні і значущі дані;
- забезпечують високі дивіденди від правильних рішень.

Існуючі підходи до аналізу

Досить довго дисципліна Data Mining не визнавалася повноцінної самостійної областю аналізу даних, іноді її називають "задвірками статистики" (Pregibon, 1997).

На сьогоднішній день визначилося кілька точок зору на Data Mining. Прихильники одного з них вважають його міражем, що відволікає увагу від класичного аналізу даних. Прихильники іншого напрямку - це ті, хто приймає Data Mining як альтернативу традиційному підходу до аналізу. Є й середина, де розглядається можливість спільного використання сучасних досягнень в області Data Mining і класичному статистичному аналізі даних. Технологія Data Mining постійно розвивається, залучає до себе все більший інтерес як з боку наукового світу, так і з боку застосування досягнень технології в бізнесі. Щорічно проводиться безліч наукових і практичних конференцій, присвячених Data Mining, одна з яких - Міжнародна конференція з Knowledge Discovery Data Mining (International Conferences on Knowledge Discovery and Data Mining).

Періодичні видання з Data Mining: Data Mining and Knowledge Discovery, KDD Explorations, ACM-TODS, IEEE-TKDE, JIIS, J. ACM, Machine Learning, Artificial Intelligence.

Матеріали конференцій: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, Machine learning (ICML), AAAI, IJCAI, COLT (Learning Theory).

Порівняння статистики, машинного навчання і Data Mining

"За останні роки, коли, прагнучи до підвищення ефективності та прибутковості бізнесу, при створенні БД всі стали користуватися засобами обробки цифрової інформації, з'явився і побічний продукт цієї активності-гори зібраних даних: і ось все більше поширюється ідея про те, що ці гори сповнені золота".

У минулому процес видобутку золота в гірничій промисловості складався з вибору ділянки землі і подальшого її просіювання велику кількість разів. Іноді шукач знаходив кілька цінних самородків або міг натрапити на золотоносну жилу, але в більшості випадків він взагалі нічого не знаходив і йшов далі до іншого багатообіцяючого місця або ж зовсім кидав добувати золото, вважаючи це заняття марною тратою часу.

Сьогодні з'явилися нові наукові методи і спеціалізовані інструменти, які зробили гірську промисловість набагато більш точною і продуктивною. Data Mining для даних розвинулася майже таким же способом. Старі методи, що застосовувалися математиками і статистиками, забирали багато часу, щоб в результаті отримати конструктивну і корисну інформацію.

Сьогодні представлено безліч інструментів, що включають різні методи, які роблять Data Mining прибутковою справою, все більш доступним для більшості компаній.

Статистика - це наука про методи збору даних, їх обробки та аналізу для виявлення закономірностей, властивих досліджуваному явищу.

Статистика є сукупністю методів планування експерименту, збору даних, їх подання та узагальнення, а також аналізу та отримання висновків на підставі цих даних.

Статистика оперує даними, отриманими в результаті спостережень або експериментів. Одна з наступних глав буде присвячена поняттю даних.

Поняття Машинного навчання

Єдиного визначення машинного навчання на сьогоднішній день немає.

Машинне навчання можна охарактеризувати як процес отримання програмою нових знань. Мітчелл в 1996 році дав таке визначення: "Машинне навчання - це наука, яка вивчає комп'ютерні алгоритми, автоматично поліпшуються під час роботи". Одним з найбільш популярних прикладів алгоритму машинного навчання є нейронні мережі.

Поняття штучного інтелекту

Штучний інтелект-науковий напрям, в рамках якого ставляться і вирішуються завдання апаратного або програмного моделювання видів людської діяльності, традиційно вважаються інтелектуальними.

Термін інтелект (intelligence) походить від латинського intellectus, що означає розумові здібності людини.

Відповідно, штучний інтелект (AI, Artificial Intelligence) тлумачиться як властивість автоматичних систем брати на себе окремі функції інтелекту людини. Штучним інтелектом називають властивість інтелектуальних систем виконувати творчі функції, які традиційно вважаються прерогативою людини.

Кожне з напрямків, що сформували Data Mining, має свої особливості. Проведемо порівняння з деякими з них.

Порівняння статистики, машинного навчання та Data Mining

Статистика

- Більш, ніж Data Mining, базується на теорії.
- Більше зосереджується на перевірці гіпотез.

Машинне навчання

- Більш евристично.
- Концентрується на поліпшенні роботи агентів навчання.

Data Mining.

- Інтеграція теорії та евристик.

• Сконцентрована на єдиному процесі аналізу даних, включає очищення даних, навчання, інтеграцію і візуалізацію результатів.

1.3 Аналіз геоданих як складових Data mining

Геоданими називають дані про об'єкти, процеси та явлениях на земній поверхні Землі, які включають як основу просторову інформацію [1, 2]. На цю основу нанизують різноманітні дані для подальшого просторового, економічного, регіонального та інших видів аналізу. Геодані є не просто да - ними, а являють собою систему даних і інформаційний ресурс [4]. Однією з особливостей геоданих є те, що вони відображають реально існуючі просторові відносини. Безліч геоданих збирається за допомогою різних технологій і систем. Дані відображають різні характеристики і властивості. Вони можуть мати різні розмірності різну кількість значущих цифр, різне число розрядів, різну точність та інше. Зібрані дані можуть зберігатися у вигляді наборів або файлів. Крім того, при зборі дані можуть організовувати пов'язані сукупності, називаються моделями [4]. Для того щоб різнорідні дані та моделі можна було обробляти в одній системі вони повинні бути впорядковані і зведені до єдиної інформаційної моделі, в якій вони будуть доповнювати один одного.

Проблеми інтелектуального аналізу

Сучасний інформаційний бум призвів до проблеми "великі дані" [5, 6]. Це потребове створення спеціальних технологій для швидкої переробки цих даних. Необхідність автоматизованого інтелектуального аналізу даних стала очевидною в першу чергу із-за величезних масивів історичної і знову збираної інформації. Важко, навіть приблизно оцінити обсяг щоденних даних, накопичуваних різними компаніями, державними, науковими та медичними організаціями. Людський розум, навіть такий тренований, як розум професійного аналітика, просто не в змозі вчасно аналізувати такі величезні інформаційні потоки. Специфіка сучасних вимог до такої переробки наступна:

- дані мають необмежений обсяг;
- дані є різнорідними (кількісними, якісними, текстовими);
- результати повинні бути конкретними і зрозумілими;
- технології для обробки сирих даних повинні бути прості у використанні.

В основу сучасної технології інтелекту- альної обробки даних покладена концепція шаблонів, що відображають фрагменти багатоаспектних відносин у даних [8].

Ці шаблони являють собою закономірності, які можуть бути компактно виражені у зрозумілій людині формі. Пошук шаблонів здійснюється методами, що не обмежені рамками апріорних пропозицій про структуру вибірки і вигляді розподілених значень аналізованих показників.

Знайдені шаблони повинні відображати неочевидні, несподівані регулярності в даних, які складають так звані приховані знання. Сирі дані містять глибинний пласт знань, при грамотній розкопці якого можуть бути виявлені справжні самородки.

Інтелектуальний аналіз даних (ІАД) (data mining) — це процес виявлення в "сирих" (первинних) великих масивах даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для прийняття рішень у різних сферах людської діяльності [9].

Інтеллектуальный аналіз даних можна розглядати також як процес трансформації неявних знань.

У загальному випадку процес ІАД складається з трьох стадій:

- виявлення закономірностей (вільнийпошук в інформаційному полі;
- використання виявлених закономер - ностей для передбачення невідомих значень (ретроспективний, поточний і прогностичне моделювання);
- аналіз винятків, призначений для виявлення і тлумачення аномалій у знайдених закономірностей (ліквідація семантичних розривів).

Всі методи інтелектуального аналізу даних поділяються на дві великі групи за принципом роботи з початковими навчальними даними.

У першому випадку вихідні дані можуть зберігатися в явному деталізованому вигляді і безпосередньо використовуватися для прогностичного моделювання і/або аналізу винятків; це так звані методи розсуджень на основі аналізу прецедентів. Головною проблемою цієї групи методів є ускладненість їх використання на великих обсягах даних, хоча саме при аналізі великих сховищ даних методи ІАД приносять найбільшу користь.

У другому випадку інформація спочатку береться з первинних даних і перетворюється в деякі формальні конструкції (їх вид залежить від конкретного методу). Згідно попередньої класифікації, цей етап виконується на стадії вільного пошуку, яка у методів першої групи в принципі відсутня.

Таким чином, для прогностичного моделювання та аналізу винятків використовуються результати цій стадії, які набагато більш компактні, ніж самі масиви вихідних даних. При цьому отримані конструкції можуть бути "прозорими" (інтерпритованими), або "чорними ящиками" (не трактованими). В обох випадках при обробці застосовують логічні інформаційні одиниці та інші інформаційні одиниці [10].

Особливо широко методи інтелектуального аналізу даних застосовуються у бізнес-додатках аналітиками і керівниками компаній. Для цих категорій користувачів розробляються інструментальні засоби високого рівня, що дозволяють вирішувати досить складні практичні завдання без спеціальної математичної підготовки. Актуальність використрання інтелектуального аналізу даних в бізнесі пов'язана з жорсткою конкуренцією, яка виникла внаслідок переходу від «ринку виробника» до «ринку споживача». В цих умовах особливо важливо якість і обгрунтованість прийнятих рішень, що вимагає строгого кількісного аналізу наявних даних. При роботі з великими обсягами накопичуваної інформації необхідно постійно оперативно відслідковувати динаміку ринку, а це практично неможливо без автоматизації аналітичної діяльності.

Типові завдання для методів інтелектуального аналізу даних наступні: прогнозування, маркетинговий аналіз, аналіз роботи персоналу, аналіз

ефективності продажу товарів поштою, профілювання клієнтів, оцінка по - тенціальних клієнтів, аналіз результатів маркетингових досліджень, аналіз роботи регіональних відділень компанії, порівняльний аналіз конкуруючих фірм.

ІАД не виключає людську участь в обробці та аналізі, але значно спрощує процес пошуку необхідних даних із сирих даних, роблячи його доступним для широкого кола аналітиків, які не є фахівцями в статистиці, математиці або програмуванні. Людська участь виражається в когнітивних аспектах участі і застосування інформаційних когнітивних моделей. Очевидно, що перераховані види завдань актуальні практично для багатьох галузей бізнесу: банківської справи і страхування (виявлення зловживань з кредитними картками, оцінка кредитних ризиків, оцінка закладних, виявлення профілей користувачів, оцінка ефективності регіональних відділень, ймовірність подачі заявки на виплату страховки та ін), фінансових ринків (прогнозування, аналіз портфелів, моделювання індексів), виробництва (прогнозування попиту, контроль якості, оцінка дизайну продукції), торгівлі та інше.

Інтелектуальний аналіз геоданих заснований на data mining, але з урахуванням особливостей просторової інформації і просторових відносин. На першому етапі інтелектуального аналізу геоданих відбувається їх формування або організація.

Організацією даних (рис.1.1) називається процедура відомості різнорідних даних і моделей в єдину несуперечливу інформаційну модель, яку в подальшому можна буде ефективно застосовувати в різних технологіях аналізу і управління. Цю особливу інформаційну модель називають інформаційною основою. Результатом організації даних є створення тільки такої інформаційної моделі, яка дозволяє організувати зберігання в базі даних і володіє синергетичним параметрами. Отже, організація геоданих дає можливості створення для БД і можливість їх автоматизованої обробки.

Геодані для їх використання повинні бути класифіковані, уніфіковані, інтегровані.

Послідовність цих процедур показана на рисунку 1.1. Першим етапом є збір інформації. Він формує так звані первинні дані. Вихідна первинна інформація включає безліч параметрів, багато з яких дублюють один одного. Зменшення числа даних про реальних об'єктах досягається застосуванням різних моделей, що зберігають основні властивості об'єктів дослідження і не містять вторинних властивостей.

Рис.1.1 Загальна схема організації геоданих

Однією з особливостей збирання даних в геоінформатиці є те, що вихідні дані можуть мати не лише різні розмірності, але і вимірюватися в різних шкалах вимірювання. Організація геоданих спрямована на об'єднання даних різних розмірностей і шкал вимірювань в єдину систему даних для їх зберігання і подальшої обробки. Імен - але це створює можливість комплексного аналізу даних, при роботі з різнорідними вихідними даними, виміряними в різних шкалах вимірювання.

Тому наступним етапом є класифікація зібраної інформації, яка служить основою подальших дій. Класифікація даних дозволяє співвіднюсить різні моделі і їх характеристики різними класами, підкласами і типів, що дає можливість систематизувати вихідні набори даних і використовувати властивості класів при аналізі конкретних даних.

Як додатковий етап класифікації геоданих в геоінформатиці є процедура локалізації даних.

