Grid і інтелектуальна обробка даних Data Mining

А.І.Петренко. проф.., доктор технічних наук

Анотація: Обговорюються відмінності Data Mining від класичних статистичних методів аналізу і OLAP-систем, розглядаються типи закономірностей, що виявляється Data Mining (асоціація, класифікація, послідовність, кластеризація, прогнозування). Описується сфера застосування Data Mining і приклад системи ADaM, працюючій в середовищі Grid.

I.Вступ: перспективи технології Data Mining

. Що недавно в Україні почали функціонувати філія Світового Центру Даних і національна Grid інфраструктура (академічний і освітянський сегменти), так що вітчизняні вчені і фахівці можуть розраховувати зараз на підвищені обсяги даних з різних галузей, що обробляються в об’єднаній мережі кластерів країни. Розвиток методів запису і зберігання даних привів до бурхливого зростання об'ємів збираної і аналізованої інформації. Об'єми даних настільки значні, що людині просто не під силу проаналізувати їх самостійно, хоча необхідність проведення такого аналізу цілком очевидна, адже в цих "сирих даних" укладені знання, які можуть бути використаний при ухваленні рішень

Для того, щоб провести автоматичний аналіз даних, використовується **Data Mining** (здобич (витягання) знань). Це нова технологія інтелектуального аналізу даних з метою виявлення прихованих закономірностей у вигляді значущих особливостей, кореляцій, тенденцій і шаблонів. Сучасні системи здобичі даних використовують засновані на методах штучного інтелекту засоби уявлення і інтерпретації, що і дозволяє знаходити розчинену в терабайтних сховищах не очевидну, але вельми цінну інформацію. Фактично, ми говоримо про те, що в процесі Data mining система не відштовхується від наперед висунутих гіпотез, а пропонує їх сама на основі аналізу.

Існує безліч визначень Data Mining, але в цілому вони співпадають у виділенні чотирьох основних ознак. Згідно визначенню, Г. Пиатецкого-Шаниро (G. Pia-tetsky Shapiro,GTE Labs), одного з ведучих світових експертів в даній області, Data Mining — дослідження і виявлення алгоритмами, засобами штучного інтелекту в "сирих даних" прихованих структур, шаблонів або залежності, яка:

- раніше не були відомі;
- нетривіальні;
- практично корисні;
- доступні для інтерпретації людиною і необхідні для ухвалення рішень в різних сферах діяльності.

Специфіка сучасних вимог до продуктивної переробки інформації наступна:

- дані мають необмежений обсяг;
- дані є різнорідними (кількісними, якісними, текстовими);
- результати повинні бути конкретний і зрозумілий;
- інструменти для обробки "сирих даних" повинні бути прості у використовуванні.

Традиційна математична статистика, що довгий час претендувала на роль основного інструменту аналізу даних, не відповідала виниклим проблемам. Головна причина — концепція усереднювання по вибірці, що приводить до операцій над фіктивними величинами. Методи математичної статистики виявилися корисними головним чином для перевірки наперед сформульованих гіпотез і для "грубого розвідувального аналізу", що становить основу оперативної аналітичної обробки даних OLAP.

В основу сучасної технології Data Mining встановлена концепція шаблонів (pattern), що відображають фрагменти багатоаспектних взаємостосунків в даних. Цими шаблонами є закономірності, властиві підвибіркам даних, які можуть бути компактно виражені у формі, зрозумілій людині. Пошук шаблонів проводиться методами, не обмеженими рамками

апріорних припущень про структуру вибірки і вид розподілів значень аналізованих показників. Причини популярності Data Mining:

- стрімке накопичення даних (рахунок йде вже на экзабайти);
- загальна комп'ютеризація бізнес-процесів;
- проникнення Інтернет у всі сфери діяльності;
- прогрес в області інформаційних технологій: вдосконалення СУБД і сховищ даних;

прогрес в області виробничих технологій: стрімке зростання продуктивності комп'ютерів, об'ємів накопичувачів, впровадження Grid систем.

Алгоритми, що використовуються в Data Mining, вимагають великої кількості обчислень. Раніше це було стримуючим чинником широкого практичного застосування Data Mining, проте сьогоднішнє зростання продуктивності сучасних процесорів зняло гостроту цієї проблеми. Тепер за прийнятний час можна провести якісний аналіз сотень тисяч і мільйонів записів. Data Mining – ***міждисциплінарна галузь*** , що виникла і розвивалася на базі таких наук як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних і ін., див. рис. 1[1]:

