Розвиток і призначення дейтамайнінгу (Data Mining)

- Поняття Data Mining
- Інтеграція OLAP-технологій та ІАД
- Data Mining і сховища даних

Поняття Data Mining

У 70-х роках минулого століття широко застосовувалася практика, коли компанії наймали аналітиків з бізнесу, котрі, використовуючи статистичні пакети подібні SAS і SPSS, виконували аналіз трендів даних і проводили їх кластерний аналіз. Як тільки стало технологічно можливим і доцільним зберігати великі обсяги даних, менеджери виявили бажання самим мати доступ до даних, подібних тим, що генеруються в пам'яті касового апарата роздрібної торгівлі й аналізувати їх. Запровадження штрихових кодів і глобальна гіпертекстова система Інтернету також зробили реальною можливість для компаній збирати великі обсяги нових даних. Однак у зв'язку з цим виникло питання про інструментальні засоби добування корисної інформації з нагромаджених обсягів «сирих» даних. Ці засоби опісля отримали назву «Data Mining» (дейтамайнінг).

Слід зауважити, що протягом багатьох років компанії проводили статистичні дослідження своїх даних. Коли статистик аналізує дані, то він спочатку висуває гіпотезу про можливий зв'язок між певними даними, а потім посилає запит до бази даних і використовує відповідні статистичні методи, щоб довести або спростувати сформульовану гіпотезу. Це підхід називається *«режимом верифікації» («verification mode»).* На противагу йому програмне забезпечення дейтамайнінгу функціонує в *«режимі відкриття» (discovery mode),* тобто виявляє приховані, часто невідомі для користувачів *шаблони (patterns)* зв'язків між даними, а не аналізує наперед створену гіпотезу щодо них.

За останні роки надзвичайно зріс інтерес до дейтамайнінгу з боку ділових користувачів, котрі вирішили скористатися перевагами даної технології для отримання конкурентної переваги в бізнесі (див. http://www.datamining.com/). Зростаюча зацікавленість щодо впровадження дейтамайнінгу (ДМ) у результаті закінчилася появою низки комерційних продуктів, кожен з яких має таку саму назву, описаний низкою подібних елементів, але фактично має неоднакові функціональні можливості й ґрунтується на різних особливих технічних підходах.

Менеджери з інформаційних технологій, що мають завдання підібрати відповідну СППР, часто безпосередньо зустрічаються зі складними питаннями стосовно реагування на потреби бізнес-користувачів через те, що засадні принципи створення дейтамайнінгу набагато складніші, ніж традиційні запити і формування звітів, крім того, вони відчувають підсилений тиск щодо часу реалізації потреб користувачів, тобто користувачі вимагають розробити дейтамайнінг якомога швидше. Проте очевидною перешкодою для розроблення і впровадження в корпораціях рішень з дейтамайнінгу є наявність багатьох різних підходів до нього, що мають свої певні властивості й переваги, у той час як фактично тільки кількома основними методами формуються основи більшості систем ДМ. У цьому контексті важливою є однозначна інтерпретація самого поняття дейтамайнінгу.

Дейтамайнінг (Data mining) — це тип аналітичних додатків які підтримують рішення, розшукуючи за прихованими шабло нами (patterns) інформацію в базі даних. Цей пошук може бути зроблений або користувачем (тобто тільки за допомогою виконання запитів) або інтелектуальною програмою, яка автоматично розшукує в базах даних і знаходить важливі для користувача зразки інформації. Відповіді на інформаційні запити подаються в бажаній для користувача формі (наприклад, у вигляді діаграм, звітів тощо).

Англомовний термін «Data mining» часто перекладається як *«добування даних»; «добування знань»; «добування інформації»; «аналіз, інтерпретація і подання інформації зі сховища даних»; «вибирання інформації із масиву даних».* У даній книзі буде використовуватися як основний термін *«дейтамайнінг»* — україномовна транскрипція початково запровадженого і однозначно вживаного в англомовній літературі терміна «Data mining».