Після того, як дані класифіковані, здійснюється їх уніфікація. Різноманітність технологій і методів збору даних породжує різноманітність типів даних, які згодом необхідно обробляти. Обробляти безліч різних даних незручно і неэффективно. Для спрощення процесу обробки, зберігання та обміну різнорідні дані приводять до єдиного структурованого вигляду, який використовується при подальшій обробці інформації. Такі дані називають уніфікованими.

Процедура відомості різнорідних видів і структур даних до єдиного вигляду і структурі називається уніфікацією. В ході уніфікації даних здійснюється побудова єдиної форми даних.

Після цих процедур можлива побудова інтегрованої моделі. Уніфікація не створює систему даних, перетворює вихідну сукупність різнорідних і неузгоджених даних в іншу, але вже більш узгоджену і менш різнорідну.

Для обробки по єдиній технологічній системі і в єдиному інформаційному середовищі моделі повинні бути об'єднані на основі правила або методу, що відповідає вимогам - ям оптимального зберігання і обробки. Таким об'єднуючим методом є інтеграція даних. інтеграція даних і створює систему даних замість сукупності даних

Необхідно відзначити, що геодані утворюють природну інформаційну систему даних [3]. Це обумовлено тим, що вони відображають реальні об'єкти і явища земної поверхні, які розташовані не довільно, а організовано і мають об'єктивні зв'язки один з одним. Можно говорити, що інформація про об'єкти і явищах земної поверхні створює якусь систему. Окремі моделі чи дані є елементами такої системи.

Інтеграцією називають відновлення та (або) підвищення якісного рівня взаємозв'язків між елементами системи, а також процесом з декількох різнорідних систем єдиної системи, з метою виключення (до технічно необхідного мінімуму) функціональної і структурної надмірності і підвищення загальної ефективності функціонування.

Таким чином, інтеграція даних призводить до встановленню додаткових зв'язків між даними і ці зв'язки можна назвати системними. Можно сказати, що саме інтеграція даних призводить до появи геоданих як системи. Можна також сказати, що інтеграція даних дає інтегровану модель геоданих. Інтегрована модель не є просто сумою інформаційних частин її утворюють. Вона, як правило, має менший обсяг фізичної пам'яти при збереженні інформаційної ємності по порівнянні з інформаційними моделями, її складовими, хоча включає дані про зв'язки та додаткову службову інформацію. Крім того, вона включає додаткові зв'язки між вихідними даними, що створює синергетичний ефект. Як наслідок з'являється мож- ливість вирішення більшої кількості завдань, в зокрема комплексного аналізу даних і кореляційного аналізу.

У реальності багато моделі можна віднести до інтегрованим. Тому говорять про ступінь інтеграції, однак іншим важливим параметром є критерій або аспект інтеграції. Він служить основою об'єднання даних у інтегровану модель. Важливою властивістю інтеграції є те, що інтеграція - це не просто об'єднання даних, а придбання цією моделлю додаткових властивостей. В результаті інтеграції даних створюється модель, що володіє додатковими властивостями або, кажучи мовою синергетики, що має синергетичний ефект. Інтегрована модель є розвитком інформаційної моделі. Вона більш складна, але з цієї причини не тільки описує інформаційні властивості об'єкта, але дозволяє проводити ефективну обробку даних, відноситимуться до досліджуваного об'єкта.

Аспект інтеграції пов'язаний з вибором критерію сталого інтеграції. В геоінформатиці є особливість аспекти інтеграції даних. Вона полягає в тому, що геодані розглядають з урахуванням трьох аспектів: просторового, часового та тематичного. Це означає, що дані, зібрані та збережені у базі геоданих (БГД), групують за трьома характеристиками: "місце", "час", "тема".

Дані, які вибирають для інтеграції, повинні бути найбільш стійкими або найменш мінливими. Тимчасові дані по визначенню змінюються і тому не

можуть служити основою інтеграції. Тематичні дані також мінливі, вони можуть змінюватися можуть зникати або з'являтися в нових видах, тому і вони не можуть служити основою інтеграції.

Просторові дані - найбільш стійкі і найменш мінливі, тому в цій групі слід шукати основу для інтеграції. Серед просторових даних найбільш стійкими (найменш мінливими) є координати. Саме вони є основою для об'єднання різних даних, тобто підста - вою для інтеграції.

Характеристика "місце" є найбільш стійкою в системі координат земної поверхні, в той час як характеристики "час" і "тема" є мінливими від об'єкта до об'єкту. Глобальна стійкість характеристики "місце" і послужила основою інтеграції інших видів інформації на цій основі.

Таким чином, якщо локалізація створює сукупність даних з вертикальними зв'язками, то інтеграція створює систему уніфікованих даних з вертикальними і горизонтальними зв'язками. Саме системність організації даних на основі інтеграції забезпечує ефективність аналізу і обробки геоданих як в геоінформатиці, так і в інших наукових напрямах. В результаті інтеграції виходить якась система даних нагадує таблицю або "універсальне відношення" з теорії реляційних баз даних. Працювати з такою однією таблицею незручно і, як випливає з теорії баз даних, її розбивають, використовуючи процедури нормалізації.

Іншими словами, в отриманій системі геоданих доцільно поставити якусь структуру для зручності аналізу та обробки. Для структуризації системи годинних застосовую процес званий стратифікацією. Стратифікація означає розбиття сукупності або системи на частини, звані стратами або шарами.

Стратифікація координатних даних основана на важливій функції координатних моделей відображати просторові властивості об'єктів. Просторові об'єкти характерні тим, що мають графічну форму представлення.

При стратифікації дані організовуються в шари, а шари групуються у відповідності з заданих темами, які відповідають об'єктам. Групування може бути з якоїсь теми, наприклад "транспорт" або "підземні комунікації".

Самий нижній шар називають елементним. Він розбиває геодані на три просторових типу. Це дані ареальних - А, лінійні -Л, точкові Т. Далі шари групуються відповідно до заданих темами, які відповідають об'єктам.

Таким чином, стратифікація це не просто структуризація геоданих, а створення інструменту аналізу і узагальнення даних на різних територіальних або адміністративно-територіально - матеріальних рівнях.

Крім того, стратифікація перетворює геодані в унікальний інформаційний ресурс. В цілому геодані можна розглядати як систему даних. Але на нижньому рівні стратифікації геодані постають у вигляді інформаційних одиниць [12]. Це дає можливість організації геоінформаційного моделювання з рівня інформаційних одиниць на глобальний рівень.

Особливістю геоданих є наявність динамічного зв'язку між графічними даними і атрибутивними даними. Зміна атрибутивних даних тягне за собою автоматичну заміну графічної інформації. Це створює добру основу для просторового аналізу і управління. Геодані організовують з урахуванням семіотичного підходу, а саме у вигляді семантичної, синтаксичної та прагматичної частин.

Семантична частина містить інформацію про об'єктах і спосіб її кодування. Синтаксична частина включає правила побудови моделей об'єктів і спосіб їх віднесення до класу відомих моделей. Прагматична частина визначає цінність інформації або дає можливість оцінити її. При відсутності будь-якого з цих трьох частин інформаційна модель геоданих не придатна для використання.

Класи систем ІАД

При інтелектуальному аналізі геоданих вони аналізуються на тих же системах. що і звичайні дані. ІАД є мульти-дисциплінарною областю, яка виникла та розвиток на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних та інше [7]. Звідси велика кількість методів і алгоритмів, реалізованих у різних діючих системах ІАД.

Багато з таких систем інтегрують в собі відразу кілька підходів. Тим не менш, як правило, в кожній системі є якась ключова компонента. Наведемо класифікацію зазначених ключових компонент з короткою характеристикою для кожного класу.

Індустріальні системи ІАД. В справжій час більшість провідних у світі виробників програмного забезпечення пропонують свої продукти та рішення в області ІАД. Як правило - це масштабовані системи, в яких реалізовані різні математичні алгоритми аналізу даних. Вони мають розвинутий графічний інтерфейс, багаті можливості візуалізації та маніпулювання з даними, надають доступ до різних джерел даних.

Предметно-орієнтовані аналітичні системи. Предметно-орієнтовані аналітичні системи дуже різноманітні. Ці системи вирішують вузький клас спеціалізованих завдань. Найбільш широкий підклас таких систем, що одержав поширення в області дослідження фінансових ринків, носить назву "технічний аналіз". Він являє собою сукупність декількох десятків методів прогнозу динаміки цін і вибору оптимальної структури інвестиційного портфеля, заснованих на особистих емпіричних моделях динаміки ринку.

Статистичні пакети. Це потужні математичні системи, призначені для статистичної обробки даних будь-якій природі. Вони включають численні інструменти статистичного аналізу, мають розвинені графічні засоби. Головний недолік систем цього класу - їх неможливо ефективно застосовувати для аналізу даних, не маючи глибоких знань в області статистики. Непідготовлений користувач повинен пройти спеціальний курс навчання.

Штучні нейронні мережі. Це широкий клас різноманітних систем, які представляють собою ієрархічні, мережні структури, у вузлах яких знаходяться так звані нейрони. Мережі тренуються на прикладах, і у багатьох випадках дають хороші результати прогнозів. Основними недоліками нейронних мереж є необхідність мати дуже великий обсяг навчальної вибірки, а також труднощі в інтерпретації результатів. Тренована нейрона мережа являє собою "розумний чорний ящик", роботу якого неможливо зрозуміти і

контролювати пакети, засновані на деревах рішень. Дерева рішень є одним з найбільш популярних підходів до вирішення задач ІАД. Цей метод використовується тільки для рішення задач класифікації. Це є його серйозним обмеженням. Результатом роботи методу є ієрархічна деревоподібна структура класифікаційних правил типу "IF...THEN...". Перевагою методу є природна здатність класифікації на безліч класів.

Системи міркувань на основі аналогічних випадків. Для того щоб зробити прогноз на майбутнє або вибрати правильне рішення, ці системи знаходять у минулому близькі аналоги наявної ситуації і вибирають ту саму відповідь, яка була для них правильною. Тому цей метод ще називається методом "найближчого сусіда". Ці системи показують дуже гарні результати в найрізноманітніших задачах.

Генетичні алгоритми. Строго кажучи, інтелектуальний аналіз даних - далеко не основною областю застосування генетичних алгоритмів, які, скоріше, потрібно розглядати як потужний засіб вирішення різноманітних комбінаторних задач та задач оптимізації. Тим не менш, генетичні алгоритми увійшли зараз у стандартний інструментарій методів ІАД. Цей метод названий так тому, що в якійсь мірі імітує процес природного добору в природі. Еволюційне програмування. В даній системі гіпотези про вигляді залежності цільової змінної від інших змінних формулюються у вигляді програм на деякій внутрішній мові програмування. Процес побудови програм будується як еволюція в світі програм (цим підхід трохи схожий на генетичні алгоритми).

Причиною зростання популярності ІАД є об'єктивність одержуваних результатів. Людині-аналітику, на відміну від машини, завжди притаманний суб'єктивізм, він у тій чи іншій мірі є заручником вже складених уявлень. Іноді це корисно, але частіше приносить велику шкоду. Геодані є одним з багатьох універсальних засобів аналізу просторових об'єктів і явища й інструментом пізнання навколишнього світу. Вони застосовуються не тільки в геоінформатиці, але і в інших наукових напрямках, включаючи штучний інтелект [8]. Проблема інтелектуального аналізу геоданих зводиться до

вирішення ряду проблем. Однак організація геоданих призводить до створення інтегрованої системи даних, що включає систему моделей і систему інформаційних одиниць. Це визначає геодані як унікальний інформаційний ресурс який застосовують в освіті і науці на виробництві для отримання нових знань.

Отже, хочеться відзначити той факт, Data Mining відносяться до дорогих програмних технологій — ціна деяких з них доходить до декількох десятків тисяч доларів. Тому до недавнього часу основними споживачами цієї технології були банки, фінансові та страхові компанії, великі торгові підприємства, а основними завданнями, що вимагають застосування Data Mining, вважалися оцінка кредитних і страхових ризиків і вироблення маркетингової політики, тарифних планів та інших принципів роботи з клієнтами. За допомогою використання Data Mining зараз ГІС-системи розвиваються небаченими темпами і відносяться до числа найбільш цікавих в комерційному плані рішень. В наші дні їх розробкою і впровадженням зайняті велика кількість різних організацій, що дозволяє говорити про конкуренцію із західними виробниками. За новими технологіями - величезні перспективи, засновані на подальшому розвитку комп'ютерних засобів обробки інформації.