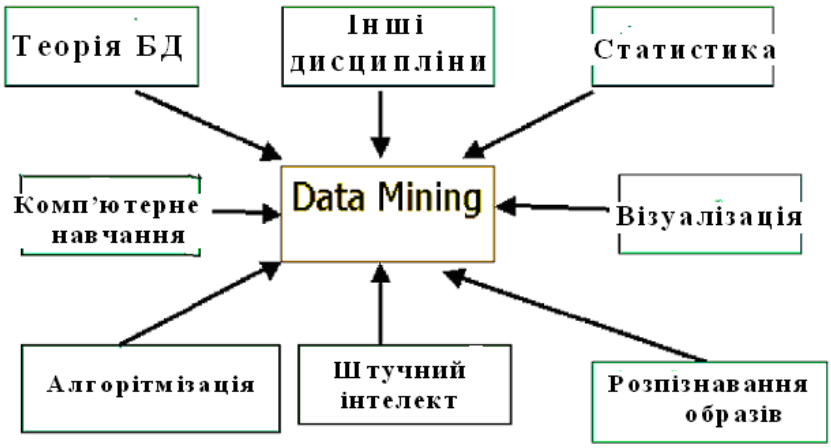

Рис. 1. Data Mining як міждисциплінарна галузь

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

- виділення типів предметних галузей з відповідними їм евристиками, формалізація яких полегшить рішення відповідних задач Data Mining, що відносяться до цих галузей;
- створення формальних мов і логічних засобів, за допомогою яких будуть формалізовані міркування і автоматизація яких стане інструментом рішення задач Data Mining в конкретних предметних галузях;
- створення методів Data Mining, здатних не тільки витягувати з даних закономірності, але і формувати деякі теорії, що 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 направлений на галузі, пов'язані з Grid системами для e-Science. Можливості e-Science характеризують обчислювальну інфраструктуру, яка складається із трьох концептуальних рівнів (рис.2) :

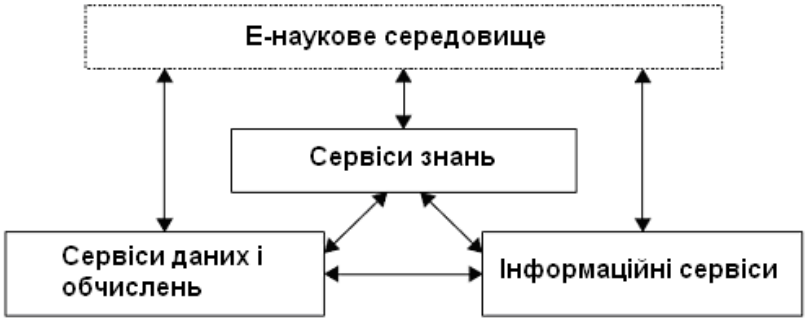

Рис..2. Трирівнева архітектура Grid-сервісів

- ***Сервіси даних/обчислень***. Цей рівень містить інформацію, яким чином розташовані обчислювальні ресурси, коли заплановано виконувати обчислення та засоби передавання даних між різними обчислювальними ресурсами. Рівень може опрацьовувати з великі обсяги даних, забезпечуючи швидкі мережі й подають різноманітні ресурси як єдиний метакомп'ютер.
- ***Інформаційні сервіси***. Цей рівень вказує, яким чином інформація продається, зберігається, хто і яким чином має до неї доступ. Тут інформація зрозуміла як дані зі значенням. Наприклад, виявлення цілого числа як подання температури процесу реакції, розпізнавання, що рядок − ім'я людини.
- ***Сервіси Знань***. Цей рівень надає спосіб, яким знання придбане, використовується, знайдено, опубліковано, щоб допомогти користувачам досягати своїх специфічних цілей. Тут знання подаються як інформація, застосована для досягнення мети, вирішення проблеми або прийняття рішення. Прикладом може бути процедура розпізнавання оператором підприємства моменту часу, коли температура реакції вимагає завершення виконання процесу.

Розглянуті поняття є складовою частиною так званої інформаційної піраміди, в підставі якій знаходяться дані, наступний рівень - це інформація, потім йде рішення, завершує піраміду рівень знання. По міру просування вгору по інформаційній піраміді об'єми даних переходять в цінність рішень, тобто цінність для знань . Як видно з рис.2, даний процес є циклічним. Ухвалення рішень вимагає інформації, яка заснована на даних. Дані забезпечують інформацію, яка підтримує рішення, і т.д.

Усі Grid-системи, які уже побудовані або будуть побудовані, містять деякі елементи всіх трьох рівнів. Ступінь важливості використання цих рівнів буде вирішуватися користувачем. Таким чином, у деяких випадках оброблення величезних обсягів даних буде домінуючим завданням, у той час як в інших випадках обслуговування знання буде основною проблемою. Дотепер більшість науково-дослідних робіт в галузі Grid концентрувалося на рівні даних/обчислень й на інформаційному рівні. У той же час все ще багато невирішених проблем, що стосуються керування широкомасштабними розподіленими обчисленнями та ефективного доступу і розповсюдження інформації з гетерогенних джерел. Вважається, що повний потенціал Grid обчислень може бути досягнутий тільки завдяки повній експлуатації функціональних можливостей та можливостей, що надаються рівнем знання, тому цей рівень необхідний для автоматизованого прямого простого доступу до операцій і взаємодій.

II. Методи і задачі Data Mining

Основна особливість Data Mining - це поєднання широкого математичного інструментарію (від класичного статистичного аналізу до нових кібернетичних методів) і

останніх досягнень у сфері інформаційних технологій. В технології Data Mining гармонійно об'єдналися строго формалізовані методи і методи неформального аналізу, тобто кількісний і якісний аналіз даних.

