УДК 004.8; 681.3.06-519.852.6

GRID ТА ІНТЕЛЕКТУАЛЬНА ОБРОБКА ДАНИХ DATA MINING

А.І. ПЕТРЕНКО

Обговорюються відмінності застосувань методів Data Mining від класичних статистичних методів аналізу і OLAP-систем. Розглядаються типи закономірностей, які виявляються цими методами у процесі розв'язання різноманітних задач (асоціація, класифікація, послідовність, кластеризація, прогнозування). Описуються сфери застосування Data Mining. Наводиться приклад системи ADaM, що працює в середовищі Grid і дистанційно обробляє наукові дані.

ВСТУП: ПЕРСПЕКТИВИ ТЕХНОЛОГІЇ DATA MINING

Нещодавно в Україні почали функціонувати світовий центр даних «Геоінформатика і сталий розвиток» і національна Grid-інфраструктура (академічний і освітянський сегменти), тому 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 викликав бурхливе зростання об'ємів збираної і аналізованої інформації. Об'єми даних настільки значні, що людина просто не спроможна проаналізувати їх самостійно, хоча необхідність проведення такого аналізу цілком очевидна, адже в цих «сирих даних» закладено знання, які можуть бути використані 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. За визначенням Г. Піатецького–Шапіро (G. Piatetsky–Shapiro,GTE Labs), одного з ведучих світових експертів у даній області, Data Mining — це дослідження і виявлення алгоритмами, засобами

штучного інтелекту в «сирих даних» прихованих структур, шаблонів або залежностей, які

1) раніше не були відомі;
2) нетривіальні;
3) практично корисні;
4) доступні для інтерпретації людиною і необхідні для ухвалення рішень в різних сферах діяльності.

Специфіка сучасних вимог до продуктивної переробки інформації:

- дані мають необмежений обсяг;
- дані є різнорідними (кількісними, якісними, текстовими);
- результати — конкретні та зрозумілі;
- інструменти для обробки «сирих даних» — прості у використанні.

Традиційна математична статистика, яка довгий час претендувала на роль основного інструменту аналізу даних, не відповідала новим проблемам. Головна причина — концепція усереднювання по вибірці, що тягне за собою операції над фіктивними величинами. Методи математичної статистики виявилися корисними, головним чином, для перевірки наперед сформульованих гіпотез і для «грубого розвідувального аналізу», який є основою оперативної аналітичної обробки даних OLAP.

Основа сучасної технології Data Mining — концепція шаблонів (pattern), що відображають фрагменти багатоаспектних взаємостосунків 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 видом 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, які 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 т.ін. (рис. 1 [1]).

Потенціал Data Mining дає «зелене світло» розширенню меж застосування цієї технології. Щодо перспектив Data Mining, то можливі такі напрями розвитку:

- виділення типів предметних галузей з їх евристиками, формалізація яких полегшить рішення відповідних задач Data Mining, що відносяться до цих галузей;

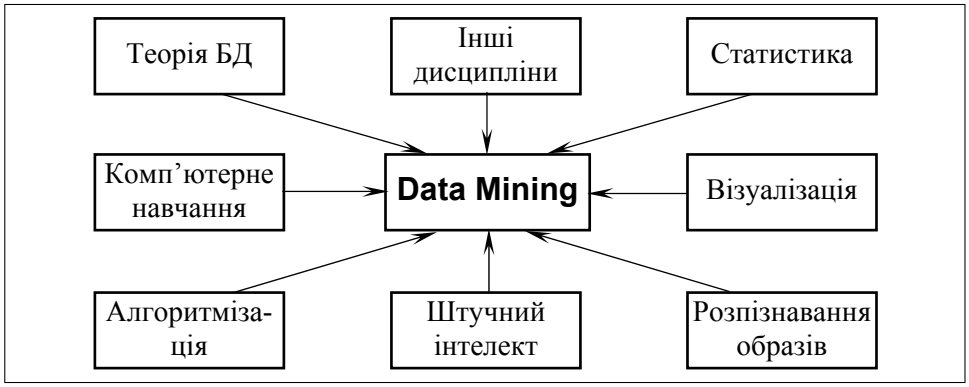

Рис. 1. Data Mining як міждисциплінарна галузь

- створення формальних мов і логічних засобів, за допомогою яких будуть формалізовані міркування і автоматизація яких стане інструментом рішення задач Data Mining у конкретних предметних галузях;
- створення методів Data Mining, здатних не тільки «витягувати» з даних закономірності, але й формувати деякі 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 у короткостроковій перспективі, то очевидно, що розвиток цієї технології здебільшого скерований на галузі, пов'язані з Grid-системами для e-Science. Можливості e-Science характеризують обчислювальну інфраструктуру, яка складається з трьох концептуальних рівнів (рис. 2).

