УДК 025.4.036: 004.9(045)

Ругайн О.В., канд. техн. наук
Малярчук В.О., канд. техн. наук
Кизима С.В.

ОСОБЛИВОСТІ ПОБУДОВИ ТА ФУНКЦІОНУВАННЯ ПОШУКОВОЇ СИСТЕМИ ЕЛЕКТРОННИХ БІБЛІОТЕК НА БАЗІ *DATA MINING*

Національний авіаційний університет

Розглянуто особливості побудови та функціонування пошукової системи електронних бібліотек на базі Data Mining. Проведено аналіз алгоритму побудови бази даних пошукової системи електронних бібліотек, наведено принципи та особливості функціонування подібних систем, а також приклади реалізації технології Data Mining у її різнопрофільних моніторингових призначеннях

Вступ

Data Mining (DM) перекладається як «здобич» або «розкопка даних». Точніше передають сенс терміну «*Data Mining*» наступні визначення:

- «виявлення знань в базах даних» (*knowledge discovery in databases*);
- «інтелектуальний аналіз даних».

DM – це технологія виявлення прихованих взаємозв'язків усередині великих баз даних. До таких баз даних можна віднести і електронні бібліотеки, число і потужність яких постійно зростають. Тому сфера застосування *DM* тут необмежена.

DM є мультидисциплінарною областю, яка розвивається на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних та ін. Звідси велика кількість методів і алгоритмів, реалізованих в різних системах *Data Mining*, що діють.

Не дивлячись на велику кількість методів *DM*, пріоритет поступово все більш зміщується у бік логічних алгоритмів пошуку в даних *if-then* правил. З їх допомогою вирішуються завдання прогнозування, класифікації, розпізнавання образів, сегментації баз даних (БД), витягання з даних «прихованих» знань, інтерпретації даних, встановлення асоціацій в БД і ін. Результати таких алгоритмів ефективні і легко інтерпретуються.

Разом з тим, головною проблемою логічних методів виявлення закономірностей є проблема перебору варіантів за прийнятний час. Відомі методи або штучно обмежують такий перебір (алгоритми *KOPA*, *Wizwhy*), або будують дерева рішень (алгоритми *CART*, *CHAID*, *Id*3, *See*5, *Sipina* і ін.), що мають принципові обмеження ефективності пошуку *if-then* правил. Інші проблеми пов'язані з тим, що відомі методи пошуку логічних правил не підтримують функцію узагальнення знайдених правил і функцію пошуку оптимальної композиції таких правил. Вдале вирішення вказаних проблем може скласти предмет нових конкурентоздатних розробок.

Етапи побудови DM-системи

Що ж потрібно для створення на базі *DM* пошукової системи електронних бібліотек у якості її спеціалізованого додатку? Щоб розробити такий продукт, необхідно виконати ряд кроків:

1. Встановити масштаби проекту (локальна, корпоративна, регіональна мережа електронних бібліотек), що визначають, які дані необхідно зібрати.

2. Розробити базу даних визначеного проекту для *DM*. Необхідна інформація може бути розподілена по декількох базах. Дані з різних баз необхідно консолідувати і усунути невідповідності. Ефективний аналіз загального фонду електронних бібліотек вимагає корпоративного сховища даних, що з погляду вкладень обходиться дешевшим, ніж використання окремих вітрин.

Відзначимо, що таке сховище надає не тільки ефективний спосіб обліку, зберігання та пошуку бібліографічних джерел, а також є інструментом для проведення унікальних досліджень в різних галузях знань,

усуває необхідність у використанні інших вітрин і стає ідеальною основою для *DM*-проекту.

Ще один важливий момент – очищення даних. Тут мається на увазі перевірка на цілісність і обробка відсутніх значень. Точність методів *DM* залежить від якості інформації, яка полягає в основі побудови такого сховища бібліографічних даних.

Слід відмітити, що перші два етапи можуть зайняти більш ніж половину бюджету часу, відведеного на весь проект.

3. Дати кількісні та якісні оцінки елементам даних. Ця робота потребує ретельної організації щодо залучення кваліфікованих експертів в кожної наочній області, що допоможе вирішити питання формування детермінованих інформаційних масивів.

4. Застосувати алгоритми *DM* для визначення відносин між даними. Як правило, для виявлення потрібних залежностей експерти використовують декілька різних алгоритмів. Одні з них підійдуть на перших етапах процесу, інші на пізніших.