РОЗДІЛ 2. Структура інформаційної системи ГІС

2.1 Характеристика джерела даних для інформаційного сховища

Розподілена обробка даних обов'язково передбачає наявність банків і баз даних. Проте база даних — це не місце, куди просто складають дані: ними потрібно користуватися, актуалізувати, змінювати формати і зв'язку і здійснювати безліч інших дій. Якщо безсистемно наповнювати базу інформацією, то через деякий час нею неможливо буде користуватися — часу на пошук потрібних даних буде йти все більше і більше, простір бази переповниться. У зв'язку з цим дані необхідно «очищати» і структурувати, а для ефективної роботи з ними потрібні системи управління роботою баз даних (Data Base Management System — СУБД). Індустрія створення баз даних і СУБД бере свій початок в 1960-е рр. і до теперішнього часу достатньо розвинена, проте термін «сховище даних» в сучасному його розумінні з'явився відносно недавно. Ідея сховищ даних виявилася затребуваною, так як у багатьох видах державної, ділової, наукової, соціальної діяльності необхідні тематично об'єднані та історично очищені сукупності даних. При цьому постійно зростала потреба в більш дешевих, точних і структурованих даних, а також більшої оперативності отримання, обробки та інтегрування даних.

До кінця 1980-х рр., коли була в повній мірі усвідомлена необхідність інтеграції корпоративної інформації та належного управління цією інформацією, з'явилися технічні можливості для створення відповідних систем, які спочатку були названі «сховищами інформації» (Information Warehouse). Лише в 1990-е рр., з виходом книги Білла Инмона, сховища отримали своє нинішнє найменування «сховища даних» (Data Warehouse — DW).

Інмон визначив сховища даних як предметно-орієнтовані, інтегровані, незмінні, підтримують хронологію набори даних, організовані для цілей підтримки управління, покликані виступати в ролі єдиного і єдиного джерела

істини, що забезпечує менеджерів і аналітиків достовірною інформацією, необхідною для оперативного аналізу та прийняття рішень [13].

В основі концепції сховищ даних лежать три основні ідеї:

1) інтеграція раніше роз'єднаних деталізованих даних (історичні архіви, дані з традиційних систем обробки документів, розрізнених баз даних, дані із зовнішніх джерел) в єдиному сховищі даних;

2) тематичне та тимчасове структурування, узгодження та агрегування;

3) поділ наборів даних, що використовуються для операційної (виробничої) обробки, і наборів даних, що застосовуються для вирішення завдань аналізу.

Дані, що поміщаються в сховище, повинні відповідати певним вимогам: предметної орієнтованості, інтегрованості, підтримки хронології та незмінюваності (табл. 2.1).

Таблиця 2.1 Вимоги до даних, який поміщають у сховище

Вимога	Характеристика
Предметна орієнтованість	Всі дані про деякої сутності (бізнес-об'єкті) з деякої предметної області збираються з різних джерел, очищаються, узгоджуються, доповнюються, агрегуються і подаються до єдиної, зручної для їх використання в бізнес-аналізі формі
Інтегрованість	Всі дані про різні бізнес-об'єктах взаємно узгоджені і зберігаються в єдиному общекорпоративном сховище
Підтримка хронології	Дані хронологічно структуровані і відображають історію за період часу, достатній для виконання завдань бізнес-аналізу, прогнозування та підготовки прийняття рішення
Незмінюваність	Вихідні (історичні) дані, після того як вони були узгоджені, верифіковані та внесені в загальнокорпоративний сховище, залишаються незмінними і використовуються виключно в режимі читання

Сховище даних виконує безліч функцій, але його основне призначення — надання точної інформації в найкоротші терміни і з мінімумом витрат. Для успішного ж просування Web-середовища електронного бізнесу потрібно, щоб доступ до інформації був недорогим і не займав багато часу.

Поняття «сховище даних» в первісному розумінні було засноване на понятті «розподіленої вітрини даних» (Distributed Data Mart — DDM). Внаслідок цього в класичному виконанні сховище даних було насамперед репозиторієм (наскрізний базою даних) інформації підприємства. Середа сховища була призначена тільки для читання і складалася з детальних і агрегованих даних, які повністю очищені і інтегровані [14]. Крім того, у сховищі зберігається велика і детальна історія даних на рівні транзакцій. З точки зору архітектурного рішення таке сховище даних реалізує свої функції через підмножину залежних вітрин даних (рис. 2.1).

Перевагами архітектури класичного сховища даних є:

- несуперечність інформації;
- один набір процесів вилучення і бізнес-логіки використання;
- загальна семантика;
- централізоване кероване середовище;
- легко створюються за шаблонами та наповнювані вітрини даних;
- єдиний репозиторій метаданих;
- різноманіття механізмів обробки і представлення даних.

До недоліків можна віднести великі витрати з реалізації, високу ресурсомісткість в масштабі всього підприємства, потреба в складних сервісних системах, ризикований сценарій розвитку, коли всі дані та металані знаходяться в одному репозиторії і у несприятливому випадку можуть бути втрачені.

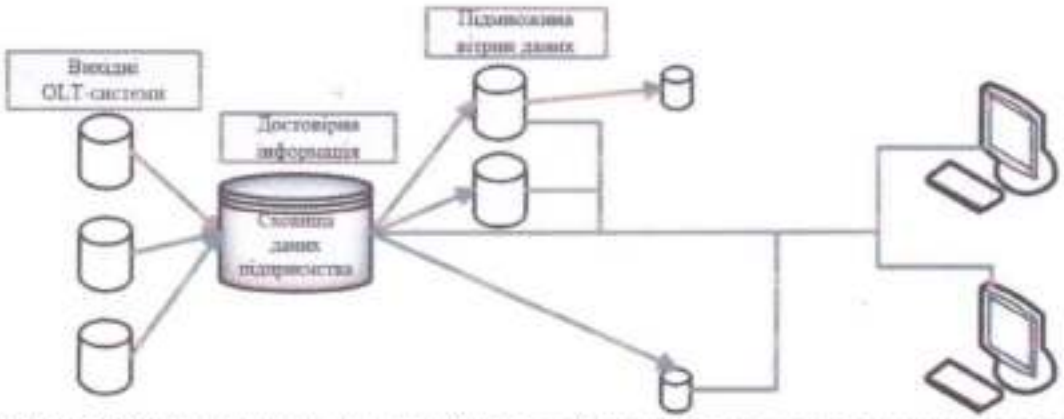

Рис.2.1 Сховище даних, що реалізує свої функції через підмножину залежних вітрин даних

Крім цього, при фільтрації і рафінуванні «сирих» даних для такого сховища зазвичай втрачається дуже багато інформації, яка може бути надзвичайно корисною при бізнес-аналізі. У зв'язку з цим виникло розуміння того, що сховище, крім механізмів вилучення даних (On-Line Transactional Processing — OLTP), репозиторію і вітрин, повинно мати відповідне простір для організації «сирих» даних і їх багатовимірного аналізу в режимі реального часу OLAP.

2.2 Аналіз системи візуалізації даних

У сучасному світі, де інформація набуває виняткове важливе значення, при цьому стаючи все більш відкритою широкої аудиторії, постає питання про пошук форм її подання, максимально доступних для сприйняття. З великої кількості різнорідної інформації, що надходить до нас сьогодні з різних джерел, часом буває важко виділити головне, а в потоці даних легко загубитися [15]. Графічна інформація сприймається в кілька разів швидше, ніж текстова. Крім того, людське сприйняття графічних образів асоціативно. Бачачи черговий рекламний плакат, ми мимоволі починаємо проводити паралелі, в нашій підсвідомості виникають асоціації з чимось уже знайомим. З розвитком сучасних ЗМІ, особливо інтернет-видань та social media, набирає популярність інфографіка — особливий вид представлення інформації, даних і знань у графічному форматі, основними перевагами якого є швидкість розуміння, наочність і доступність (рис. 2.2).

Рис.2.2 Вигляд інфографіки, із продемонстрованим співвідношенням

Для цільової аудиторії інфографіка є ефективним засобом подання статистичної інформації, комерційних звітів, бізнес-планів, аналітики та ін Далі в статті будуть розглянуті програмні засоби, що дозволяють використовувати графічні елементи візуалізації даних у веб-додатках[16].

Класифікація різновидів графічного подання

Графіки, як одні з найбільш простих елементів, що використовуються для відображення залежності одного набору даних від іншого. Графіки бувають декількох видів (рис.2.3). Лінійні представляють собою об'єднані лінією набори точок, відповідних значень по осях. Графіки розсіювання показують розподіл сукупності точок, відповідних значень по осях.

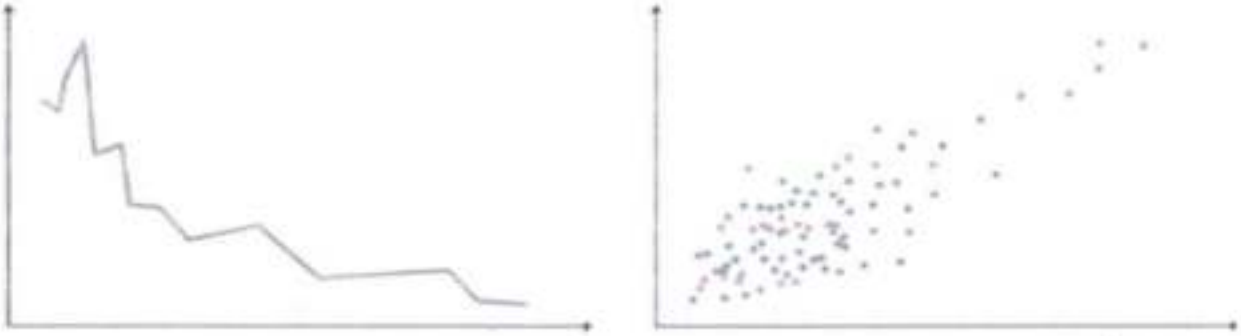

Рис. 2.3 Лінійний графік (ліворуч), графік розсіювання (праворуч)

Потрібно сказати, що графіки є підтипом діаграм. Крім цього, діаграми бувають стовпчастими (гістограми), круговими, кільцевими, пелюстковими (рис. 2.4) і т. д. Циклічні діаграми показують ключові кроки процесу, який містить набір повторюваних дій.

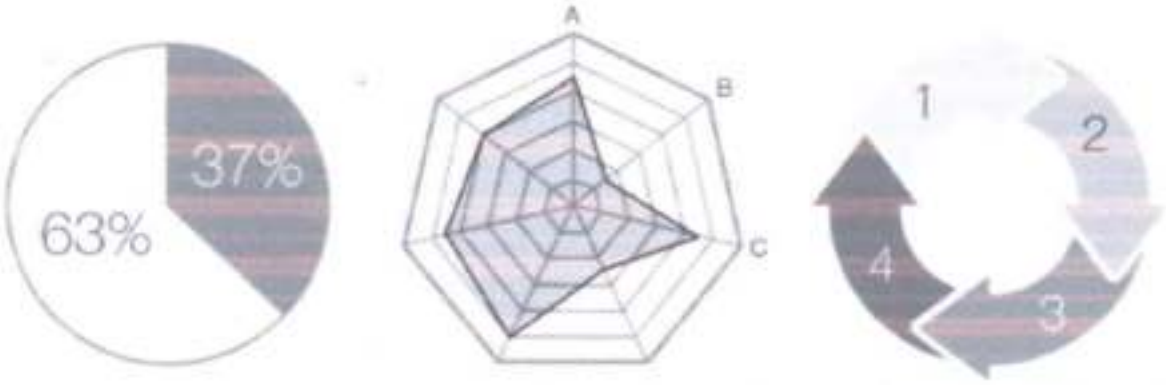

Рис.2.4 Діаграми: секторна діаграма, пелюсткова, циклічна (зліва направо)

Окремо варто відзначити такі типи діаграм, як теплова карта і плоске дерево. Теплова карта порівнює значення всередині набору даних, зафарбовуючи їх одним з кольорів спектру. Плоске дерево, що представляє ієрархію набору даних, в якій елементи є батьківськими або дочірніми по відношенню один до одного, відображається у вигляді набору вкладених прямокутників, кожний з яких є гілкою (рис. 2.5).

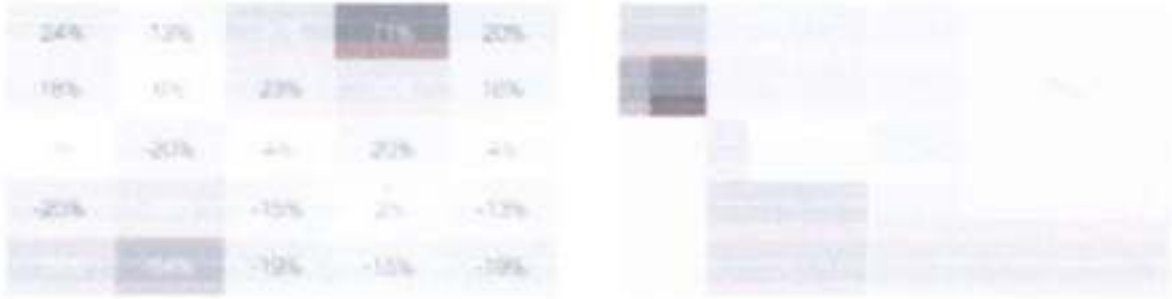

Рис. 2.5 Теплова карта (ліворуч), плоске дерево (праворуч)

Дерева і діаграми зв'язків дещо відрізняються від згаданих вище уявлень. Вони покликані показати структуру (ієрархію) набору даних, відобразити взаємозв'язок окремих складових. Ці види подання широко використовуються для візуалізації таких математичних структур, як графи та мережі. Візуальне відображення графа являє собою сукупність вузлів, з'єднаних між собою лініями — ребрами. Дерева і ментальні карти є приватними випадками графів (рис.2.6).