До методів і алгоритмів Data Mining відносяться наступні: штучні нейроні мережі, дерева рішень, символьні правила, методи найближчого сусіда і к- найближчого сусіда, метод опорних векторів, байесові мережі, лінійна регресія, кореляційно-регресійний аналіз; ієрархічні методи кластерного аналізу, неієрархічні методи кластерного аналізу, у тому числі алгоритми к-середніх і к-медіани; методи пошуку асоціативних правил, у тому числі алгоритм Apriori; метод обмеженого перебору, еволюційне програмування і генетичні алгоритми, різноманітні методи візуалізації даних і безліч інших методів.

Більшість аналітичних методів, що використовуються в технології Data Mining - це відомі математичні алгоритми і методи. Новою в їх застосуванні є можливість їх використовування при рішенні тих або інших конкретних проблем, обумовлена новими можливостями технічних і програмних засобів, що з'явилися. Слід зазначити, що більшість методів Data Mining була розроблена в рамках теорії штучного інтелекту. Єдиної думки щодо того, які задачі слід відносити до Data Mining, немає. Більшість авторитетних джерел перераховує наступні: *класифікація, кластеризація, прогнозування, асоціація, візуалізація, аналіз і виявлення відхилень, оцінювання, аналіз зв'язків, підведення підсумків.* Розглянемо деякі з них.

Класифікація (Classification). Це найпростіша і поширена задача Data Mining. В результаті рішення задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних - класи; по цих ознаках новий об'єкт можна віднести до того або іншого класу. Для вирішення задачі класифікації можуть використовуватися методи: найближчого сусіда (Nearest Neighbor); к-ближайшего сусіда (к-Nearest Neighbor); байесові мережі (Bayesian Networks); індукція дерев рішень; нейроні мережі (neural networks).

Кластеризація (Clustering) Кластеризація є логічним продовженням ідеї класифікації. Це задача складніша, особливість кластеризації полягає в тому, що класи об'єктів спочатку не визначені. Результатом кластеризації є розбиття об'єктів на групи. Прикладом методу задачі кластеризації є особливий вид нейроних мереж (карти Кохонена), що само організовуються без вчителя..

Асоціація (Associations). В ході рішення задачі пошуку асоціативних правил відшукуються закономірності між зв'язаними подіями в наборі даних. Відмінність *асоціації* від двох попередніх задач Data Mining: пошук закономірностей здійснюється не на основі властивостей аналізує мого об'єкту, а між декількома подіями, які відбуваються одночасно.Самий відомий алгоритм рішення задачі пошуку асоціативних правил - алгоритм Apriori.

Послідовність (Sequence), або послідовна асоціація *(sequential association)* Послідовність дозволяє знайти тимчасові закономірності між транзакціями. Задача послідовності подібна асоціації*, але* її метою є встановлення закономірностей не між одночасно наступаючими подіями, а між подіями, зв'язаними в часі (тобто що відбуваються з деяким певним інтервалом в часі. Цю задачу Data Mining також називають задачею знаходження послідовних шаблонів (sequential *pattern).*Правило послідовності*:* після події X через певний час відбудеться подія У.

Прогнозування (Forecasting). В результаті рішення задачі прогнозування на основі особливостей існуючих даних оцінюються пропущені або ж майбутні значення цільових чисельних показників. Для вирішення таких задач широко застосовуються методи математичної статистики, нейроні мережі і ін.

Візуалізація (Visualization, Graph Mining) В результаті візуалізації створюється графічний образ аналізованих даних. Для вирішення задачі візуалізації використовуються графічні

методи, що показують наявність закономірностей в даних. Приклад методів візуалізації - представлення даних в 2-D і 3-D вимірюваннях.
Підведення підсумків (Summarization) - задача, мета якої - опис конкретних груп об'єктів з аналізованого набору даних та інші.
Задачі Data Mining, залежно від моделей, що використовуються, можуть бути ***дескриптивними*** і ***прогнозуючими***. В результаті рішення описових (descriptive) задач аналітик одержує шаблони, що описують дані, які піддаються інтерпретації. Ці задачі описують загальну концепцію аналізованих даних, визначають інформативні, підсумкові, відмітні особливості даних.
Прогнозуючі (predictive) задачі грунтуються на аналізі даних, створенні моделі, прогнозі тенденцій або властивостей нових або невідомих даних.

III. Класифікація стадій Data Mining

Data Mining може складатися з двох або трьох стадій :
Стадія 1. Виявлення закономірностей (вільний пошук**).**
Стадія 2. Використовування виявлених закономірностей для прогнозу невідомих значень (прогностичне моделювання).
На додаток до цих стадій іноді вводять стадію оцінювання (валідації) , наступну за стадією вільного пошуку. Мета валідації - перевірка достовірності знайдених закономірностей. Проте, ми вважатимемо валідацію частиною першою стадії, оскільки в реалізації багатьох методів, зокрема, нейроних мереж і дерев рішень, передбачений розподіл загальної множини даних на навчальні і перевірочні, і останні дозволяють перевіряти достовірність отриманих результатів.
Стадія 3. Аналіз виключень **-** стадія призначена для виявлення і пояснення аномалій, знайдених в закономірностях.