5. Дослідити співвідношення, виявлені на попередніх етапах, на застосовність в масштабах проекту. Експерти в наочній області визначають, чи є ті або інші відносини дуже специфічними або дуже загальними і вказує, в яких областях слід продовжити аналіз.

6. Узагальнити результати у вигляді систематизованого масиву даних, в якому будуть враховані всі відносини, що інтерпретуються.

Мета першого прототипу проекту полягає в тому, щоб скоротити кількість помилок в базі даних (маються на увазі перший, другий, третій і п'ятий етапи).

Для усвідомлення всіх тонкощів досліджуваних даних іноді потрібно декілька ітерацій. Для пізніших прототипів важливі третій, четвертий і п'ятий етапи.

Також на розподіл часу для *DM* проекту впливають і інші чинники: тип кінцевого застосування, наявність і стан сховища даних. Для якісної реалізації пошукової функції проекту більше часу доведеться витратити на перших трьох етапах, а для ефективного аналізу отриманих результатів – на останніх трьох.

Роботи по оптимізації сховища даних електронних бібліотек для пошукових систем діляться на роботу з внутрішніми і зовнішніми чинниками.

До внутрішніх чинників можливо долучити наступні:

1. Складання семантичного ядра (основні ключові запити за проектом).

2. Роботи по реорганізації структури сховища і подачі інформації.

3. Оптимізація і побудова метатегів.

4. Оптимізація контенту сховища (його наповнення).

5. Коректування програмного коду адресних елементів сховища і виправлення допущених помилок.

6. Розробка системи перехресних внутрішніх посилань.

7. Підключення статистики, аналіз якісних і кількісних характеристик відвідуваності сховища користувачами його пошукової системи.

8. Коректування внутрішніх чинників, з урахуванням результатів.

До зовнішніх можливо долучити наступні чинники:

1. Реєстрація в каталогах і рейтингах.

2. Посторінкова реєстрація в глобальних пошукових системах.

3. Аналіз індексування сховища пошуковими системами.

4. Аналіз поточних позицій сховища по запитах з семантичного ядра, а також додатковим запитам.

5. Взаємодія з адміністраторами Інтернет-ресурсів для розміщення інформації про проект.

6. Тематичний обмін посиланнями та інші.

Ключове завдання пошукового просування – добитися максимально високого ступеня рельовантності сховища запитам цільової аудиторії, а, отже, збільшення його відвідуваності

Особливості функціонування пошукової DM-системи

Розглянемо деякі визначення техно-огії DM с точки зору її функціонування. *DM* – це процес аналізу, виділення і представлення деталізованих (*detailed data*) даних неявної конструктивної інформації. У випадку, який досліджується у даній роботі, під неявною конструктивною інформацією розуміємо інформацію, яка схована у кожному елементі (бібліографічному джерелі) сховища електронних бібліотек, долучених до проекту. Також *DM* – це процес виділення (*selecting*), дослідження і моделювання великих об'ємів даних для виявлення невідомих до цього структур (*patterns*) з метою їх дослідження за відповідними критеріями. *DM* – це процес, мета якого – виявити нові значущі кореляції, зразки і тенденції в результаті просіювання великого об'єму даних, що зберігаються, з використанням методик розпізнавання зразків плюс застосування статистичних і математичних методів (*Gartner Group*). *DM* – це процес автоматичного виділення дійсної, ефективної, раніше невідомої і абсолютно зрозумілій інформації з великих баз даних і використання її для досягнення мети дослідження.

На побутовому рівні це звучить приблизно так: «Ви мучите дані, поки вони не признаються».

Відзначимо, що процес виявлення знань не є повністю автоматичним. Цей процес вимагає кваліфікованої роботи користувача. Коротше кажучи, користувач повинен знати, що він шукає, ґрунтуючись на власних гіпотезах. У результаті часто замість підтвердження наявної гіпотези процес пошуку викликає появу нових гіпотез. Все це позначається терміном *discovery-driven data mining* (*DDDM*), і терміни *Data Mining, knowledge discovery*, якій в загальному випадку відносяться до *DDDM*.

Отже, можна реалізувати той або інший пакет для *DM*, але якщо користувач не вміє правильно користуватися її пошуковим апаратом, то результат не буде адекватним очікуваному.