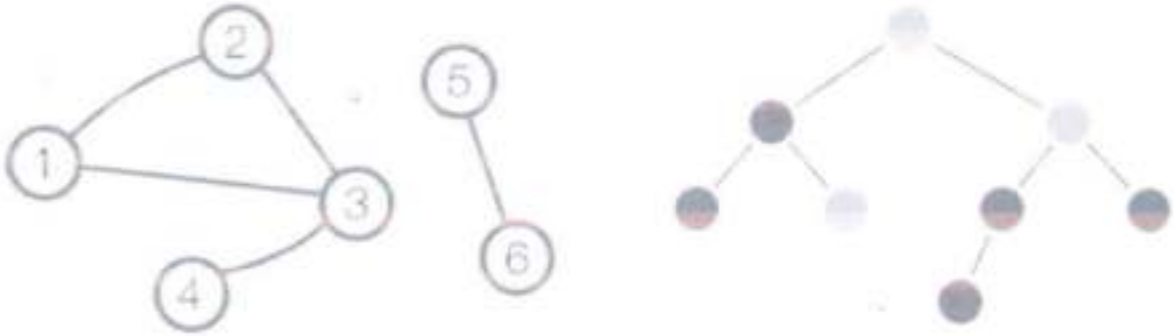

Рис. 2.6 Граф (ліворуч), дерево (праворуч)

Ще одним споживаним інструментом візуалізації є карти. Карти здатні в наочному вигляді уявити рівень безробіття по регіонах, позначити області, що постраждали від паводків і т. д. Іншими словами, карти і картограми дають змогу швидко виявити географічні і часові закономірності, чого не можна досягти за допомогою лише зведених таблиць з даними. Крім того, за допомогою тематичних маркерів на карті можна показати розташування будь-яких об'єктів, наприклад велосипедних парковок або офісів компаній.

2.3 Структура інформаційного сховища

ІХ являє собою базу узагальненої інформації, формується з безлічі зовнішніх і внутрішніх джерел, на основі якої виконуються статистичні групування та інтелектуальний аналіз даних.

В основі ІХ лежить поняття багатовимірного інформаційного простору або гіперкуба, в комірках якої зберігаються аналізовані числові показники () Вимірами (осями) гіперкуба є ознаки аналізу. При зберіганні ознаки аналізу відокремлюються від фактичних даних.

Основними складовими структури сховищ даних є таблиця фактів (fact table) і таблиці вимірювань (dimension tables).

Таблиця фактів є основною таблицею сховища даних. Як правило, вона містить відомості про об'єкти або події, сукупність яких буде надалі аналізуватися. Зазвичай говорять про чотирьох найбільш часто зустрічаються типи фактів [17]. До них відносяться:

• факти, пов'язані з транзакціями (Transaction facts). Вони засновані на окремих подіях (типовими прикладами яких є телефонний дзвінок або зняття грошей з рахунку за допомогою банкомату);

• факти, пов'язані з «моментальними знімками» (Snapshot facts). Засновані на стані об'єкта (наприклад, банківського рахунку) в певні моменти часу, наприклад на кінець дня або місяця. Типовими прикладами таких фактів є обсяг продажу за день або денна виручка;

• факти, пов'язані з елементами документа (Line-item facts). Засновані на тому чи іншому документі (наприклад, рахунку за товар або послуги) і містять детальну інформацію про елементи цього документа (наприклад, кількість, ціну, відсоток знижки);

• факти, пов'язані з подіями або станом об'єкта (Event or state facts). Представляють виникнення події без подробиць про нього (наприклад, просто факт продажу або факт відсутності такої без інших подробиць).

Таблиця фактів, як правило, містить унікальний складовою ключ, що об'єднує первинні ключі таблиць вимірів. Найчастіше це цілочисельні значення або значення типу «дата/час» — адже таблиця фактів може містити сотні тисяч або навіть мільйони записів, і зберігати в ній повторювані текстові описи, як правило, невигідно — краще помістити їх на менші за обсягом таблиці вимірів. При цьому як основні, так і деякі неключевые поля повинні відповідати майбутнім вимірам OLAP-куба [18]. Крім цього таблиця містить одне або кілька числових полів, на підставі яких в подальшому будуть отримані агрегатні дані. Таблиця фактів, яка може бути побудована на основі бази даних Northwind, наведено на рис. 2.7.

TimeKey	CustomerKey	ShipperKey	ProductKey	EmployeeKey	RequiredDate	LineItemFreight	LineItemTotal	LineItemQuantity	LineItemDiscount
85	4	11	5	01.08.1996	14.3904	168	12	0	
85	4	42	5	01.08.1996	11.992	98	10	0	
85	4	72	5	01.08.1996	5.996	174	5	0	
79	1	14	6	16.08.1996	2.1321	167.4	9	0	
79	1	51	6	16.08.1996	9.476	1696	40	0	
34	2	41	4	05.08.1996	10.971	77	10	0	
34	2	51	4	05.08.1996	38.3985	1404	35	222.6	
34	2	65	4	05.08.1996	16.4565	252	15	37.8	
84	1	22	3	05.08.1996	6.0492	100.8	6	5.04	
84	1	57	3	05.08.1996	15.123	234	15	11.7	
84	1	65	3	05.08.1996	20.164	336	20	0	
76	2	20	4	06.08.1996	19.54	2592	40	129.6	
76	2	33	4	06.08.1996	12.2125	50	25	2.5	
76	2	60	4	06.08.1996	19.54	1088	40	0	
34	2	31	3	24.07.1996	11.404	200	20	0	

Рис.2.7 Таблиця фактів

В даній таблиці, відповідні дані майбутнього куба відповідають перші шість полів, а агрегатним даними — останні чотири.

Зазначимо, що для багатовимірного аналізу придатні таблиці фактів, що містять як можна більш докладні дані (тобто відповідні членам нижніх рівнів ієрархії відповідних вимірювань). В даному випадку краще взяти за основу факти продажу товарів окремим замовникам, а не суми продажів для різних країн — останні все одно будуть вирахувані OLAP-засобом. Виняток можна зробити, мабуть, тільки для клієнтських OLAP-засобів, оскільки в силу ряду обмежень вони не можуть маніпулювати великими обсягами даних.

Відзначимо, що в таблиці фактів немає ніяких відомостей про тому, як групувати записи обчислення агрегатних даних. Наприклад, у ній є ідентифікатори продуктів чи клієнтів, але відсутня інформація про те, до якої категорії відноситься даний продукт або в якому місті знаходиться даний клієнт. Ці відомості, в подальшому використовуються для побудованийія иерархий у вимірах куба, містяться в таблицях вимірювань.

Таблиці вимірювань містять незмінні або рідко змінювані дані. У переважній більшості випадків ці дані представляють собою по одному запису для кожного члена нижнього рівня ієрархії в вимірі. Таблиці вимірювань також містять як мінімум одне описове поле (зазвичай з ім'ям члена вимірювання) і, як правило, ціле ключове поле для однозначної ідентифікації члена вимірювання. Якщо майбутнє вимір, засноване на даній таблиці

вимірювань, містить ієрархію, то таблиця вимірювань також може містити поля, що вказують на «батьків» цього члена в цій ієрархії. Нерідко (але не завжди) таблиця вимірювань може містити поля, що вказують на «прабатьків», та інших «предків» у даній ієрархії (зазвичай це характерно для збалансованих ієрархій), а також додаткові атрибути членів вимірювань, що містилися у вихідній оперативній базі даних (наприклад, адреси та телефони клієнтів).

Кожна таблиця вимірювань повинна знаходитися у відношенні «один-до-багатьох» з таблицею фактів [19].

Відзначимо, що швидкість росту таблиць вимірів повинна бути незначною в порівнянні зі швидкістю зростання таблиці фактів; наприклад, додавання нового запису в таблицю вимірювань, що характеризує товари, виробляється тільки при появі нового товару, що не продавався раніше. Структура таблиці вимірювань наведена на рис.2.8.

Рис. 2.8 Таблиця вимірювань

Один вимір куба може міститися як в одній таблиці (в тому числі і при наявності декількох рівнів ієрархії), так і в кількох зв'язаних таблиць, що відповідають різним рівням ієрархії в вимірі. Якщо кожне вимірювання міститься в одній таблиці, така схема сховища даних носить назву «зірка» (star schema). Приклад такої схеми наведено на рис. 2.9.

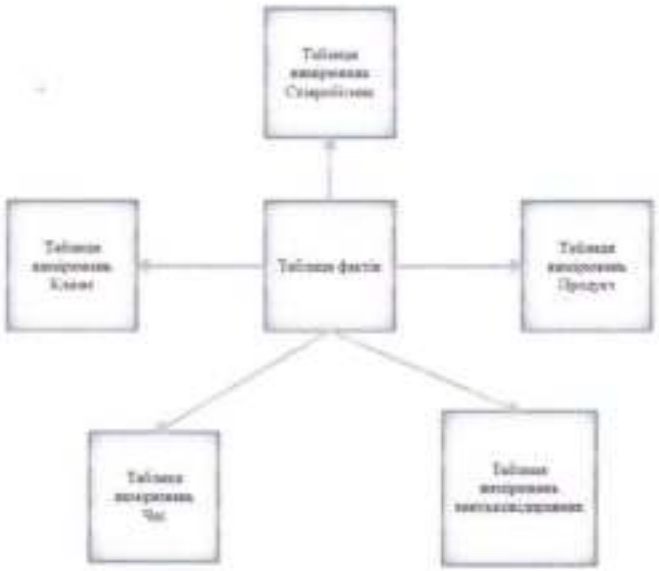

Рис.2.9 Приклад схеми «зірка»

Якщо ж хоча б один вимір міститься в кількох пов'язаних таблицях, така схема сховища даних носить назву «сніжинка» (криву schema). Додаткові таблиці вимірювань в такій схемі, зазвичай відповідні верхнім рівням ієрархії вимірювання, перебувають у співвідношенні «один до багатьох» з головною таблицею вимірювань, що відповідає нижньому рівню ієрархії. Приклад схеми «сніжинка» наведений на рис.2.10.

Рис.2.10 Приклад схеми «сніжинка»

Відзначимо, що навіть при наявності иерархических вимірювань з метою підвищення швидкості виконання запитів до сховища даних нерідко перевага віддається схемою «зірка».

Говорячи про вимірювання, слід згадати про те, що значення, можуть мати різні рівні деталізації.

Отже, таким чином, дані, занурені в сховище, організуючись в інтегровану цілісну структуру, що володіє природними внутрішніми зв'язками, набувають нових властивостей, що надають їм статус інформації.

Вони є основою для побудови аналітичних систем і систем підтримки прийняття рішень. Саме тому технології інформаційних сховищ орієнтовані на керівників, відповідальних за прийняття рішень.

РОЗДІЛ 3. Реалізація підсистеми аналітичної обробки даних

3.1 Створення джерела даних

Методи інтелектуального аналізу інформації часто розглядають як природний розвиток концепції сховищ даних. Головна відмінність сховища від бази даних полягає в тому, що їх створення і експлуатація переслідують різну мету. База даних відіграє роль помічника в оперативному управлінні організацією. Це щоденні задачі отримання актуальної інформації: бухгалтерські звітності, облік договорів, тощо. Сховище даних накопичує всі необхідні дані для здійснення задач стратегічного управління в середньостроковому і довгостроковому періоді. Наприклад, продаж товару і генерація рахунку проводяться з використанням бази даних, а аналіз динаміки продажів за декілька років, що дозволяє спланувати роботу з постачальниками - за допомогою сховища даних.

Сховище даних (Data Warehouse) - це систематизована інформація з різнорідних джерел, яка є необхідною для обробки з метою ухвалення стратегічно важливих рішень.

Сховище будується на основі клієнт-серверної архітектури, СУБД і утиліт підтримки прийняття рішень. Дані, що надходять у сховище, стають доступні тільки для читання.

Властивості сховища даних:

- предметна орієнтація (інформацію організовано відповідно до основних аспектів діяльності);
- інтегрованість даних (дані в сховище надходять з різних джерел і відповідно агрегуються);
- стабільність, інваріантність у часі (записи в DW ніколи не змінюються, являючи собою відбитки даних, зроблені у певний час);

- мінімізація збитковості інформації (перед завантаженням у сховища дані фільтруються, зберігаються у певній послідовності, а також формується деяка підсумкова інформація).