Вільний пошук (Discovery). На стадії вільного пошуку здійснюється дослідження набору даних з метою пошуку прихованих закономірностей. Попередні гіпотези щодо виду закономірностей тут не визначаються. **Закономірність** (law) - істотний і постійно повторюється взаємозв'язок, що визначає етапи і форми процесу становлення, розвитку різних явищ або процесів.
Система Data Mining на цій стадії визначає шаблони, для отримання яких в системах OLAP, наприклад, аналітику необхідно обдумувати і створювати множину запитів. Тут же аналітик звільняється від такої роботи - шаблони шукає за нього система. Особливо корисно застосування даного підходу в надвеликих базах даних, де уловити закономірність шляхом створення запитів достатньо складно, для цього вимагається перепробувати безліч різноманітних варіантів. Вільний пошук представлений такими діями:

- виявлення закономірностей умовної логіки (conditional logic);
- виявлення закономірностей асоціативної логіки (associations and affinities);
- виявлення трендів і коливань (trends and variations).

Описані дії в рамках стадії вільного пошуку виконуються при допомозі :

- індукції правил умовної логіки (задачі класифікації і кластеризації, опис в компактній формі близьких або схожих груп об'єктів);
- індукції правил асоціативної логіки (задачі асоціації і послідовності і витягування при їх допомозі інформація);
- визначення трендів і коливань (початковий етап задачі прогнозування).

На стадії вільного пошуку також повинна здійснюватися валідація закономірностей, тобто перевірка їх достовірності на частини даних, які не брали участь у формуванні закономірностей.

Прогностичне моделювання (Predictive Modeling) .Друга стадія Data Mining - прогностичне моделювання - використовує результати роботи першої стадії. Тут знайдені закономірності використовуються безпосередньо для прогнозування. Прогностичне моделювання включає такі дії:

- прогноз невідомих значень (outcome prediction);
- прогнозування розвитку процесів (forecasting).

В процесі прогностичного моделювання розв'язуються задачі класифікації і прогнозування. При рішенні задачі класифікації результати роботи першої стадії (індукції правил) використовуються для віднесення нового об'єкту з певною упевненістю до одного з відомих, наперед визначених класів на підставі відомих значень. При рішенні задачі прогнозування результати першої стадії (визначення тренда або коливань) використовуються для прогнозу невідомих (пропущених або ж майбутніх) значень цільової змінної (змінних).

Порівняємо вільного пошуку і прогностичного моделювання з погляду логіки Вільний пошук розкриває загальні закономірності. Він по своїй природі *індуктивний*. Закономірності, отримані на цій стадії, формуються від часткового до загального. В результаті ми одержуємо деяке загальне знання про деякий клас об'єктів на підставі дослідження окремих представників цього класу.
Прогностичне моделювання, навпаки, *дедуктивне*. Закономірності, отримані на цій стадії, формуються від загального до часткового. Тут ми одержуємо нове знання про деякий об'єкт або ж групі об'єктів на підставі:

- знання класу, до якого належать досліджувані об'єкти;
- знання загального правила, діючого в межах даного класу об'єктів.

Аналіз виключень (forensic analysis). На третій стадії Data Mining аналізуються виключення або аномалії, виявлені в знайдених закономірностях. Дія, виконувана на цій стадії, - виявлення відхилень (deviation detection). Для виявлення відхилень необхідно визначити норму, яка розраховується на стадії вільного пошуку.Стадія аналізу виключень може бути використана як очищення даних .

IУ. Практичні реалізації Data Mining

Сьогодні у світі існують декілька комерційних і фірмових систем (PolyAnalyst,Weka, Orange Canvas, SVM^{light6}, Cognos і інші)[4,8]. В даний час вартість масових систем від $1000 до $10000. Кількість інсталяцій масових продуктів, судячи за наявними відомостями, сьогодні досягає десятків тисяч.
Особливості Data Mining систем розглянемо на прикладі системи ADaM (Algorithm Development and Mining System), розробленої Центром Інформаційних Технологій і Систем (ITSC) в Університеті Алабами , яка використовується для дистанційної обробки наукових даних технологіями Data Mining [6]. Створені засоби Data Mining складаються з взаємодіючих компонентів, які можна для різних прикладних задач включати у спеціалізовані додатки. ADaM містить понад 100 компонентів, які можуть бути конфігуровані так, щоб за замовленням користувача створювати необхідні процеси інтелектуального аналізу даних. Нові компоненти можуть бути легко додані, щоб пристосувати систему до інших проблем науки.
Кожний компонент ADaM підтримується C, C++, або іншим програмним інтерфейсом додатку (API), загальними інструментальними засобами опису (Perl, Python, сценарії оболонок) і кінець кінцем інтерфейсом WEB-сервісів, що забезпечує використання Web і Grid додатків. Компоненти ADaM – універсальні модулі інтелектуального аналізу даних (mining) і обробки зображень, які можуть бути легко пристосовані для численних рішень і задач. Приклади компонентів ADaM наведені в табл.1.