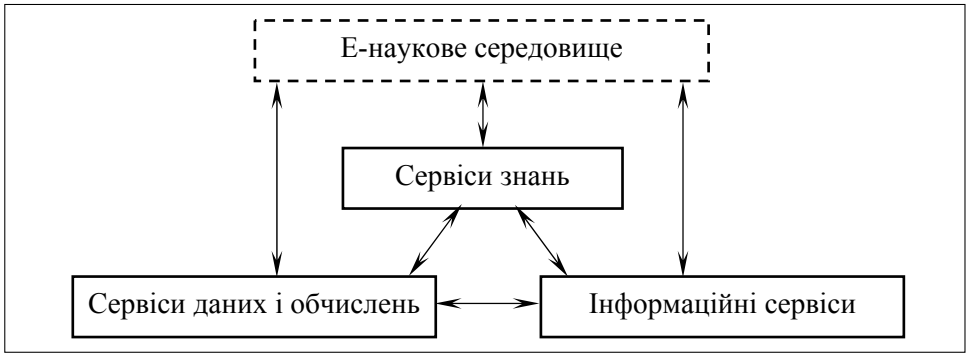

Рис. 2. Трирівнева архітектура Grid-сервісів

1. **Сервіси даних/обчислень**. Цей рівень містить інформацію про розташування обчислювальних ресурсів, виділених на обчислення, та про засоби передавання даних між різними обчислювальними ресурсами. Він може опрацьовувати великі обсяги даних, забезпечуючи швидкі мережі, і надавати різноманітні ресурси як єдиний метакомп'ютер.

2. **Інформаційні сервіси**. Вказує, яким чином інформація передається, зберігається, хто має до неї доступ. Тут інформація виступає як дані зі значенням. Наприклад, виявлення цілого числа як температури процесу реакції, розпізнавання, що рядок — ім'я людини.

3. **Сервіси знань**. Надає спосіб, яким знання придбане, використовується, знайдено, опубліковане, 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 основі якої знаходяться дані, наступний рівень — інформація, потім йде рішення, завершує піраміду рівень знання. При просуванні вгору по інформаційній піраміді об'єми даних переходять в цінність рішень, тобто цінність знань. Як видно з рис. 2, даний процес є циклічним. Ухвалення рішень вимагає інформації, заснованої на даних. Дані забезпечують інформацію, що підтримує рішення, і т.д.

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

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. У технології Data Mining гармонійно об'єдналися строго формалізовані методи і методи неформального аналізу, тобто кількісний і якісний аналізи даних.

До методів і алгоритмів Data Mining належать: штучні нейронні мережі, дерева рішень, символьні правила, методи найближчого сусіда і *k*-найближчого сусіда, метод опорних векторів, байесові мережі, лінійна регресія, кореляційно-регресійний аналіз; ієрархічні методи кластерного аналізу, неієрархічні методи кластерного аналізу, у тому числі алгоритми *k*-середніх і *k*-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, які використовуються в технології Data Mining, — це відомі математичні алгоритми і методи. Новим є те, що їх можна застосовувати при рішенні тих або інших конкретних проблем. Це обумовлено новими властивостями технічних і програмних засобів. Слід зазна-

чити, що більшість методів Data Mining розроблено в рамках теорії штучного інтелекту.

Єдиної думки щодо того, які задачі слід відносити до Data Mining, немає. Більшість авторитетних джерел називає такі: класифікація, кластеризація, прогнозування, асоціація, візуалізація, аналіз і виявлення відхилень, оцінювання, аналіз зв'язків, підведення підсумків. Розглянемо деякі з них.

Класифікація (Classification). Найпростіша і поширеніша задача Data Mining. У результаті рішення цієї задачі виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних — класи. За цими ознаками новий об'єкт можна віднести до того або іншого класу. Для розв'язання задачі класифікації можуть використовуватися методи найближчого сусіда (Nearest Neighbor), *к*-найближчого сусіда (*к*-Nearest Neighbor), байесові мережі (Bayesian Networks), індукція дерев рішень, нейронні мережі (neural networks).