В сховищах даних надмірність даних є мінімальною (приблизно 1%), оскільки:

- при завантаженні у сховище дані сортуються і фільтруються;
- інформація у сховищах зберігається в хронологічному порядку, що майже повністю виключає перекриття даних;
- при завантаженні у сховище дані зводяться до єдиного формату, включаючи обчислення підсумкових (агрегованих) показників.

Сервери багатовимірних баз даних можуть зберігати дані по-різному, крім агрегованих показників формується ще й додаткова інформація: поля часу, дати; адресні посилання, таблиці метаданих тощо. Це приводить до значного збільшення інформації. Вхідний масив розміром 200 Mb може розростись до об'єму 5 Gb. Сховище даних повинне бути оптимально організованою базою даних, яка забезпечує максимально швидкий і оперативний пошук інформації.

Вітрина даних - це спрощений варіант сховища даних, що містить лише тематично орієнтовані, агреговані дані.

Глобальне сховище даних складається з трьох рівнів:

- сховище агрегованих даних;
- вітрини даних, які базуються на інформації зі сховища даних;
- клієнтські робочі місця, на яких встановлено засоби оперативного аналізу даних.

У розпорядженні виробників прикладних програмних засобів є три різні технології роботи з базами даних:

- DAO (Data Access Objects) - доступ до локальних баз даних;
- RDO (Remote Data Objects) - доступ до віддалених баз даних;
- ADO (ActiveX Data Objects) - доступ до Widows-додатків через Інтернет. В основному використовується з міркувань безпеки.

Одним з перспективних напрямів удосконалення доступу до даних є гнучке конфігурування системи, коли розподіл між клієнтською і серверною частинами можливий за допомогою використання механізму віддалених процедур. Поряд з потоками даних існують і потоки метаданих, які розміщуються в депозитарії. Він дає змогу визначити семантичну структуру додатка у вигляді опису термінів предметної галузі, їхні взаємозв' язки й атрибути [20]. Метадані - це дані про дані, які визначають джерело, приймач та алгоритм трансформації даних під час перенесення їх від джерела до приймача.

Метадані містять:

- описи структур даних та їхніх взаємозв'язків;
- інформацію про джерела даних і про ступінь їх вірогідності;
- інформацію про власників даних, права доступу;
- схему перетворення стовпців вхідних таблиць у стовпці кінцевих таблиць;
- правила підсумовування, консолідації та агрегування даних;
- інформацію про періодичність оновлення даних;
- каталог використаних таблиць, стовпців та ключів;
- фізичні атрибути стовпців;
- кількість табличних рядків та обсяг даних;
- часові ярлики (дата та час створення/модифікації записів);
- статистичні оцінки часу виконання запитів.

Контроль модифікації (versioning) полягає у властивості метаданих відслідковувати зміни в структурі даних та їх значення в часі.

Функціональна архітектура сховища даних містить наступні компоненти:

- сховище даних;
- клієнтська частина системи (дизайнери сховища, засоби розробки додатків, засоби адміністрування, інструменти аналізу даних, завантаження

словника метаданих з XML-файлу у сховище і експорт його зі сховища в XML-файл;

- сервер обміну даними (Data Exchange Server) - набір програм імпорту/експорту даних зі сховища й каталогів для організації обміну даними із зовнішніми OLTP-системами;
- бібліотеки прикладних класів: ACL (Application Class Library), VCL (Visual Component Library), Win Lite.

Наповнення інформаційних сховищ відбувається в декілька етапів:

- екстракція (витяг) - імпорт даних у сховище з інформаційних підсистем, виробничих відділів та інших джерел;
- трансформація - консолідування, агрегування даних, розбиття їх на фракції, коригування та трансформування у відповідні формати;
- завантаження - у сховище, синхронізація з датою або зовнішніми подіями.

Обслуговування інформаційних сховищ полягає в: копіюванні баз даних, налаштуванні, тиражуванні, надсиланні застарілих баз даних до архіву, управлінні правами користувачів, створенні та редагуванні графічних діаграм баз даних, тощо [9].

Типи архівації у сховищах поділяють на:

- звичайна;
- копіювальна;
- додаткова;
- диференціальна;
- щоденна.

Архівні магнітні носії зберігають у вогнетривких сейфах або за межами обчислювального центру. Крім того, розробляється план архівації компонентів сервера баз даних. Сучасні сервери автоматично підтримують копію свого каталогу на кожному сервері вузла. Цей процес називається реплікацією каталогів (directory replication).

Звичайна архівація каталогів на всіх серверах здійснюється раз на тиждень у вихідні дні, а диференціальна - щодня в робочі дні. У річному архіві, як правило, зберігаються дані останнього тижня місяця. Усі зміни в каталозі сервера, а також в особистих і загальних сховищах записуються у файли, які називаються журналами трансакцій (transaction log files).

Під час виконання додаткової архівації каталогу або інформаційного сховища архівуванню підлягають лише журнали трансакцій.

Для ефективної роботи зі сховищем даних, необхідно зібрати максимум інформації про процес. Наприклад, для прогнозування обсягів продажів можуть бути використані бази даних облікових систем компанії, маркетингові дані, відгуки клієнтів, дослідження конкурентів і т.п.

Для оцінки діяльності організації використовується інформація:

- хронологія продажів;
- стан складу на кожний день - якщо спад продажів буде пов'язаний із відсутністю товару на складі, а не через відсутність попиту;
- відомості про ціни конкурентів;
- зміни у законодавстві;
- загальний стан ринку;
- курс долара, інфляція;
- відомості про рекламу;
- відомості про відношення до продукції клієнтів;
- різного роду специфічну інформацію. Наприклад, для продавців морозива - температуру, а для фармакологічних складів - санітарно-епідеміологічний стан, тощо.

Проблема полягає в тому, що зазвичай в системах оперативного обліку більша частина цієї інформації відсутня, а наявна - неповна або спотворена. Кращим варіантом в цьому випадку буде створення сховища даних, куди б з певною заданою періодичністю надходила вся необхідна інформація, заздалегідь систематизована і очищена рис.3.1.

Рис.3.1 Структура сховища даних

Ефективна архітектура сховища даних організовується таким чином, щоб бути складовою частиною інформаційної системи управління підприємством.

Найбільш поширений випадок, коли сховище організовано за типом "зірка", де в центрі розміщуються факти і агрегатні дані, а "променями" є виміри. Кожна "зірка" описує певну дію, наприклад, продаж товару, його відвантаження, надходження коштів й інше:

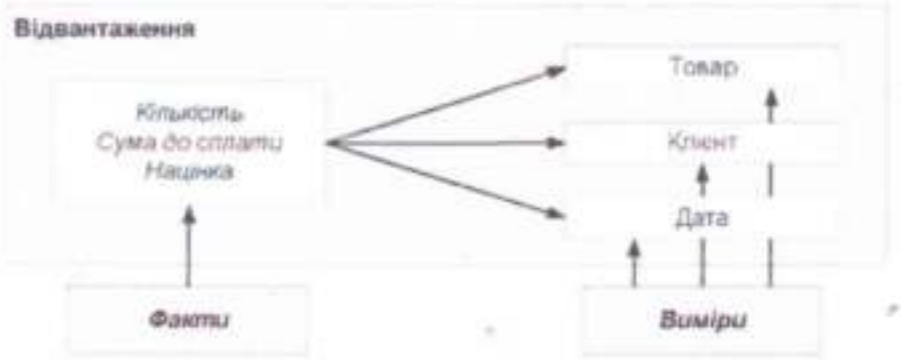

Рис.3.2 Схема організації сховища даних за типом "зірка"

Як правило, дані копіюються в сховище з оперативних баз даних і інших джерел відповідно до певних правил.

3.2 Етапи представлення джерела даних

Велику роль при аналізі даних відіграє володіння деякими спеціальними способами подання отриманих даних в наочній — короткій і схематизованій — формі.

Табличний спосіб зображення даних дозволяє представити якісні та кількісні дані з коротким супровідним пояснюючим текстом. Таким текстом служать назва таблиці, що розкриває зв'язок між числовими рядами, і внутрішні заголовки таблиці (вказують вимірювані ознаки, місце, час, одиниці виміру і т.п.).

Матриця являє собою різновид таблиці з рядками і рядами (стовпцями), що мають будь-які функціонально-логічні зв'язки. В результаті в матриці виявляється наявність або відсутність зв'язку між різними факторами педагогічного процесу.

Графіки ще більш наочно, ніж таблиці, відображають зміну експериментальних даних. Графіки будуються в прямокутній системі координат, в якій на осі "X" відзначається значення вибірки, а по осі "Y" — значення, порядок ознаки, частота події.

Композиція графіка - це поєднання всіх його елементів. Графік повинен привертати увагу, забезпечуючи в той же час легкість його прочитання і засвоєння. Важливим завданням композиції графіка є художня та естетична сторона його оформлення.

Правила побудови графіків:

1. Необхідно провести ретельний відбір з наявного цифрового статистичного матеріалу тих даних, які будуть зображені на графіку (далеко не всі отримані дані слід зображати графічно).

2. Вибрати той вид графіка, який на думку дослідника найбільш яскраво буде відображати отримані дані.

3. Назва графіка має бути ясним і повним, що відображає зміст і мають при необхідності особливі пояснення.

4. Написи і легенда розташовуються, як правило, в нижній або правій частині діаграми.

5. Цифри шкали слід наносити зліва і знизу або вздовж осей. Горизонтальну шкалу (по осі абсцис) необхідно будувати зліва направо, вертикальну (по осі ординат) - знизу вгору.

6. Якщо числові дані не включені в діаграми, бажано їх представити поруч в табличній формі.

7. Густота координатної сітки повинна бути оптимальною, що не утрудняє читання графіка.

8. Допускається кількість різних кольорів на графіку не більше трьох.

9. Якщо графіки відображають серію спостережень, рекомендується ясно позначати всі точки, відповідні окремим спостереженням.

Види графіків:

1. Лінійний графік-передає зміни в деяких мірних числах, наприклад, зміна середніх оцінок контрольних робіт, проведених в одному класі протягом навчального року.

2. Гістограма являє собою різновид графіка в якому по осі "Y" відкладаються частотні (інтервальні) значення будь-якої угруповання, в результаті чого графік стає "ступінчастим".

3. Полігон частот - на базі полігону частот будується гістограма, різниця між ними полягає в тому, що в Полігоні частота інтервалу зведена до його центру, а при гістограмі частоти зображують рівномірно в межах всього інтервалу.

4. Кумулятивний графік частоти – накопичувальний розподіл частоти) - частота окремих інтервалів сукупності розглядається кумулятивно, тобто до частоти кожного інтервалу додаються частоти всіх попередніх інтервалів.

5. Діаграми зіставляють кількісну інформацію у вигляді площ різних фігур (коло, прямокутник, сектор, циліндр, бульбашки та ін.).

Графи - особливий вид графічного відображення даних результатів; це фігура, що складається з точок (вершин), з'єднаних відрізками (ребрами).

Вершини графа можуть позначати різні компоненти педагогічного процесу, параметри, фактори, а ребра — відносини і зв'язки між ними. Графи (як моделі) часто застосовуються на етапі прогнозування експерименту, а на узагальнюючому етапі з ними зіставляються результати. Найпростішим прикладом графа служить "дерево" цілей.

3.3 Реалізація завдань візуалізації

В Excel 2016 Power Map вбудувана в програму і називається 3D-карта (3D-map). В якості вихідних даних, оберемо вихидні ресурси із статистики України. А саме - обсяги та середні ціни культур зернових та зернобобових, реалізованих підприємствами у січні-листопаді 2020 року (рис 3.3), і обсяги та середні ціни насіння культур олійних, реалізованих підприємствами у січні-листопаді 2020 року рис 3.4.