Таблиця 1. Компоненти ADaM

Методи класифікації • Bayes Classifier • Naïve Bayes Classifier • Bayes Network Classifier • CBEA Classifier • Decision Tree Classifier • SEA classifier • Very Fast Decision Tree Classifier • Back Propagation Neural Network • k-Nearest Neighbor Classifier • Multiple Prototype Minimum Distance Classifier Recursively Splitting Neural Network ***Методи кластеризації*** • DBSCAN • Hierarchical Clustering • Isodata • k-Means • k-Mediods • Maximin	***Методи оцінки властивостей*** • Backward Elimination • Forward Selection • Principal Components • RELIEF (filter-based feature selection) • Removing Attributes • Checking Range ***Методи розпізнавання образів*** • Accuracy Measures • Data Cleaning • k-Fold Cross Validation • Vector Magnitude • Merging Patterns • Normalization • Sampling • Subsetting • Statistics • Cleaning Outliers • Comparing Image File • Comparing ASCII files • Discretization	• Magnitude Computation ***Методи асоціації*** • Apriori ***Методи оптимізації*** • Genetic Algorithm • Hill Climbing • Simulated Annealing ***Базові перетворення образжень*** • Arithmetic Operations(+-*/) • Collaging • Cropping • Image Difference • Image Normalization • Image Moments • Equalization • Inverse • Quantization • Relative Level Quantization • Resampling • Rotation • Scaling • Statistics • Thresholding • Vector Plot	***Визначення форм, сегментів, границь*** • Boundary Detection • Polygon Circumscription • Making Region • Marking Region ***Методи фільтрації*** • Dilation • Energy Erosion • Fast Fourier Transfer • Median and Mode Filters • Pulse Coupled Neural Network • Spatial Filter ***Визначення елементів текстур*** • Association Rules • Fractal Dimension • Gabor Filter • GLCM (Gray Level Concurrence Matrix) • GLRL (Gray Level Run Length) • Markov Random Field Computing

ITSC є партнером NSF (National Science Foundation) дослідницького проекту у сфері IT під назвою LEAD (Linked Environments for Atmospheric Discovery – Зв'язані оточення для дослідження атмосфери). Формування користувачем з окремих компонентів ADaM завдання на інтелектуальну обробку показано на рис.3, а візуалізацію змодельованого тернадо –на рис.4.

Онтологія − це засіб опису семантики проблемної області за допомогою словника і підібраної специфікації існуючих в ній відношень та обмежень, що забезпечують інтеграцію словника. Інформаційні онтології створюються завжди з конкретною метою − рішення конструкторських задач; вони оцінюються більше щодо використання, ніж повноти. Онтології – це фундаментальні блоки для будівництва семантичної Grid. Їх визначають як: «розширення існуючої Grid, де інформації та сервісам надаються конкретні значення, покращені можливості для об'єднаної роботи людей та комп'ютерів». Для проекту LEAD створена онтологія, що забезпечує семантичні метадані для наборів даних і яка служить як освітній сервіс, ресурс знань і список посилань для громадськості. ITSC проводить дослідження по створенню національної кібернетичної інфраструктури для виконання широкомасштабних наукових досліджень і проектування.