Кластеризація (Clustering). Логічне продовження ідеї класифікації. Ця 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 нейронних мереж (карти Кохонена), які самоорганізуються без вчителя.

Асоціація (Associations). Відшукуються закономірності між зв'язаними подіями в наборі даних. Відмінність асоціації від двох попередніх задач Data Mining: пошук закономірностей здійснюється не на основі властивостей об'єкту, що аналізується, а між декількома подіями, що відбуваються одночасно.Самий відомий алгоритм рішення задачі пошуку асоціативних правил — алгоритм apriori.

Послідовність *(*Sequence), або послідовна асоціація *(sequential association)*. Дозволяє знайти тимчасові закономірності між транзакціями. Задача послідовності подібна асоціації, але її метою є встановлення закономірностей не між одночасними подіями, а між подіями, які відбуваються з деяким певним інтервалом у 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). На основі особливостей існуючих даних оцінюються пропущені або ж майбутні значення цільових чисельних показників. Для вирішення таких задач широко застосовуються методи математичної статистики, нейронні мережі і т. ін.

Візуалізація (Visualization, Graph Mining). Створюється графічний образ аналізованих даних. Для вирішення цієї задачі використовуються графічні методи, які показують наявність закономірностей у даних. Приклад методів візуалізації — представлення даних в 2-D і 3-D вимірюваннях.

Підведення підсумків (Summarization). Опис конкретних груп об'єктів з аналізованого набору даних та ін.

Задачі Data Mining, залежно від моделей можуть бути ***дескриптивними*** і ***прогнозуючими***. У результаті розв'язання описових (descriptive) задач аналітик одержує шаблони, які описують дані, що піддаються інтерпретації. Ці задачі надають загальну концепцію аналізованих даних, визначають інформативні, підсумкові, відмітні особливості даних. Прогнозуючі (predictive) задачі грунтуються на аналізі даних, створенні моделі, передбаченні тенденцій або властивостей нових або невідомих даних.

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 першої стадії, оскільки в реалізації багатьох методів (зокрема, нейронних мереж і дерев рішень) передбачено розподіл загальної множини даних на навчальні і перевірочні, і останні дозволяють контролювати достовірність отриманих результатів.

Стадія 3. Аналіз виключень. Виявлення і пояснення аномалій, знайдених у закономірностях.

Вільний пошук (Discovery). Дослідження набору даних з метою пошуку прихованих закономірностей. Попередні гіпотези щодо виду закономірностей тут не визначаються. **Закономірність** (law) — істотний і постійно повторюваний взаємозв'язок, що визначає етапи і форми процесу становлення та розвитку різних явищ або процесів.

Система Data Mining на цій стадії визначає шаблони, для отримання яких в системах OLAP, наприклад, аналітику необхідно обмірковувати і створювати множину запитів. Тут же аналітик звільняється від такої роботи — шаблони шукає за нього система. Особливо корисно застосовувати даний підхід у надвеликих базах даних, де встановити закономірність шляхом створення запитів достатньо складно, для цього необхідно перепробувати безліч різноманітних варіантів. Вільний пошук — це такі дії:

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

Описані дії в рамках стадії вільного пошуку виконуються при допомозі:

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

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

Прогностичне моделювання (Predictive Modeling). Друга стадія Data Mining. Використовує результати роботи першої стадії. Тут знайдені закономірності використовуються безпосередньо для прогнозування. Прогностичне моделювання — це такі дії:

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

У процесі прогностичного моделювання розв'язуються задачі класифікації і прогнозування. При розв'язанні задачі класифікації результати роботи першої стадії (індукції правил) використовуються для віднесення нового об'єкта з певною ймовірністю до одного з відомих, наперед визначених класів на підставі заданих 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 вільний пошук і прогностичне моделювання з точки зору логіки. Вільний пошук розкриває загальні закономірності. Він по своїй природі *індуктивний*. Закономірності, отримані на цій стадії, формуються від часткового до загального. У 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 групи об'єктів на підставі:

- знання класу, до якого належать досліджувані об'єкти, та
- загального правила, що діє в межах даного класу об'єктів.