Обсяги та середні ціни культур зернових та зернобобових, реалізованих підприємствами у січні-листопаді 2020 року[1]

	Культури зернові та зернобобові, усього				[illegible]			
	реалізовано		середні ціни реалізації		реалізовано		середні ціни реалізації	
	т	у % до відповідного періоду 2019	грн за т	у % до відповідного періоду 2019	т	у % до відповідного періоду 2019	грн за т	у % до відповідного періоду 2019
Україна	**30871698,3**	**97,2**	**4891,4**	**119,9**	**14287429,1**	**91,8**	**4976,7**	**121,6**
Вінницька	2271884,2	73,2	4750,4	119,2	827284,4	85,6	5058,4	121,2
Волинська	527866,1	88,1	4673,7	117,8	312808,4	89,0	4862,0	118,1
Дніпропетровська	1848412,8	104,5	4801,0	124,5	1098175,7	115,9	4870,0	125,3
Донецька	948869,3	108,2	4859,3	123,2	747287,8	102,6	4931,0	123,7
Житомирська	1286502,0	77,8	4582,8	118,2	309701,7	80,1	5029,1	128,2
Закарпатська	87454,2	139,1	4183,4	111,2	13184,8	129,8	4113,4	120,3
Запорізька	1598272,7	88,1	4881,4	120,0	1072778,1	84,5	5002,2	121,3
Івано-Франківська	474808,3	98,2	4573,9	118,6	233800,8	120,6	4682,9	124,2
Київська	2437489,8	78,5	4547,5	116,5	817842,8	84,8	4977,4	120,2
Кіровоградська	1872448,0	83,8	4720,2	119,2	864888,9	88,9	5181,8	122,4
Луганська	790728,0	121,8	4554,2	127,5	617971,5	102,7	4682,9	127,9
Львівська	623481,0	85,7	4759,4	118,3	343554,5	88,1	4858,9	120,2
Миколаївська	1184390,2	89,0	4823,8	117,0	810108,4	82,8	5183,8	118,9
Одеська	941128,2	77,1	4941,1	113,6	469017,0	48,9	5142,1	114,8
Полтавська	3525447,2	93,5	4784,4	119,1	803137,5	88,0	4857,8	118,1
Рівненська	448152,3	67,1	4686,7	121,4	188728,5	81,5	4772,8	118,8
Сумська	2884934,2	92,4	4728,3	120,1	612303,8	101,2	4839,3	121,2
Тернопільська	1227897,3	82,1	4803,3	123,4	551855,8	89,8	4980,0	121,8
Харківська	2468741,1	111,9	4782,3	125,1	1512594,1	113,2	4881,7	124,2
Херсонська	1228885,3	113,4	4898,8	121,1	708554,7	102,8	5066,5	123,1
Хмельницька	1891247,1	78,8	4758,6	120,3	729418,5	78,0	5085,5	124,2
Черкаська	2034813,6	72,7	4828,2	117,8	907754,4	82,9	5119,6	120,0
Чернівецька	113598,8	74,8	4577,5	121,4	41988,8	81,8	4418,7	122,7
Чернігівська	3698091,4	88,7	4428,5	117,6	319851,8	90,0	4977,8	125,2
м. Київ	727482,2	99,0	4423,8	111,8	226189,6	91,5	5049,2	123,0

Рис 3.3 Статистика обсягів та середні ціни культур зернових та зернобобових, реалізованих підприємствами у січні-листопаді 2020 року

Рис.3.4 Обсяги та середні ціни насіння культур олійних, реалізованих підприємствами у січні-листопаді 2020 року

Вихідні дані краще зберігати у вигляді таблиці Excel, а не в звичайному діапазоні. Це дозволить автоматично підхоплювати нові значення при їх додаванні.

Для наочної візуалізації даних візьмемо з таблиці обсяги та середні ціни культур зернових та зернобобових, реалізованих підприємствами у січні-листопаді 2020 року, реалізацію тон Культур зернових та зернобобових та окремо реалізацію тон пшениці які належать до певних міст України. Так само візьмемо процентне співвідношення культур зернових і зернобобових і так само співвідношення реалізації пшениці в порівнянні з 2019 роком.

Аналогічно будемо брати дані з таблиці обсягів і середніх цін насіння культур олійних, реалізованих підприємствами в січні-листопаді 2020 року.

Активуємо будь-яку клітинку, та у вкладці Вставка - натискаємо 3D-карта рис.3.5.

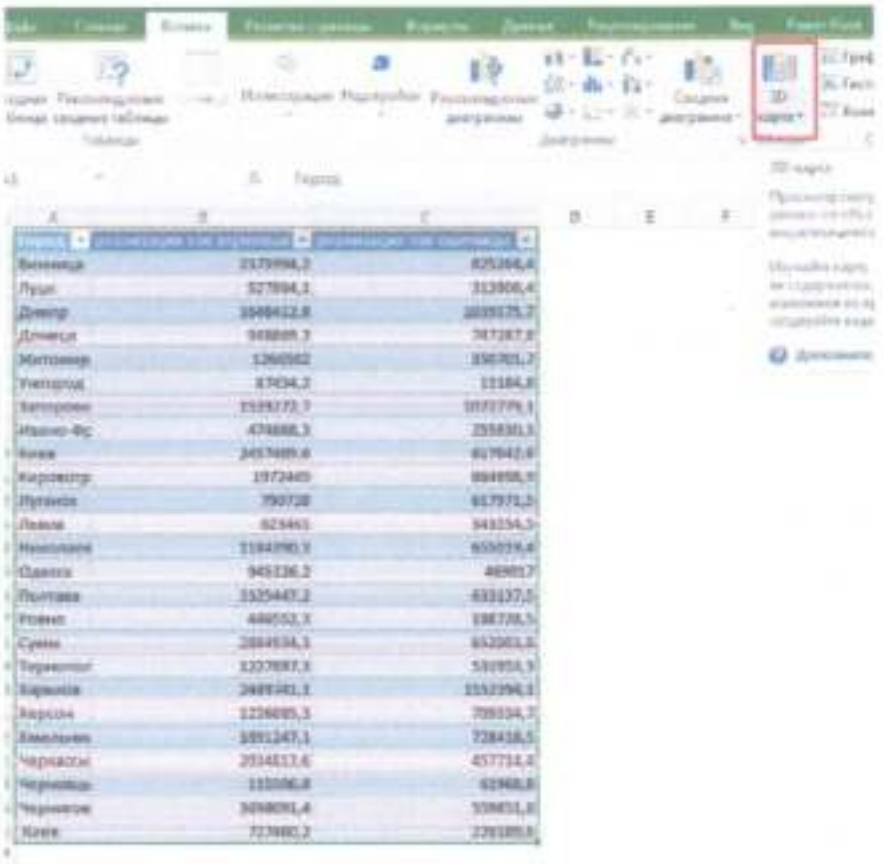

Рис.3.5 3D - карта

Відкривається вікно створення і редагування 3D-карти, рис. 3.6.

Рис 3.6 Вигляд 3D - карти

У верхній частині вікна програми знаходиться стрічка, що складається з однієї вкладки - Головна. В цій вкладці знаходяться команди додавання, видалення, редагування. Зліва знаходиться панель зі сценами карт.

Праворуч панель керування полями даних, за якими малюється карта. У центрі ми бачимо карту поточної сцени (у нас вона поки тільки одна) з усіма чинними налаштуваннями. За замовчуванням це глобус. По ньому можна переміщатися, як в навігаторі. Для цього використовується миша або клавіатура. Справа внизу кнопки навігації.

Для додавання даних передбачена панель праворуч, рис.3.7.

Рис. 3.7 Панель керувани даними

Підемо зверху вниз. На карті, можна використовувати кілька шарів. Кожен шар показує певну інформацію і настроюється незалежно. Шар можна перейменувати в більш зрозумілу назву, ніж Шар 1, Шар 2 і т. д. Шари можна приховувати, редагувати, видаляти (див. піктограми праворуч від назви). Якщо потрібно створити ще один шар, натискаємо Додати шар.

Нижче в розділі Дані вибирається тип карти. Рівні значень можуть відображатися у вигляді стовпчиків, стовпчиків з накопиченням, стовпчиків поруч, бульбашок, теплової та регіональної карти.

Потім йдуть поля для заповнення карти.

Розташування. У цій області указується поле з географічною ознакою (геолокація). Excel всіляко намагається нам допомогти і сам вибирає стовпець, де знаходиться географічний ознака.

Рис.3.8 Поле з географічною ознакою

Для Excel важливим є тип географічної змінної. Це можуть бути країни, міста, адреси, координати та ін. В нашому простому прикладі Excel сам привласнив тип Країна/регіон.

Ті назви, які Excel розпізнав, відразу відзначаються на карті (якщо сам показник ще не вибрано, то у вигляді точок), а трохи вище області Розташування показаний відсоток розпізнаних назв, рис.3.9.

Рис.3.9 Розташування показаний відсоток розпізнаних назв

Отже, зорієнтувалися на місцевості. Переходимо до даних для візуалізації.

Висота. У цій області вибирається показник, який потрібно відобразити на карті (або кілька). Виберемо реалізацію тон пшениці.

Так як за замовчуванням в якості типу обрана гістограма (з накопиченням), то на карті (глобусі) з'являться стовпці, рис. 3.10.

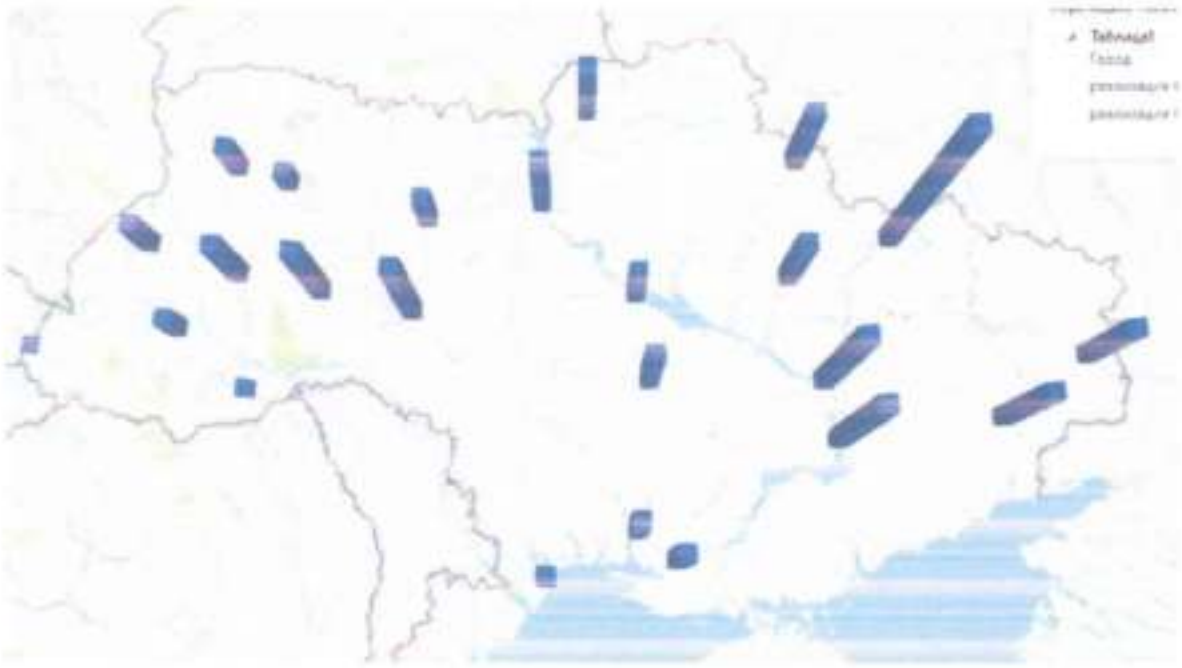

Рис.3.10 Обрана гістограма (з накопиченням)

Взагалі кажучи, 3d-гістограма спотворює реальні пропорції. Тим більше, якщо дивитися зверху. Якщо раптом вам потрібні саме вертикальні стовпчики, то хоча б змінити кут огляду.

Вихідні дані не завжди представлені у вигляді компактної (зведеної таблиці. Джерелом може бути база даних, де кожен рядок відображає операцію і, відповідно, багато ознак повторюються, у тому числі і географічні. В цьому випадку значення потрібно аргументувати певним способом. Як правило, використовується підсумовування. Задається в переліку, що відкривається близько назви (як у зведеній таблиці), рис.3.11.

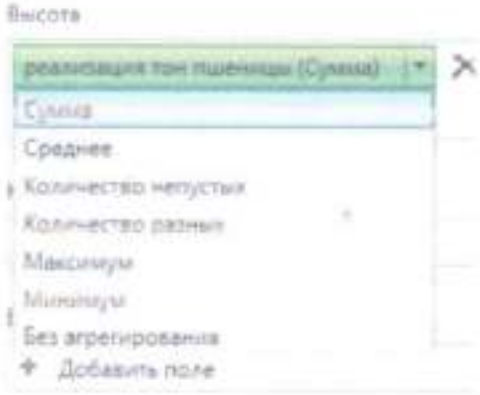

Рис. 3.11 Використання підсумовування

Іншими словами, всі значення обраного показника, що відповідають певній країні, підсумовуються, а результат показується на карті. Як видно із списку, то можна агрегувати, розраховуючи середнє, мінімум, кількість та ін.

Отже, з даними визначилися, виглядають вони за замовчуванням. Зробимо їх більш наочними. Універсальних правил тут, звичайно, немає, оскільки наочність залежить від самих даних. Для зрозумілого відображення показників, будемо використовувати плоску карту. Для заміни глобуса на карту використаємо кнопку Плоска карта, рис. 3.12.

