обміну та управлінням з боку операційної системи. Ця функція є зовнішньою по відношенню до нерозпаралеленої програми, і час її реалізації – це чисті втрати продуктивності. Можлива компенсація цих втрат за рахунок більш раціонального кешування, не усуває природу самих втрат. Здійснення міжпроцесорного обміну за допомогою передачі повідомлень призводить до великих втрат продуктивності навіть в разі, коли всі елементи системи працюють на частоті процесорів.

1. *Lazou C.* Cray's Adaptive Supercomputing – A Paradigm Shift. March 24, 2006. http://www.hpcwire.com/features.
2. *Трахтенгерц Э.А.* Программное обеспечение параллельных процессов. М.: Наука, 1987 г. – 272 с.
3. *Booth, Nick.* NEC claims 10-Petaflop supercomputing breakthrough. March, 2008. http://www.theinquirer.net/gb/inquirer/news
4. *Сігарьов О.О., Душеба В.В.* Динаміка прискорення реалізації паралельного алгоритму в масштабованих системах з масовим паралелізмом // Моделювання та інформаційні технології. Зб. наук. пр. ІПМЕ ім. Г.Є. Пухова НАН України. – Вип. 79. – К.: 2017. – С.9-16.

http://doi.org/10.5281/zenodo.3859677

Поступила 3.10.2019р.

УДК 009.4

Б.М. Гавриш [1]
Б.В. Дурняк [1]
О.В. Тимченко [1, 2]

АНАЛІЗ СУЧАСНИХ ІНСТРУМЕНТІВ DATA MINING

Abstract. The article analyzes modern Data Mining tools. A detailed description of each Data Mining tool is given. The principles of operation of these tools are considered, the main criteria for comparison are presented. The pros and cons of each Data Mining tool are listed. Conclusions are made about the effectiveness of the DMST tool for analytical projects.

Keywords: DMST, Intelligent Processing, Data Mining, Mathematical Packages, Business Analytics, Data Analysis, Mat Package.

[1] Українська академія друкарства, Львів
[2] University of Warmia and Mazury Olsztyn, Poland

Вступ

Традиційна математична статистика, яка довгий час претендувала на роль основного інструменту аналізу даних, не може вирішити нові проблеми, які виникли. Головна причина – концепція усереднення за вибіркою, яка веде до операцій над фіктивними величинами. Методи математичної статистики виявилися корисними головним чином для перевірки заздалегідь сформульованих гіпотез (Verification-Driven Data Mining) та для «грубого» попереднього аналізу, що становить основу оперативного аналітичного опрацювання даних (OnLine Analytical Processing, OLAP).

В основу сучасної технології Data Mining (Discovery-Driven Data Mining) покладено концепцію шаблонів (патернів), які відображають фрагменти багатоаспектних взаємовідносин у даних. Ці шаблони – закономірності, властиві для підвибірок даних, які можуть бути компактно виражені у зрозумілій формі для людського сприйняття. Пошук шаблонів проводиться методами, не обмеженими рамками апріорних припущень про структуру, вибірку і види розподілу значень аналізованих показників.

Основна частина

Перспективи технології Data Mining

Потенціал Data Mining дає широкі можливості для розширення меж застосування інформаційних технологій. Щодо перспектив Data Mining можливі наступні напрямки розвитку:

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

Якщо розглядати майбутнє Data Mining в короткостроковій перспективі, то очевидно, що розвиток цієї технології здебільшого прямує 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 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 як нових

видів лікування різних захворювань, так і нового розуміння природи всесвіту.

Однак Data Mining приховує в собі і потенційну небезпеку – адже щораз більша кількість інформації стає доступною через всесвітню мережу, зокрема і відомості приватного характеру, і існує можливість щораз більше інформації видобути з неї.