Добування даних — це процес фільтрування великих обсягів даних для того, щоб підбирати відповідну до контексту задачі інформацію. Вживається також термін *«Data surfing»* (дослі дження даних в Інтернеті). Корпорація IBM визначає ДМ, як «процес екстракції з великих баз даних заздалегідь невідомої, важливої інформації, що дає підстави для дій та використання її для розроблення критичних бізнесових рішень». Інші визначення не пов'язують ні з обсягом бази даних, ні з тим, чи використовується підготовлена інформація в бізнесі, але переважно ці умови загальні.

Інструментальні засоби добування даних використовують різноманітні методи, включаючи доказову аргументацію (case-based reasoning), візуалізацію даних, нечіткі запити й аналіз, нейромережі та інші. Доказову аргументацію (міркування за прецедентами) застосовують для пошуку записів, подібних до якогось певного запису чи низки записів. Ці інструментальні засоби дають змогу користувачеві конкретизувати ознаки подібності підібраних записів. За допомогою візуалізації даних можна легко і швидко оглядати графічні відображення інформації в різних аспектах (ракурсах). Ці та інші методи частково були розглянуті раніше, а детальніше будуть розглянуті далі.

Дейтамайнінг як процес виявлення в загальних масивах даних раніше невідомих, нетривіальних, практично корисних і доступних для інтерпретації знань, необхідних для прийняття рішень у різних галузях людської діяльності, практично має нічим не обмежені сфери застосування. Але, насамперед, методи ДМ нині більше всього заінтригували комерційні підприємства, що створюють проекти на основі сховищ даних (Data Warehousing), хоча наявність сховища даних не є обов'язковою умовою здійснення дейтамайнінгу. Досвід багатьох таких підприємств свідчить, що рівень рентабельності від застосування дейтамайнінгу може досягати 1000 %. Наприклад, відомі повідомлення про економічний ефект, за якого прибутки у 40—70 раз перевищували первинні витрати, і становили від 350 до 750 тис. дол. Є відомості про проект у 60 млн. дол., який окупився всього за 4 місяці. Інший приклад — річна економія 700 тис. дол. за рахунок упровадження дейтамайнінгу в мережі універсамів у Великобританії.