Рис.3.12 Заміна глобуса на карту

Після її натискання, глобус розгорнеться в звичайну карту, рис. 3.13.

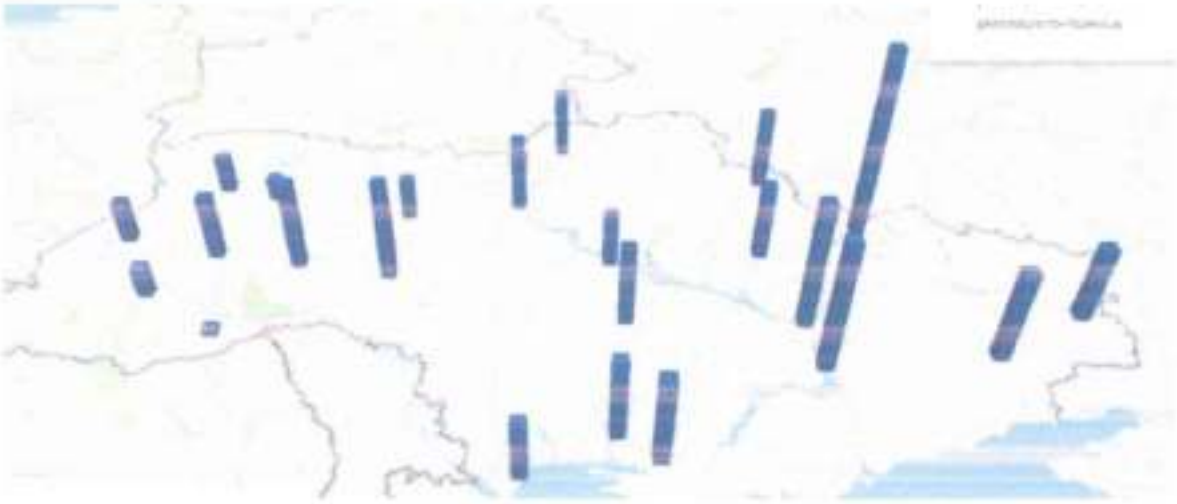

Рис.3.13 Вигляд плоскої карти

Наступний момент – це підписи. На стрічці є спеціальна кнопка.

Рис.3.14 Вибір підписів

На жаль, регулювати підписи неможливо. Вони додаються у відповідності з масштабом. Тому слід обережно користуватися цією опцією.

Додамо в область Висота дані про реалізацію тон зернових культур, рис 3.15.

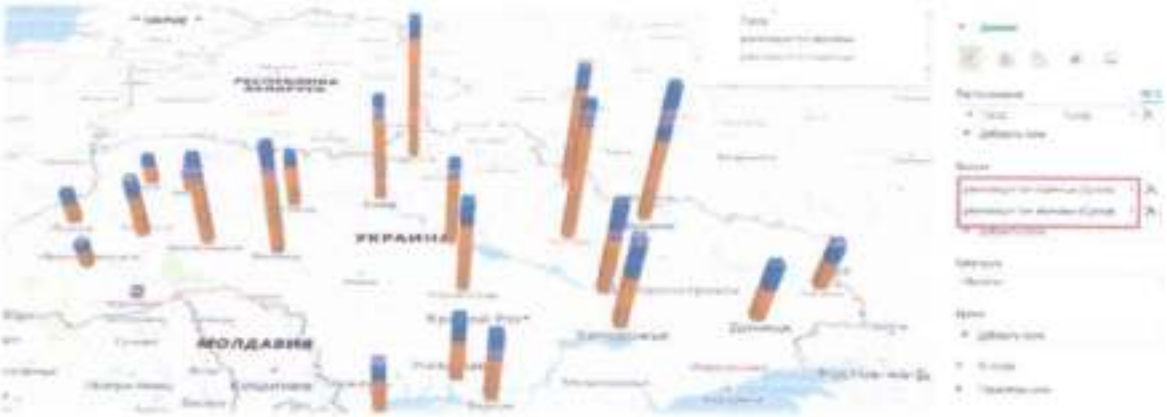

Рис.3.15 Додання показників

Так як використовується гістограма з накопиченням, то два показника просто склалися, що не підходить для відображення данних. Тому виберемо гістограму з угрупованням, рис 3.16.

Рис.3.16 Гістаграма з угруповуванням

Отримаємо наступний вигляд, рис.3.17.

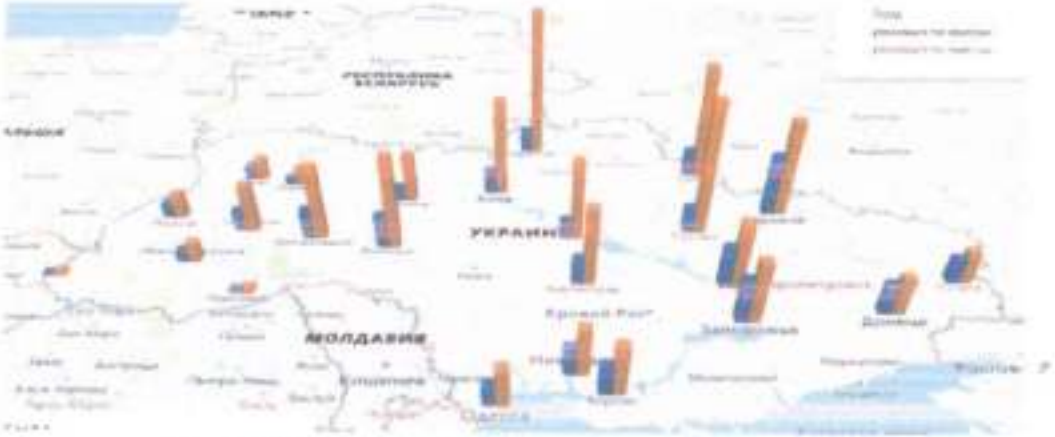

Рис.3.17 Вигляд гістаграми з угруповуванням

Для гарного розуміння зображених даних, додамо на карту плоску діаграму, рис.3.18.

Рис.3.18 Плоска діаграма

Діаграма поміщається прямо на карті.

Наступна візуалізація – бульбашкова діаграма. Діаметр кружка відповідає значенню показника, рис.3.19.

Рис. 3.19 Бульбашкова діаграма

Анологічні дії робимо з іншими обраними показниками, і відображаємо їх на карті, рис.3.20.

Город	реализация семян олейных в соотношении с 19 годом	реализация в соотношении с 19 годом бобов
Винница	674660,3	116134,1
Луцк	250483,9	56456,7
Днепр	836122,9	6510,2
Донецк	449510,7	0
Житомир	379860,6	105864,8
Ужгород	25195,3	21332,1
Запорожье	701656,9	28136,9
Ивано-Фр	225461,3	50884,8
Киев	741260,1	201362,8
Кировогр	781288,1	38752,2
Луганск	518698,7	0
Львов	271548,6	95897,3
Николаев	516308,6	8361,5
Одесса	434587,1	2303,7
Полтава	759191,2	155907,6
Ровно	155335,2	55477,7
Сумы	721888,7	159310,2
Тернополг	454904,7	119499
Харьков	955771,5	25328,5
Херсон	508197,2	157152,6
Хмельниц	763389,1	246156,7
Черкассы	317625,1	63191,4
Черновцы	98933,2	37138,5
Чернигов	684214,2	66841,2
Киев	256089,5	109374

Рис.3.20 Показники реалізації насіння культур олійних, та бобів

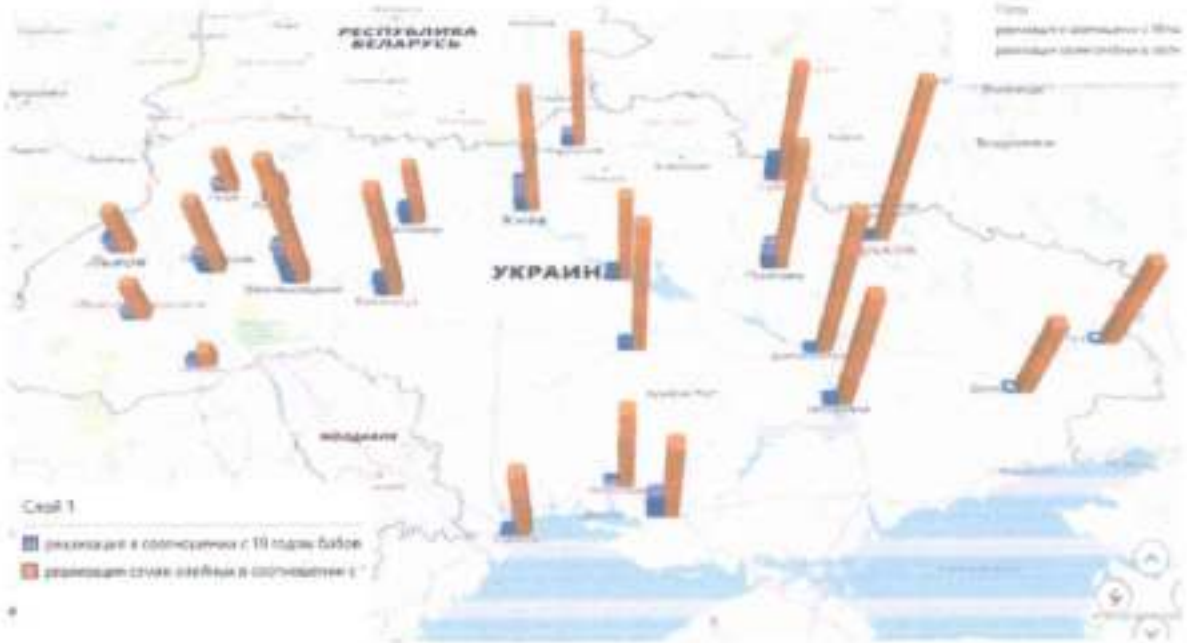

Рис.3.21 Вигляд гістаграми з угруповуванням

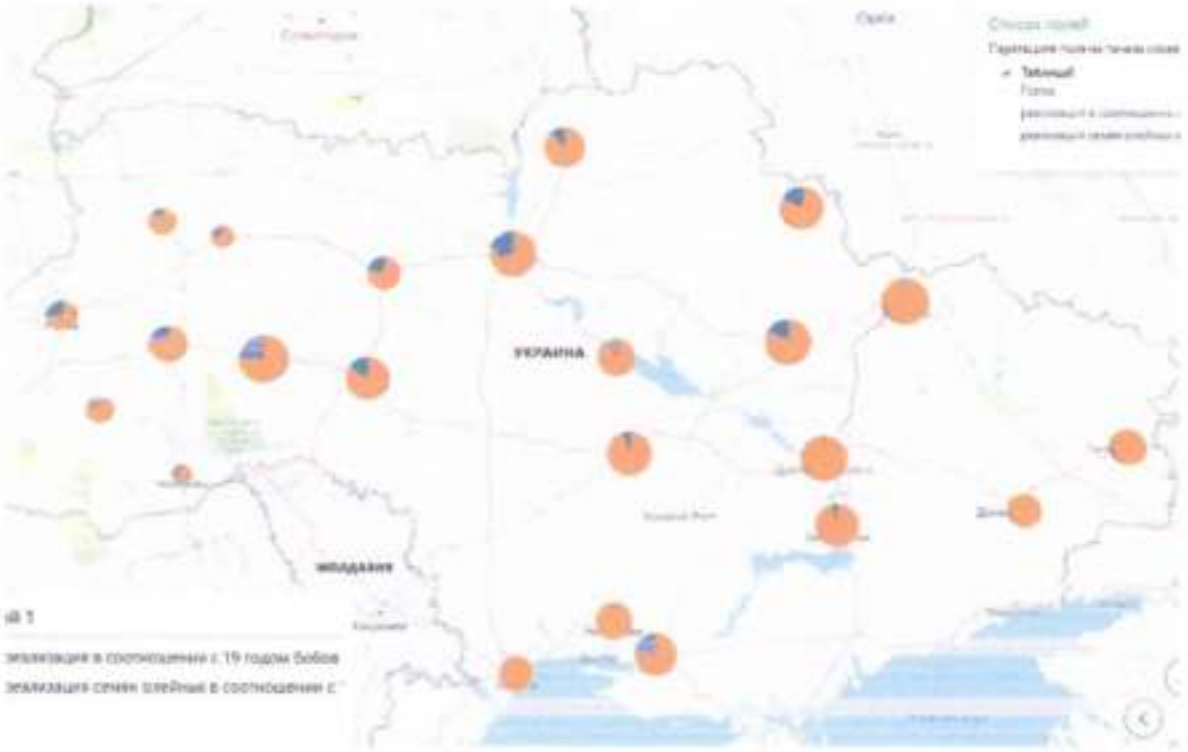

Рис. 3.22 Бульбашкова діаграма

Винница	73,2	80,6
Луцк	88,1	89
Днепр	104,3	110,9
Донецк	105,5	102,6
Житомир	77,8	80,1
Ужгород	139,1	156,6
Запорожье	99,1	94,5
Ивано-Франков	98,2	120,6
Киев	76	94,9
Кировоград	83,8	86,9
Луганск	101,8	102,7
Львов	80,7	88,1
Николаев	89	82,8
Одесса	57,1	48,9
Полтава	95,3	98
Ровно	65,1	81,2
Сумы	92,4	101,2
Тернополь	92,1	96,9
Харьков	111	113,2
Херсон	110,4	102,9
Хмельницкий	75,8	79
Черкассы	72,7	82,9
Черновцы	74,9	61,8
Чернигов	88,7	90
Киев	99	91,3