Аналіз виключень (forensic analysis). Третя стадія Data Mining. Аналізуються виключення або аномалії, виявлені в знайдених закономірностях. Дія, виконувана на цій стадії, — виявлення відхилень (deviation detection), для чого необхідно визначити норму, що розраховується на стадії вільного пошуку.Стадія аналізу виключень може бути використана як очищення даних.

ПРАКТИЧНІ РЕАЛІЗАЦІЇ DATA MINING

Сьогодні у світі існують декілька комерційних і фірмових систем (PolyAnalyst,Weka, Orange Canvas, SVM^{light6}, Cognos та ін.)[4, 8]. Вартість масових систем від $1000 до $10000. Кількість інсталяцій масових продуктів 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 із взаємодіючих компонентів. Їх можна для різних прикладних задач включати у спеціалізовані додатки. ADaM містить понад 100 компонентів, які можуть бути конфігуровані так, щоб на замовлення 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 об-

робки зображень, які можуть бути легко пристосовані до численних рішень і задач. Приклади компонентів ADaM наведено нижче.

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 Cluster ing • Isodata • *k*-Means • *k*-Mediods • Maximin	***Методи оцінки властивостей*** • Backward Elimina tion • Forward Selection • Principal Compo nents • RELIEF (filter-based feature selection) • Removing Attributes • Checking Range ***Методи розпізнавання образів*** • Accuracy Measures • Data Cleaning • *k*-Fold Cross Valida tion • Vector Magnitude • Merging Patterns • Normalization • Sampling • Subsetting • Statistics • Cleaning Outliers • Comparing Image File • Comparing ASCII files • Discretization	• Magnitude Compu tation ***Методи асоціації*** • Apriori ***Методи оптимізації*** • Genetic Algorithm • Hill Climbing • Simulated Annealing ***Базові перетворення зображень*** • 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 Circum scription • Making Region • Marking Region ***Методи фільтрації*** • Dilation • Energy Erosion • Fast Fourier Trans-fer • 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

Центр 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 торнадо — на рис. 4.

Онтологія — це засіб опису семантики проблемної області за допомогою словника і підібраної специфікації існуючих в ній відношень та обмежень, які забезпечують інтеграцію словника. Інформаційні онтології створюються завжди з конкретною метою — рішення конструкторських задач — і 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, яка забезпечує семантичні метадані для наборів даних і 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.