Традиційні методи аналізу даних (статистичні методи) і OLAP (OnLine Analytical Processing) в основному орієнтовані на перевірку заздалегідь сформульованих гіпотез (verification-driven data mining) і на «грубий» розвідувальний (попередній) аналіз, що становить основу оперативного аналітичного опрацювання даних, в той час як одне з основних положень Data Mining – пошук неочевидних закономірностей. Інструменти Data Mining можуть знаходити такі закономірності самостійно а також самостійно будувати гіпотези про взаємозв'язки. Оскільки саме формулювання гіпотези щодо залежностей є найскладнішим завданням, перевага Data Mining в порівнянні з іншими методами аналізу є очевидною.

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

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

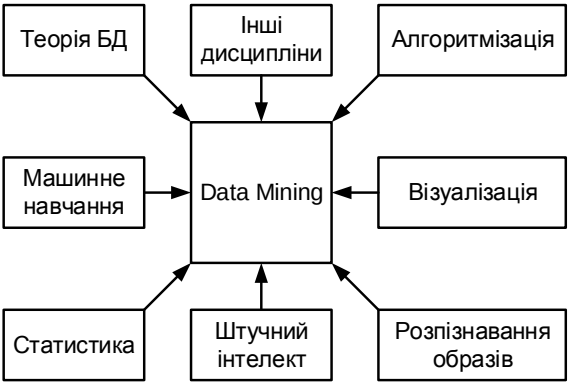

Рис.1. Базові дисципліни які використовує Data Mining

Сьогодні з'явилися нові наукові методи і спеціалізовані інструменти, для того щоб Data Mining для даних розвивалася ефективним способом (рис.1). Старі методи, що застосовувалися математиками і статистиками, забирали

багато часу, щоб в результаті отримати конструктивну і корисну інформацію.

Ринок інструментів Data Mining визначається широтою цієї технології і внаслідок цього – величезним різноманіттям програмного забезпечення.

Оскільки обсяг даних постійно зростає, то це призводить до того, що завдання аналізу стають щораз складнішими. Для вирішення завдань, які поділяються на загальні і приватні, необхідний постійний пошук нових, нестандартних і доступних знань для аналізу. Найважливішим інструментом пошуку таких знань є глибокий і всебічний аналіз даних, що описують процеси і явища, які відбуваються в аналітичних системах, з використанням сучасних інформаційних технологій. В даний момент Data Mining це найбільш багатообіцяючий напрямок інформаційних технологій. Існує наступне визначення Data Mining: набір різних методів і алгоритмів для виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних до інтерпретації знань, необхідних для прийняття рішень в різних сферах людської діяльності [1, 5].

Термін Data Mining отримав свою назву з двох понять: пошуку цінної інформації у великій базі даних (data) і видобутку гірської руди (mining). Обидва процеси вимагають або просіювання величезної кількості сирого матеріалу, або розумного дослідження і пошуку цінностей.

Для ефективної організації пошуку знань, необхідних для підтримки прийняття рішень в аналітичних системах, найбільш ефективним підходом є реалізація комплексних DM-проектів з глибокою інтеграцією аналітичних інструментів в робочі процеси.

Використовувані для цих цілей системи інтелектуального аналізу повинні відповідати таким вимогам:

1. Підтримка експорту/імпорту даних. Підтримка вивантаження/завантаження даних з різних вихідних областей. Об’єднання даних в сховищі даних.
2. Підтримка технології «клієнт-сервер» для опрацювання даних на віддалених серверах.
3. Підтримка/створення різних звітів.
4. Підтримка різних алгоритмів та методів інтелектуального опрацювання даних (методи математичної статистики, алгоритми бізнес-аналізу, машинне навчання тощо).
5. «Дружній», зручний графічний інтерфейс, який буде придатний для великого сектора користувачів.
6. Підтримка потужних засобів візуалізації даних. Система повинна містити широкий набір візуалізаторів вихідних даних, проміжних та кінцевих результатів, а також структуру побудованих моделей.

Даним вимогам відповідають 6 груп інструментів Data Mining:

1. Інструменти DM (DMFT – Data Mining Field Tools) – дані інструменти спрямовані на особливу прикладну область.