Рис.3.23 Показники реалізації насіння зернових культур, та пшениці

Рис.3.24 Вигляд гістаграми з угруповуванням

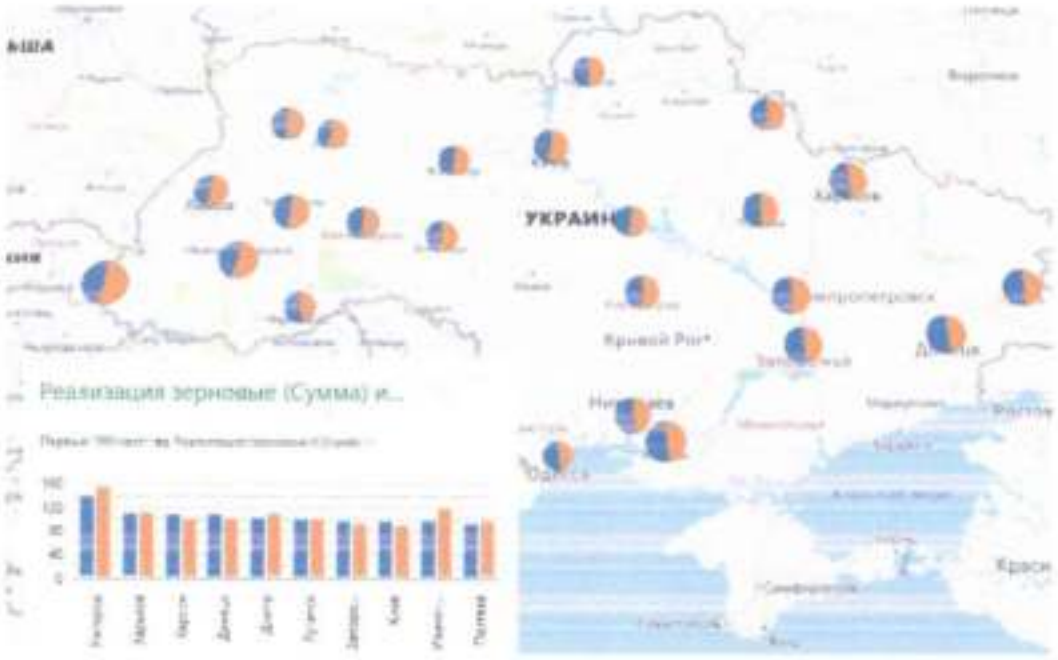

Рис. 3.25 Бульбашкова діаграма

Город	Реализация Семян культур, из них бобы сои
Винница	70,5
Луцк	99
Днепр.	79,5
Донецк	88,9 к
Житомир	80,5
Ужгород	82,4
Запорожье	100,6
Ивано-Франковск	100,8
Киев	78,8
Кировоград	85,1
Луганск	107,8 к
Львов	93,4
Николаев	75,9
Одесса	60,8
Полтава	90,2
Ровно	68,9
Сумы	93,1
Тернополь	100,7
Харьков	98
Херсон	76,1
Хмельницкий	92,7
Черкассы	78
Черновцы	121,4
Чернигов	83,9
Киев	82,5

47,5
97,4
55,3
53,3
87,6
187,1
99,6
61,4
54,9
99,3
94,5
65,3
77,2
69,2
83,2
95,8
67,6
72,8
88,1
45,7
181,1
37,3
76,6

Рис.3.26 Обсяги насіння культур олійних, реалізованих підприємствами

Рис.3.27 Вигляд гістаграми з угруповуванням

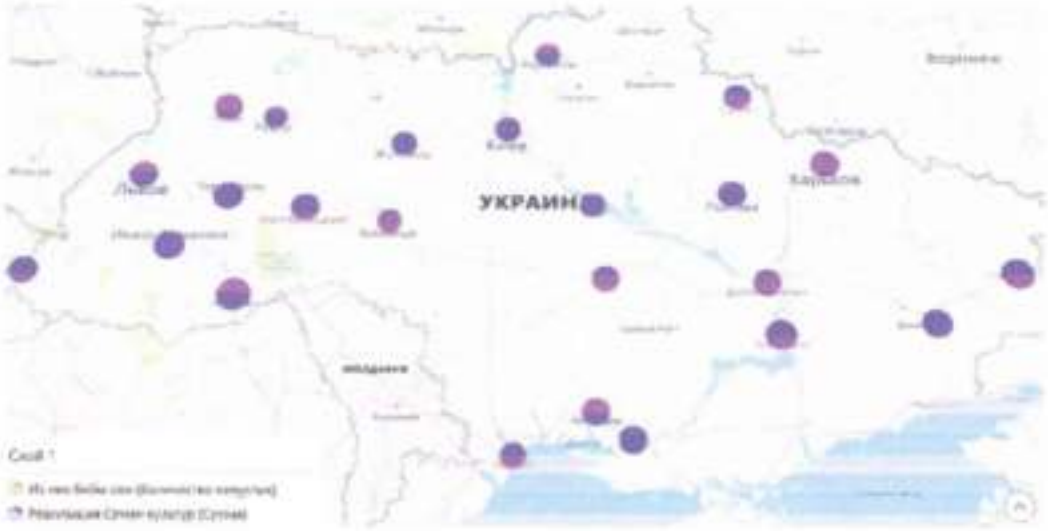

Рис.3.28 Бульбашкова діаграма

Із наочного прикладу спостерігається те, що технологія ГІС дає користувачеві набагато більші можливості для створення і обробки картографічної інформації, які в багатьох випадках не передбачені традиційними методами паперових технологій.

На екран виводиться кілька вікон з різними тематичними картами для їх спільного візуального аналізу; електронні карти легко масштабуються з можливістю автоматизованої генералізації; спеціальні засоби редагування дозволяють швидко змінювати підписи, умовні позначення та загальне розташування картографічного зображення. Найбільш компактними і звичним способом подання географічної інформації залишаються карти.

Таблиці і графіки, що включають різні характеристики об'єктів (атрибути) або їх співвідношення, можуть використовуватися як самостійні або додаткові до інших засобів візуалізації.

Візуальна інформація краще сприймається і дозволяє швидко й ефективно донести до глядача власні думки та ідеї. Дуже зручно використовувати для відображення різних статистичних даних як в розглянутому прикладі вище - сільського господарства, для презинтацій, доповідей, так само і для кращого аналізу і сприйняття великих даних.

Фізіологічно, сприйняття візуальної інформації є основною для людини. Із численних досліджень, виявляються факти, які підтверджують, що:

- 90% інформації людина сприймає через зір;
- 70% сенсорних рецепторів знаходяться в очах;
- близько половини нейронів головного мозку людини задіяні в обробці візуальної інформації;
- на 19% менше при роботі 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 запам'ятовуються докладні деталі візуальної інформації.

Отже, як підсумок із всього вище сказаного, візуалізація на відміну від табличних або текстових звітів допомагає: швидко "читати" інформацію, обробляти її, приймати усвідомлені рішення, підкріплені даними.

Візуалізація допомагає фахівцям правильно організовувати і аналізувати інформацію: діаграми, схеми, малюнки, карти пам'яті сприяють засвоєнню великих обсягів інформації, дозволяють легко запам'ятовувати і простежувати взаємозв'язки між блоками інформації; дає можливість зв'язати отриману інформацію в цілісну картину про те чи інше явище або об'єкт; швидко охопити великий обсяг інформації; відтворити і реконструювати різні процеси і події.

ВИСНОВКИ

Отже, Data Mining відносяться до дорогих програмних технологій — ціна деяких з них доходить до декількох десятків тисяч доларів. Тому до недавнього часу основними споживачами цієї технології були банки, фінансові та страхові компанії, великі торгові підприємства, а основними завданнями, що вимагають застосування Data Mining, вважалися оцінка кредитних і страхових ризиків і вироблення маркетингової політики, тарифних планів та інших принципів роботи з клієнтами. За допомогою використання Data Mining зараз ГІС-системи розвиваються небаченими темпами і відносяться до числа найбільш цікавих в комерційному плані рішень. В наші дні їх розробкою і впровадженням зайняті велика кількість різних організацій, що дозволяє говорити про конкуренцію із західними виробниками.

Таким чином, дані, занурені в сховище, організуючись в інтегровану цілісну структуру, що володіють природними внутрішніми зв'язками, набувають нових властивостей, що надають їм статус інформації.

Вони є основою для побудови аналітичних систем і систем підтримки прийняття рішень. Саме тому технології інформаційних сховищ орієнтовані на керівників, відповідальних за прийняття рішень.

Сьогодні візуалізація даних має свій ряд переваг, які допомагають сприймати інформацію. Простіше і швидше зробити висновок аналізуючи графічне представлення, дивлячись на графік, де один зі стовпців або одна з точок взаємодії знаходиться набагато вище всіх інших; більше залученої аудиторії; висока залученість читачів; краще розуміння даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Болбаков Р. Г. аналіз пізнання в науці та освіті. Перспективи науки і освіти, 2014, № 4, С. 15-19.

2. «Візуалізація даних». URL: https://www.owox.ru/blog/articles/data-visualization/ (дата звернення: 7.11.2020).

3. «Методи Data Mining».URL: https://allbest.ru (дата звернення: 5.11.2020)

4. Поляков В.І. «Предметно-орієнтовані економічні інформаційні системи». 2013, 244 с.

5. Соловйов В.В. «Застосування моделі інформаційної ситуації в геоінформатиці наук про Землю». № 01. 54-58 с. 2012.

6. Цвітков В. Я. «Інформаційне поле». Журнал «Life Science». № 11(5). 2014. 551-554 с.

7. Черняк Л. «Великі дані — нова теорія і практика. Відкрита система. СУБД». №10. 2011 .18-25 с.

8. Цвітков В. А., Лобанов А. А. «Великі дані як інформаційний бар'єр». Європейський дослідник, Т. (78), № 7-1, с. 1237-1242.2014.

9. «Державна служба статистики України» URL: http://ukrstat.gov.ua/ (дата звернення 18.11.2020).

10. «Поняття статистики». URL: http://isprom.blogspot.com/2011/12/data-mining-data-mining-data-mining.html (дата звернення: 16.11.2020).

11. «Інформаційні сховища». URL: https://studme.org/263282/informatika/informatsionnye_hranilischa (дата звернення: 16.11.2020).

12. Савіних В.П., Цвітков В. Я. «Геодезія як системний інформаційний ресурс». Том 84, № 9, 826-829 с. 2014.

13. Геоінформаційні системи: навчальний посібник / Л. А. Павленко. – Х.: Вид. ХНЕУ, 2013. –220-245 с.

14. Геоінформаційні системи. – Електронний навчальний посібник / Під ред. Є.М. Крижановського.- Вінниця: ВНТУ, 2014 -121с.

15. Геоінформаційні технології у територіальному управлінні: наук.-практ. конф. 17-18 верес. 2015 р. – Одеса: ОРІДУ НАДУ, 2015 – 98-100 с.

16. Геоінформаційні системи і бази даних: монографія / В. І. Зацерковний, В. Г. Бурачек, О. О. Железняк, А. О. Терещенко. – Ніжин: НДУ ім. М. Гоголя, 2014. – 305 с.

17. Зацерковний В. І. Аналіз стану топографо-картографічного забезпечення як джерела даних для регіональної ГІС / В. І. Зацерковний // Вісник ЧДТУ. Серія "Технічні науки". – 2012. – № 1 (55). – С. 186–193.

18. Харків. нац. ун-т міськ. госп-ва ім. О. М. Бекетова; уклад.: А. А.Євдокімов – Харків: ХНУМГ ім. О. М. Бекетова, 2015. – 19 с.

19. Даценко, Л. Основи геоінформаційних систем і технологій у школах світу. Краєзнавство, географія, туризм. – 2015. – № 46. – С. 15-21.

20. Даценко, Л.М. Викладання основ геоінформаційних систем і технологій у старших класах загальноосвітніх навчальних закладів / Л.М. Даценко / Національне картографування: стан, проблеми та перспективи розвитку: зб. наук. пр. – К.: ДНВП "Картографія", 2015. – Вип. 4. – С. 260-263.