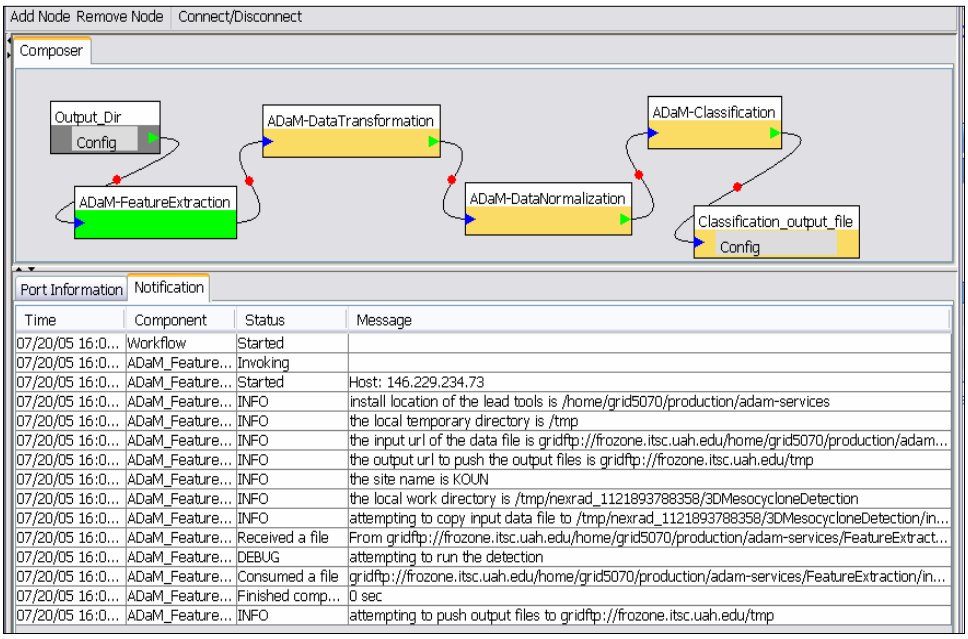

Рис. 3. Приклад формування завдання для Data Mining

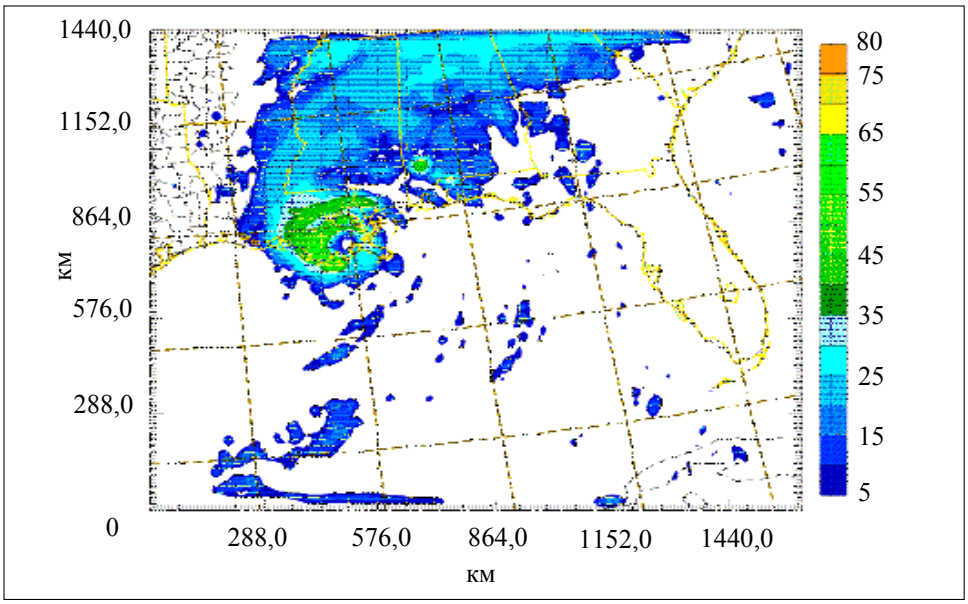

Рис. 4. Вихідна інформація Data Mining

Спільно з академічними установами, Урядом і промисловістю ITSC встановлює регіональну оптичну мережу, пов'язану із національними і міжнародними дослідницькими і освітніми мережами. Дослідження ITSC в об-

числювальних мережах високої продуктивності містять розробку паралельних алгоритмів і оцінку продуктивності та 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 давачів, використовуються для відстежування сотень цільових об'єктів при моделюванні військових дій.

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 основні сфери застосування технології Data Mining більш детально: наука, бізнес, роздрібна торгівля і Web-напрям [1,5,7].

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 для 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, 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.

У ***молекулярній генетиці і генній інженерії*** виділяють окремий напрям Data Mining, який має назву «аналіз даних у мікромасивах (Microarray Data Analysis*, MDA)*». Деякі застосування цього напряму:

- нова молекулярна мета для терапії;
- рання і більш точна діагностика;
- поліпшення та індивідуальний підбір видів лікування;
- фундаментальні біологічні відкриття.

Приклади використовування Data Mining — молекулярний діагноз деяких найсерйозніших захворювань; відкриття того, що генетичний код дійс-

но може передбачати вірогідність захворювання; винахід 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 у 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 деяких специфічних особливостей побудови з'єднань, які можуть складатися із тисячі елементів.

Основні задачі Data Mining у ***промисловому виробництві*** :

- комплексний системний аналіз виробничих ситуацій;
- короткостроковий і довгостроковий прогнози 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 та 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;
- візуалізація результатів аналізу, підготовка 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 прогнозувати популярність певних характеристик і знання того, які характеристики звичайно замовляються у сукупності. Виробникам потрібно також передбачати число клієнтів, що подадуть гарантійні заявки, і середню вартість заявок. Авіакомпанії можуть знайти групу клієнтів, яких даними заохочувальними заходами можна спонукати літати більше. Наприклад, одна авіакомпанія виявила категорію клієнтів, які здійснювали багато польотів на короткі відстані, не накопичуючи достатньо миль для вступу до їх клубів, тому вона змінила правила прийому в клуб, щоб заохочувати число польотів так само, як і милі.