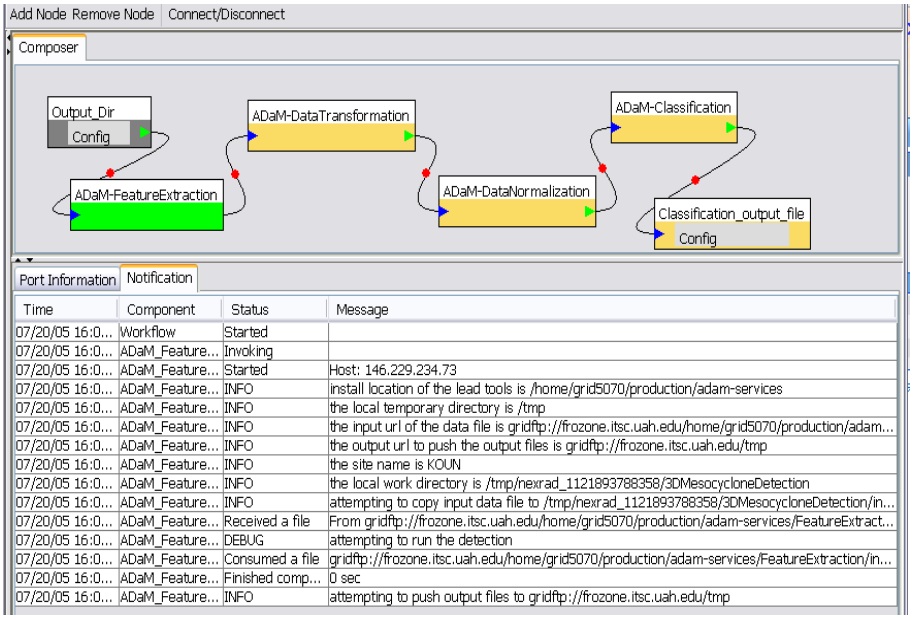

Рис.3. приклад формування завдання для Data Mining

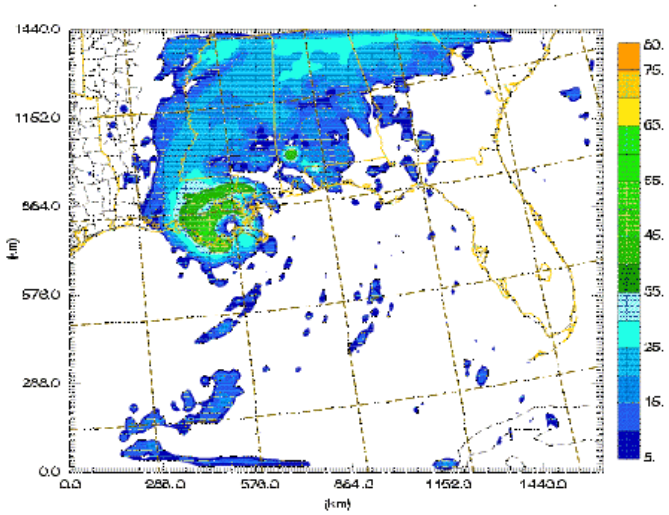

Рис.4. Вихідна інформація Data Mining

Спільно з академічними установами, урядом і промисловістю, ITSC встановлює регіональну оптичну мережу із зв'язністю з національними і міжнародними дослідницькими і освітніми мережами. Дослідження ITSC в обчислювальних мережах високої продуктивності включають розробку паралельних алгоритмів і оцінку продуктивності і регулювання обчислювальних кластерів і паралельних файлових систем. ITSC розробляє алгоритми реального часу для об'єднання даних і трасування для дуже великих сенсорних мереж. Мережі, що налічують більше мільйона різнорідних датчиків, використовуються для відстежування сотень цільових об'єктів на моделювання театрі військових дій.

У. Сфери застосування Data Mining

Слід відразу визначити , що область використовування Data Mining нічим не обмежена - вона скрізь, де є які-небудь дані. Можна виділити два напрями застосування систем Data Mining: як масового продукту і як інструменту для проведення унікальних досліджень.

Зараз технологія Data Mining використовується практично у всіх сферах діяльності людини, де накопичені ретроспективні дані. Розглянемо чотири основні сфери застосування технології Data Mining більш детально: наука, бізнес, роздрібна торгівля і Web-напрям [1,5,7].

5.1. Data Mining для наукових досліджень і промисловості.

Одна з наукових областей застосування технології Data Mining - ***біоінформатик***а, напрям, метою якого є розробка алгоритмів для аналізу і систематизації генетичної інформації. Отримані алгоритми використовуються для визначення структур макромолекул, а також їх функцій, з метою пояснення різних біологічних явищ.
Не дивлячись на консервативність ***медицини*** в багатьох її аспектах, технологія Data Mining останніми роками активно застосовується для різних досліджень і в цій сфері людської діяльності. Традиційно для постановки медичних діагнозів використовуються експертні системи, які побудовані на основі символьних правил, що поєднують, наприклад, симптоми пацієнта і його захворювання. З використанням Data Mining за допомогою шаблонів можна розробити базу знань для експертної системи.

В області ***фармацевтики*** методи Data Mining також мають достатньо широке застосування. Це задачі дослідження ефективності клінічного застосування певних препаратів, визначення груп препаратів, які будуть ефективні для конкретних груп пацієнтів. Актуальними тут також є задачі просування лікарських препаратів на ринок.
Молекулярна генетика і генна інженерія
В ***молекулярній генетиці і генній інженерії*** виділяють окремий напрям Data Mining, який має назву аналіз даних в микромассивах (Microarray Data Analysis*, MDA).* Деякі застосування цього напряму:

- нова молекулярна мета для терапії;
- рання і більш точна діагностика;
- поліпшені і індивідуально підібрані види лікування;
- фундаментальні біологічні відкриття.

Приклади використовування Data Mining - молекулярний діагноз деяких найсерйозніших захворювань; відкриття того, що генетичний код дійсно може передбачати вірогідність захворювання; відкриття деяких нових ліків і препаратів.
Основні поняття, якими оперує Data Mining в областях "Молекулярна генетика і генна інженерія" - маркери, тобто генетичні коди, які контролюють різні ознаки живого організму. На фінансування проектів з використанням Data Mining в даних сферах виділяють значні фінансові кошти.