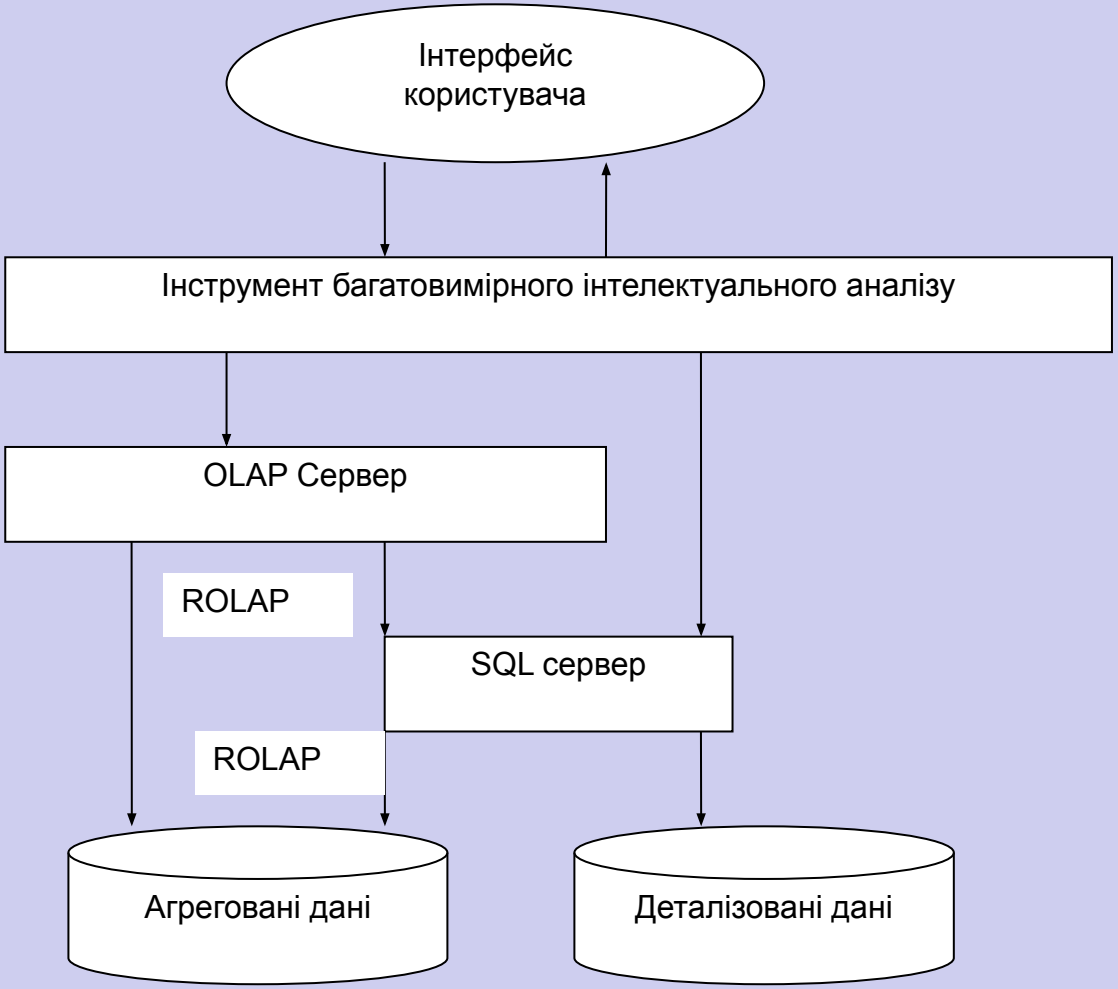
Дейтамайнінг являє собою велику цінність для керівників і аналітиків у їх повсякденній діяльності. Ділові люди"усвідомили, що за допомогою методів ДМ вони можуть отримати відчутні переваги в конкурентній боротьбі.

Інтеграція OLAP-технологій та ІАД

Оперативна аналітична обробка та інтелектуальний аналіз даних - дві складові частини процесу підтримки прийняття рішень. Але сьогодні більшість систем OLAP загострює увагу тільки на забезпеченні доступу до багатовимірних даних, а більшість засобів ІАД, що працюють у сфері закономірностей, мають справу з одновимірними перспективами даних. Ці два види аналізу повинні бути тісно об'єднані, тобто системи OLAP повинні фокусуватися не тільки на доступі, але і на пошуку закономірностей. Як відмітив N. Raden, "багато компаній створили ... прекрасні сховища даних, ідеально розклавши по поличках гори невживаної інформації, яка сама по собі не забезпечує ні швидкою, ні достатньо грамотної реакції на ринкові події".

- Вчений K. Parsaye вводить складений термін "OLAP Data Mining" (багатовимірний інтелектуальний аналіз) для позначення такого об'єднання інший науковець J. Han пропонує ще простішу назву - "OLAP Mining", і пропонує декілька варіантів інтеграції двох технологій.
- "Cubing then mining". Можливість виконання інтелектуального аналізу повинна забезпечуватися над будь-яким результатом запиту до багатовимірного концептуального уявлення, тобто над будь-яким фрагментом будь-якої проекції гіперкуба показників.
- "Mining then cubing". Подібно даним, витягнутим з сховища, результати інтелектуального аналізу повинні представлятися в гіперкубічній формі для подальшого багатовимірного аналізу.
- "Cubing while mining". Цей гнучкий спосіб інтеграції дозволяє автоматично активізувати однотипні механізми інтелектуальної обробки над результатом кожного кроку багатовимірного аналізу (переходу між рівнями узагальнення, витягання нового фрагмента гіперкуба і т. д.).