В цілому технологію функціонування *DM* достатньо точно визначає Григорій Піатецький-Шапіро – один із засновників цього напряму [1]: *DM* – це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних, доступних інтерпретації знань, необхідних для ухвалення рішень в різних сферах людської діяльності. *DM* являє собою синтетичну область, що ввібрала в себе досягнення штучного інтелекту, статистики, чисельних математичних методів, евристичні підходи. *DM* – найважливіша пошукова ланка, від збору даних до оцінки результатів дії, а також є інструментом для проведення унікальних досліджень в різних галузях знань. Результати *Data Mining* (емпіричні моделі, класифікаційні правила, знайдені кластери і т.п.) можна потім інкорпорувати в існуючі інституції наукових досліджень і використовувати для аналізу тенденцій розвитку технічних, технологічних та наукових напрямків.

DM забезпечує вирішення шести завдань: класифікація, кластеризація, регресія, асоціація, послідовність, відхилення.

Тому *DM* – є не один, а сукупність великого числа різних методів виявлення знань. Вибір методу часто залежить від типу наявних даних і від того, яку інформацію ви намагаєтеся отримати. Деякі методи перераховані нижче:

-об'єднання (*association*, іноді використовують термін *affinity*, що означає схожість, структурну близькість) – виділення структур, що повторюються в тимчасовій послідовності. Виявляє правила, по яких присутність одного набору елементів корелюється з іншим. Мета – знайти закономірності серед великого числа транзакцій;

-аналіз тимчасових рядів (*sequence based analysis*, інша назва – *sequential association*) – дозволяє знайти тимчасові закономірності між транзакціями;

- кластеризація (*clustering*) – угрупування записів, що мають однакові характеристики, наприклад, по близькості значень полів. Можуть використовуватися статистичні методи або нейромережі. Кластеризація часто розглядається як перший необхідний крок для подальшого аналізу даних;

- класифікація (*classification*) – віднесення запису до одного із заздалегідь певних класів;
- оцінювання (*estimation*);
- нечітка логіка (*fuzzy logic*);
- статистичні методи, що дозволяють знаходити криву, найближче розташовану до набору точок даних;
- генетичні алгоритми (*genetic algorithms*);
- фрактальні перетворення (*fractalbased transforms*);
- нейронні мережі (*neural networks*) – дані пропускаються через шари вузлів, «навчених» розпізнаванню тих або інших структур.

До *DM* можна додати ще візуалізацію даних – побудова графічного образу з даних, використання кольору. Це допомагає при загальному аналізі даних побачити аномалії та структури масивів даних, що досліджуються.

З перерахованого видно, що ця область дуже обширна, щоб одна людина могла освоїти всі методи і бути фахівцем у всій цій області.

Приклади реалізації технології Data Mining

В реальних системах *DM* тісно інтегрована з сховищами даних (*Data Warehousing – DW*) і можна сказати, що *DW* забезпечують роботу *Data Mining*, а *Data Mining* виправдовують *DW*. Наприклад, коли для *DM* потрібні нові дані, їх додають через *DW*.

Згідно архітектурі *Oracle DM* функції *DM* переносяться в базу даних (*DW*). У комплекті *Oracle Data Mining Suite Release* дані вибираються з бази даних, де відбувається все маніпулювання з даними, розкопування і залік (*mining and scoring*). Користувачі взаємодіють з цим програмним забезпеченням через графічний інтерфейс *Windows*.

Зараз існує декілька користувацьких пошукових систем, яки функціонують на принципах *DM*. Наприклад, продукт *Oracle Personalization,* якій розроблений на основі концепції бази даних в оперативній пам'яті для обробки з потрібною швидкістю великих об'ємів даних, зв'язаних з *web*. Реалізації алгоритмів *Naive Bayes* і *Associations algorithms* для паралельної комп'ютерної архітектури забезпечує проникнення в суть процесів Інтернет-користувачам як в режимі реального часу, так і неоперативному режимі. Засоби *DM* в *Oracle Personalization* реалізовані на стороні сервера і використовують можливості паралельних багатопроцесорних (*SMP*) систем, так що можна використовувати загальну потужність безлічі комп'ютерів і виконувати аналіз в «*n*» разів швидше і «розкопувати» в «*n*» разів більше даних *Web*-сайтів. Механізми рекомендацій обслуговують *Web*-сайти підприємства в режимі реального часу.