Технологія Data Mining активно використовується в дослідженнях ***органічної і неорганічної хімії***. Одне з можливих застосувань Data Mining в цій сфері - виявлення яких-небудь специфічних особливостей будови з'єднань, які можуть включати тисячі елементів.
Основні задачі Data Mining в ***промисловому виробництві*** :

- комплексний системний аналіз виробничих ситуацій;
- короткостроковий і довгостроковий прогноз розвитку виробничих ситуацій;
- вироблення варіантів оптимізаційних рішень;
- прогнозування якості виробу залежно від деяких параметрів технологічного процесу;
- виявлення прихованих тенденцій і закономірностей розвитку виробничих процесів;
- прогнозування закономірностей розвитку виробничих процесів;
- виявлення прихованих чинників впливу;

- виявлення і ідентифікація раніше невідомих взаємозв'язків між виробничими параметрами і чинниками впливу;
- аналіз середовища взаємодії виробничих процесів і прогнозування зміни її характеристик;
- вироблення оптимізаційних рекомендацій по управлінню виробничими процесами;
- візуалізацію результатів аналізу, підготовку попередніх звітів і проектів допустимих рішень з оцінками достовірності і ефективності можливих реалізацій.
- Наприклад, при збірці автомобілів виробники повинні враховувати вимоги кожного окремого клієнта, тому їм потрібна можливість прогнозування популярності певних характеристик і знання того, які характеристики звичайно замовляються разом; виробникам потрібно також передбачати число клієнтів, які подадуть гарантійні заявки, і середню вартість заявок. Авіакомпанії можуть знайти групу клієнтів, яких даними заохочувальними заходами можна спонукати літати більше. Наприклад, одна авіакомпанія знайшла категорію клієнтів, які скоювали багато польотів на короткі відстані, не накопичуючи достатньо миль для вступу до їх клубів, тому вона таким чином змінила правила прийому в клуб, щоб заохочувати число польотів так же, як і милі.

5.2. ***Data Mining для вирішення бізнес-задач.***

Досягнення технології Data Mining використовуються в банківській справі для вирішення наступних поширених задач:

- *виявлення шахрайства з кредитними картками*. Шляхом аналізу минулих транзакцій, які згодом виявилися шахрайськими, банк виявляє деякі стереотипи такого шахрайства;
- *сегментація клієнтів*. Розбиваючи клієнтів на різні категорії, банки роблять свою маркетингову політику більш цілеспрямованої і результативної, пропонуючи різні види послуг різним групам клієнтів;
- *прогнозування змін клієнтури*. Data Mining допомагає банкам будувати прогнозні моделі цінності своїх клієнтів, і відповідним чином обслуговувати кожну категорію.

У сфері ***електронної комерції*** Data Mining застосовується для формування рекомендаційних систем і рішення задач класифікації відвідувачів Web-сайтів. Така класифікація дозволяє компаніям виявляти певні групи клієнтів і проводити маркетингову політику відповідно до знайдених інтересів і потреб клієнтів. Технологія *Data Mining* для електронної комерції тісно пов'язана з технологією Web Mining .

У сфері ***маркетингу*** Data Mining знаходить дуже широке застосування для відповідей на основні питання маркетингу "Що продається?", "Як продається?", "Хто є споживачем?" Інший поширений набір методів для вирішення задач маркетингу - методи і алгоритми пошуку асоціативних правил. Також успішно тут використовується пошук тимчасових закономірностей.

5.3. У сфері ***роздрібної торгівлі*** сьогодні збирається докладна інформація про кожну окрему купівлю, використовуючи кредитні картки з маркою магазина і комп'ютеризовані системи контролю. Ось типові задачі, які можна вирішувати за допомогою Data Mining у сфері роздрібної торгівлі:

- аналіз середовища взаємодії виробничих процесів і прогнозування зміни її характеристик; *аналіз купівельної корзини* (аналіз схожості), призначений для виявлення товарів, які покупці прагнуть придбавати разом. Знання купівельної корзини необхідне для поліпшення реклами, вироблення стратегії створення запасів товарів і способів їх розкладки в торгових залах;
- *дослідження тимчасових шаблонів* допомагає торговим підприємствам ухвалювати рішення про створення товарних запасів. Воно дає відповіді на питання типу

"Якщо сьогодні покупець придбав відеокамеру, то через який час він найімовірніше купить нові батареї і плівку?"

- *створення прогнозуючих моделей* дає можливість торговим підприємствам взнавати характер потреб різних категорій клієнтів з певною поведінкою, наприклад, купуючих товари відомих дизайнерів або відвідуючих розпродажі. Ці знання потрібні для розробки точно направлених, економічних заходів щодо просування товарів.