2. Data Mining для вирішення бізнес-задач

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

- *Виявлення шахрайства з кредитними картками*. Шляхом аналізу минулих транзакцій, які згодом були визнані шахрайськими, банк визначає деякі стереотипи такого шахрайства.
- *Сегментація клієнтів*. Розділяючи клієнтів на різні категорії, банки здійснюють свою маркетингову політику більш цілеспрямовано і результативно, 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 ***електронній 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.

3. Роздрібна торгівля. Збирається докладна інформація про кожну окрему купівлю із використанням кредитних карток з маркою магазину і комп'ютеризованих систем 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 придбати сукупно. Знання купівельної корзини необхідне для поліпшення реклами, вироблення стратегії створення запасів товарів і способів їх розкладки у 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 дізнаватися про характер потреб різних категорій клієнтів з певною поведінкою, наприклад, тих, хто купує товари відомих дизайнерів або відвідує розпродажі. Ці знання потрібні для розробки точно направлених економічних заходів щодо просування товарів.

4. Web Mining

Web Mining можна перекласти як «здобич даних у 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 клієн-

том Web-магазина, яка група цих клієнтів приносить найбільший дохід, які інтереси певного відвідувача або групи відвідувачів.

Технологія Web Mining містить методи, здатні на основі даних сайту знайти нові, раніше невідомі знання і надалі використовувати їх 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 реалізації Web Mining перед розробниками виникає два типи задач: перший — збір даних, другий — використовування методів персоніфікації. У результаті збору деякого об'єму персоніфікованих ретроспективних даних про конкретного клієнта система накопичує інформацію про нього і може рекомендувати йому, наприклад, певні набори товарів або послуг. На основі інформації про всіх відвідувачів сайту Web-система може виявити 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»), у яких збираються дані більш ніж з одного джерела. Будуються вони комбінуванням функціональності різних програмних інтерфейсів і джерел даних.

Машапи вже застосовуються як

- сервіси агрегування (інформацію з різних джерел розміщують в одному місці);
- збирачі даних (із даних з різних джерел створюють новий сервіс (тобто агрегування));
- контролери змісту (відслідковують, фільтрують, аналізують та дозволяють пошук сервісів);
- сервісні збирачі.

5. Text Mining (інтелектуальний аналіз текстів)

Text Mining містить 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 реалізують цю задачу, повинні деяким чином оперувати природною людською мовою і при цьому розуміти семантику аналізованого тексту.

6. Call Mining (інтелектуальний аналіз дзвінків)

Технологія Call Mining об'єднує в собі розпізнавання мови, її аналіз і Data Mining. Її мета — спрощення пошуку даних в аудіоархівах, які містять записи переговорів між операторами і клієнтами. За допомогою цієї технології оператори можуть знаходити недоліки в системі обслуговування клієнтів, а також можливості збільшення продажів і виявляти тенденції в зміні контингенту клієнтів. Аналітики відзначають, що за останні роки інтерес до систем на основі Call Mining значно зріс. Це пояснюється тим, що менеджери вищої ланки компаній, які 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

Важлива позиція Data Mining — нетривіальність 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 знайдені 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 ефект 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 у їх повсякденній діяльності.

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

ЛІТЕРАТУРА

1. *Чубукова И.А.* Data Mining: учебное пособие. — М.: Интернет-ун-т информ. технологий. БИНОМ. Лаборатория знаний, 2006. — 382 с. (http://www.intuit.ru/department/database/datamining/).
2. *Data Mining*: учебный курс (+CD) / В. Дюк и др. — СПб.: Питер, 2001. — 368 с.
3. *Knowledge* Discovery Through Data Mining: What Is Knowledge Discovery? — Tandem Computers Inc., 1996. — 306 p.
4. *Кречетов Н.* Продукты для интеллектуального анализа данных // Рынок программных средств. — 1997. — № 14–15. — С. 32–39.
5. *Средства* добычи знаний в бизнесе и финансах / М.Киселев и др. // Открытые системы. — 1997. — № 4. — С. 41–44.
6. *Data* Mining and Image Processing Toolkits. — http://datamining.itsc.uah. edu/adam/.
7. *Методы* и модели анализа данных OLAP и Data Mining / Ф. Барсегян, М. Куприянов, В. Степаненко, И. Холод. — СПб.: БХВ. — 2008. — 267 с.
8. *Data* Mining, Web Mining, Text Mining, and Knowledge Discovery. — http://www.kdnuggets.com.

Надійшла 14.03.2008