На жаль, дуже небагато виробників надають сьогодні достатньо могутні засоби інтелектуального аналізу багатовимірних даних в рамках систем OLAP. Проблема також полягає в тому, що деякі методи ІАД (байєсівські мережі, метод найближчого сусіда) непридатні для завдань багатовимірного інтелектуального аналізу, оскільки засновані на визначенні схожості деталізованих прикладів і не здатні працювати з агрегованими даними .


Інтерфейс
користувача
Інструмент багатовимірного інтелектуального аналізу
OLAP Сервер
ROLAP
SQL сервер
ROLAP
Агреговані дані
Деталізовані дані

Дуже часто виникає питання про різницю між засобами інтелектуального аналізу і OLAP-системами (On-Line Analytical Processing) - засобами оперативної аналітичної обробки.

OLAP - це частина технологій, скерованих на підтримку прийняття рішення. Звичайні засоби формування запитів і звітів описують саму базу даних. Технологія OLAP використовується для відповіді на задані питання. При цьому користувач сам формує гіпотезу про дані чи відношення між даними і після цього використовує серію запитів до бази даних для підтвердження чи відхилення цих гіпотез. Засоби Data Mining відрізняються від засобів OLAP тим, що замість перевірки передбачуваних взаємозалежностей, вони на основі наявних даних можуть будувати моделі, що дозволяють кількісно оцінити ступінь впливу досліджуваних факторів. Крім того, засоби інтелектуального аналізу дозволяють робити нові гіпотези про характер невідомих, але реально існуючих відношень у даних.

Сучасні технології інтелектуального аналізу опрацьовують інформацію з метою автоматичного пошуку шаблонів, характерних для яких-небудь фрагментів неоднорідних багатомірних даних. На відміну від оперативної аналітичної обробки даних у Data Mining тягар формулювання гіпотез і виявлення незвичайних шаблонів перекладено з людини на комп'ютер.

- **Приклади формулювань задач при використанні методів OLAP і Data Mining**

OLAPData Mining Які середні показники травматизму для людей, що палять і не палять?Які фактори найкраще передбачають нещасні випадки?Які середні розміри телефонних рахунків існуючих клієнтів у порівнянні з рахунками колишніх клієнтів (що відмовилися від послуг телефонної компанії)?Які характеристики відрізняють клієнтів, що, цілком ймовірно, збираються відмовитися від послуг телефонної компанії?Яка середня величина щоденної купівлі по вкраденій та не вкраденій кредитній картці?Які схеми купівлі характерні для шахрайства з кредитними картками?

Data Mining і сховища даних

Для успішного проведення всього процесу знаходження нових знань необхідною умовою є наявність сховища даних.

Отже, **сховище даних** - це предметно-орієнтований, інтегрований, прив'язаний до часу, незмінний збір даних для підтримки процесу прийняття управлінських рішень. Предметна орієнтація означає, що дані об'єднані в категорії і зберігаються відповідно до тих областей, що вони описують, а не до їх застосувань. Інтегрованість означає, що дані задовольняють вимогам усього підприємства (у його розвитку), а не єдиної функції бізнесу. Тим самим сховище даних гарантує, що однакові звіти, згенеровані для різних аналітиків, будуть містити однакові результати.

Прив'язка до часу означає, що сховище можна розглядати як сукупність "історичних" даних: можна відновити картину на будь-який момент часу. Атрибут часу завжди є явно присутнім у структурах сховища даних. Незмінність означає, що, потрапивши один раз у сховище, дані вже не змінюються на відміну від оперативних систем, де дані обов'язані бути присутніми тільки в останній версії, оскільки постійно змінюються. У сховище дані лише долучаються. Для рішення переліченого ряду задач, що неминуче виникають при організації й експлуатації інформаційного сховища, повинно існувати спеціалізоване програмне забезпечення. Сучасні засоби адміністрування сховища даних мають забезпечити ефективну взаємодію з інструментарієм знаходження нового знання.