5.4. Web Mining

Web Mining можна перевести як "здобич даних в Web". Web здатний визначати інтереси і переваги кожного відвідувача сайтів , спостерігаючи за його поведінкою, що є серйозною і критичною перевагою конкурентної боротьби на ринку електронної комерції. Системи Web Mining можуть відповісти на багато питань, наприклад, хто з відвідувачів є потенційним клієнтом Web-магазина, яка група клієнтів Web-магазина приносить найбільший дохід, які інтереси певного відвідувача або групи відвідувачів .
Технологія Web Mining охоплює методи, які здатні на основі даних сайту знайти нові, раніше невідомі знання і які надалі можна буде використовувати на практиці. Іншими словами, технологія Web Mining застосовує технологію Data Mining для аналізу неструктурованої, неоднорідної, розподіленої і значної за об'ємом інформації, що міститься на Web-вузлах. При використовуванні Web Mining перед розробниками виникає два типи задач. Перша торкається збору даних, друга - використовування методів персоніфікації. В результаті збору деякого об'єму персоніфікованих ретроспективних даних про конкретного клієнта, система накопичує певні знання про нього і може рекомендувати йому, наприклад, певні набори товарів або послуг. На основі інформації про всіх відвідувачів сайту Web-система може виявити певні групи відвідувачів і також рекомендувати їм товари або ж пропонувати товари в розсилках.
В останні роки з’явилися Web-додатки типу ***Машап*** (від англ. mash-up — «змішувати»), які поєднують дані більш ніж з одного джерела і будується комбінуванням функціональності різних програмних інтерфейсів і джерел даних. Машапи вже застосовуються як:
- сервіси агрегування: збирають інформацію з різних джерел та розміщують їх в одному місці;
− збирачі даних: збирають дані з різних джерел, щоб створити новий сервіс (тобто агрегування);
− контролювачі змісту: відслідковують, фільтрують, аналізують та дозволяють пошук сервісів;
− сервісні збирачі.

5.5. Text Mining ("здобич" аналіз текстів)

Text Mining охоплює нові методи для виконання семантичного аналізу текстів, інформаційного пошуку і управління. На відміну від технології Data Mining, яка передбачає аналіз впорядкованої в якусь структуру інформації, технологія Text Mining аналізує великі і надвеликі масиви неструктурованої інформації. Програми, що реалізовують цю задачу, повинні деяким чином оперувати природною людською мовою і при цьому розуміти семантику аналізованого тексту. .

5.6. Call Mining ("здобич" і аналіз дзвінків)

Технологія Call Mining об'єднує в себе розпізнавання мови, її аналіз і Data Mining. Її мета - спрощення пошуку даних в аудіо архівах, що містять записи переговорів між операторами і клієнтами. За допомогою цієї технології оператори можуть знаходити недоліки в системі обслуговування клієнтів, знаходити можливості збільшення продажів, а також виявляти тенденції в обігу клієнтів. Аналітики відзначають, що за останні роки

інтерес до систем на основі Call Mining значно зріс. Це пояснюється тим фактом, що менеджери вищої ланки компаній, що працюють в різних сферах, у тому числі в області фінансів, мобільного зв'язку, авіабізнесу, не хочуть витрачати багато часу на прослуховування дзвінків з метою узагальнення інформації або ж виявлення яких-небудь фактів порушень.

Висновки

Важливе положення Data Mining - не тривіальність розшукуваних шаблонів. Це означає, що знайдені шаблони повинні відображати неочевидні, несподівані (unexpected) регулярності в даних, складові так званих прихованих знань (hidden knowledge). До суспільства прийшло розуміння, що сирі дані (raw data) містять глибинний пласт знань, при грамотній розкопці якого можуть бути знайдений справжні самородки.

Сфера застосування Data Mining нічим не обмежена - вона скрізь, де є які-небудь дані. Але в першу чергу методи Data Mining сьогодні заінтригували комерційні підприємства. Досвід багатьох таких підприємств показує, що віддача від використовування Data Mining може досягати 1000%. Наприклад, річна економія мережі універсамів Великобританії за рахунок упровадження Data Mining складає 700 тис. Data Mining представляє велику цінність для керівників і аналітиків в їх повсякденній діяльності. .

Настала черга вчених і інженерів опонувати Data Mining як інструмент для проведення наукових досліджень (генетика, хімія, медицина, нанотехніка і ін.). Розробники національної Grid інфраструктури України зв'язують майбутнє Data Mining з їх використанням в якості Grid інтелектуальних додатків, вбудованих в віртуальні чи корпоративні сховища даних, а також в мережу Світових Центрів Даних. Але міждисциплінарна задача потребує об’єднання зусиль українських фахівців (може, в межах відповідної державної Програми) , які працюють в вузах і академічних інститутах і які добре знаються в математичних методах і мають досвід створення багатьох унікальних алгоритмів обробки інформації, щоб створити сучасну Data Mining систему з видатними можливостями.

Література

1.Чубукова И. А. Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. — 382 с. (http://www.intuit.ru/department/database/datamining/)

*2.*Дюк В. Data Mining: учебный курс (+CD)/.Дюк В., Самойленко А. — СПб: Изд. Питер, 2001. — 368 с.

3.Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? - Tandem Computers Inc., 1996.

4.Кречетов Н.. Продукты для интеллектуального анализа данных. - //Рынок программных средств, N14-15_97, с. 32-39.

5. Киселев М. Средства добычи знаний в бизнесе и финансах/Киселев М., Соломатин Е.. . - //Открытые системы, № 4, 1997, с. 41-44.

6**.** Data Mining and Image Processing Toolkits (http://datamining.itsc.uah.edu/adam/)

7.Барсегян Ф. Методы и модели анализа данных OLAP и DataMining./Барсегян Ф., Куприянов М., Степанеенко В., Холод И. -СПб БХВ-Петербург,2008.

.8. Data Mining, Web Mining, Text Mining, and Knowledge Discovery (http://www.kdnuggets.com)