2. Інструменти для бізнес-аналітики (DMBT – Data Mining Business Tools) – не орієнтовані на роботу із завданнями Data Mining, але підтримують методи інтелектуального опрацювання даних (наприклад, алгоритми кластеризації, класифікації для бізнес аналізу).
3. Інструменти DM (RDMT – Research Data Mining Tools) – дані інструменти використовуються для розробки нових експериментальних алгоритмів і методів інтелектуальної розробки даних.
4. Математичні пакети (DMMP – Data Mining Mat Package) – дані пакети не були орієнтовані для Data Mining, але вони містять величезну кількість алгоритмів і методів, які дозволяють здійснювати функції інтелектуального аналізу даних [2].
5. Інструменти DM (SDMT – Specialties Data Mining Tools) – дані інструменти використовуються для певних видів або методів інтелектуального опрацювання даних.
6. Інтеграційні пакети (IDMT–Integration Data Mining Tool) – набори алгоритмів, які утворюють або окремі програмні засоби, або пакети розширення.

Таблиця 1

Порівняльна характеристика інструментів DM

Інструменти DM	Експорт/ Імпорт	Підтримка клієнт-сервер	Наявність звітів	Підтримка різноманітних алгоритмів	Візуалізація
DMFT	так	ні	ні	ні	так
DMBT	так	так	так	ні	ні
RDMT	ні	ні	ні	так	ні
DMMP	так	ні	ні	так	ні
SDMT	так	ні	ні	ні	так
IDMP	ні	ні	ні	так	ні
DMST	так	так	так	так	так

7. «Набори» інтелектуального опрацювання даних (DMST – Data Mining Suite Tools) – підтримують цілий спектр алгоритмів та методів інтелектуального опрацювання даних. Орієнтовані на роботу з різними даними (багатовимірні дані, структуровані і неструктуровані дані).

Для більш доступного сприйняття даної інформації зроблена таблиця 1, в якій проведений порівняльний аналіз інструментів інтелектуального опрацювання даних для реалізації аналітичних DM-проектів.

На основі даних, зазначених в таблиці, можна зробити висновок, що сформульовані раніше вимоги відповідають «наборам» інструментів DMST.

Висновки

Отже, можна зробити висновок, що для комплексної реалізації аналітичних проектів потрібно використовувати інструменти Data Mining Suit Tools, оскільки DMST дають можливість використовувати повний набір засобів інтелектуального аналізу даних. Вони організовують ефективний пошук знань в базах даних, підтримують технологію «клієнт-сервер», яка дозволить проводити ефективний пошук знань в базах даних (в локальних, віддалених). До мінусів можна віднести високу вартість даних інструментів.

1 Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? – Tandem Computers Inc., 1996.

2 Data Mining and Image Processing Toolkits – Режим доступу: http:// datamining.itsc.uah.edu/adam/)

3 Data Mining, Web Mining, Text Mining, and Knowledge Discovery – Режим доступу: http://://www.kdnuggets.com)

4 *Майер-Шенбергер В.* Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Виктор Майер-Шенбергер, Кеннет Кукьер ; пер. с англ. Инны Гайдюк. – М. : Манн, Иванов и Фербер, 2014. – 240 с

5 *Коэн Дж.* Могучие способности: новые приемы анализа больших данных [Електронний ресурс] / Джеффри Коэн, Брайен Долэн, Марк Данлэп, Джозеф Хеллерстейн, Кейлэб Велтон; пер. с англ. Сергей Кузнецов. – Режим доступу: http://citforum.ru/database/articles/mad_skills/.

6 History and evolution of big data analytics [Електронний ресурс]. – Режим доступу: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html

7 *Ronen Sh.* Links that speak: The global language network and its association with global fame [Електронний ресурс] / Shahar Ronen, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker, César A. Hidalgo // PNAS, Vol. 111, No. 52, 2014. – Режим доступу: http://stevenpinker.com/files/pinker/files /pnas_hildago_et_al_global_language_ network_2014.pdf.

http://doi.org/10.5281/zenodo.3859679

Поступила 23.09.2019р.