Другий продукт, *Oracle Data Miner*, розширює концепцію *Oracle Personalization* стосовно типових застосувань *DM* і пропонує аналогічний призначений для користувача інтерфейс, а також адаптовані «методології», розроблені для вирішення заздалегідь певних проблем бізнесу, – наприклад, моделі формування відповідей і лояльності (*loyalty and response modeling*). *Oracle Data Miner* пропонує інтерфейс до більшості стандартних функцій *Oracle Data Mining Suite Release* 3.7.

З продуктом *Oracle Data Miner* співробітники, ведучі маркетинг на основі бази даних можуть прийняти значення за умовчанням, щоб підготувати дані для аналізу і отримати специфічну для додатку допомогу і ради. Результати *DM*, наприклад, «звіт про невизначені ризики» («*churn jeopardy report*»), можуть бути автоматично створені і показані з використанням засобів корпорації *Oracle* для формування запитів і звітів.

Компанія Мегапьютер проводить і пропонує на ринку сімейство продуктів для *DM* – *Polyanalyst On-line Tutor*. Система *Polyanalyst* призначена для автоматичного аналізу числових і текстових даних з метою виявлення в них раніше невідомих, нетривіальних, практично корисних і доступних розумінню закономірностей, необхідних для ухвалення оптимальних рішень в бізнесі та в інших областях людської діяльності.

В даний час *Polyanalyst* є одній з наймогутніших систем *Data Mining* в світі, реалізованих для *Intel* платформ і операційних систем *Microsoft Windows*.

Аналогічні системи *Data Mining* таких провідних виробників, як *IBM* (*Intelligent Miner, Data Miner*), *Silicon Graphics* (*SGI Miner*), *Integral Solutions* (*Clementine*), *SAS Institute* (*SAS*) працюють на середніх і великих машинах і коштують десятки і навіть сотні тисяч доларів. Завдяки унікальній технології «Еволюційного програмування», і іншим інноваційним математичним алгоритмам, *Polyanalyst* поєднує в себе високу продуктивність «великих систем» з низькою вартістю, властивою програмам для *Windows*.

Polyanalyst – один з небагатьох комерційних продуктів, в якому реалізовані не тільки методи аналізу числових даних, але і алгоритми *Text Mining*, – аналізу текстової інформації. Протягом своєї більш, ніж 10-річній історії, пакет безперервно розвивається, компанія-виробник додає нову функціональність, нові математичні модулі, планується портація системи на *Unix* платформи.

Polyanalyst набув широкого поширення в світі. Більше 500 інсталяцій в 20 країнах світу, серед користувачів системи значний список складають найбільші світові корпорації: *Boeing, 3m, Chase Manhattan Bank, Dupont, Siemens* та інші. *Polyanalyst* – універсальна система *Data Mining*, вона з успіхом застосовується в різних областях: у вирішенні бізнес-задач (*direct marketing, cross-selling, customer retention*), в соціологічних дослідженнях, в прикладних наукових і інженерних завданнях, в банківській справі, в страхуванні і медицині.

Висновки

Застосування *DM*-технологій для побудови пошукової системи електронних бібліотек є привабливим з різних точок зору. Реалізація цього проекту обіцяє суттєві переваги по відношенню до інших моніторингових систем. Особливо слід зазначити, що *DM*-система одночасно з виконанням пошукових функцій може бути інструментом для проведення унікальних досліджень в різних галузях знань. Однак у зв'язку з великими витратами прийняття остаточного рішення відносно інвестування цього проекту в значній мірі залежить від конкурентоспроможності його фінансово-економічних показників на етапах розробки, побудови та експлуатації.

Список літератури

1. *Gregory Piatetsky-Shapiro*. Data mining and knowledge discovery 1996 to 2005. – Published online: 27 January 2007. – 7 p.

2. *Дюк В., Самойленко А*. Data Mining. – СПб: Издательский дом "Питер", 2001. – 368 с.

3. *Антоненко І., Баркова О.* Електронні ресурси як об'єкт каталогізації: історія питання, термінологія, форматне забезпечення // Бібл. вісн. – 2004. – №2. – С. 11–22.

4. *Армс В*. Электронные библиотеки: (Пер. с англ). – М.: ПИК ВИНИТИ, 2001. – 274 с.