УДК 004.891

к.т.н., доц. **Джулій В.М.** (ХмНУ)
к.т.н., доц. **Чешун В.М.** (ХмНУ)
к.т.н., доц. **Кривцун В.І.** (ХмНУ)
Солодєєва Л.В. (ВІКНУ)

АНАЛІЗ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОЇ ОБРОБКИ ІНФОРМАЦІЇ

У статті проведено аналіз методів інтелектуальної обробки інформації. Показано, що інтелектуальний аналіз даних - це не тільки виконання деяких складних запитів до даних, що зберігаються в базі даних, незалежно від того, чи використовуються бази даних на основі документів, чи прості неструктуровані файли. Необхідно працювати з даними, форматувати або реструктурувати їх, визначити формат інформації, на якому буде грунтуватися метод і аналіз. Якщо інформація знаходиться в потрібному форматі, можна застосовувати різні методи (окремо або в сукупності), які не залежать від необхідної базової структури даних або набору даних.

Таким чином, визначено, що існуючі програмні та апаратні засоби не вирішують повністю задачу інформаційного пошуку в корпоративних сховищах. Широко поширені системи інформаційного пошуку в корпоративному сегменті в слабкій мірі враховують зміст оброблюваних документів і взаємозв'язок їх змісту з семантикою предметної галузі промислового підприємства.

Ключові слова: інтелектуальна обробка інформації, класифікація, інформаційне джерело, аналіз даних, асоціація, прогнозування, дерево рішень

Вступ. Ефективність роботи сучасного промислового підприємства в інформаційному суспільстві залежить від швидкості і якості задоволення потреб у службовій інформації кожного з працівників. Інформаційні сховища корпоративних інформаційних систем можуть досягати величезних розмірів, що сильно ускладнює пошук. Необхідна інформація часто розподілена по різних інформаційних системах всередині підприємства, її інтеграція ускладнена через неоднозначність використовуваної термінології, специфічної структури компонентів інформаційних сховищ, різного рівня компетентності співробітників підприємства. Виникає парадоксальна ситуація, коли користувач має доступ до інформації, але не може її отримати.

Згідно з даними досліджень, співробітники, що використовують персональний комп'ютер для виконання посадових обов'язків, в середньому витрачають 9,5 годин на тиждень на пошук інформації. Для деяких областей діяльності пошук може займати до 50% робочого часу працівника. Неможливість знайти і використовувати потрібну інформацію призводить до того, що співробітники перестворюють її, витрачаючи на це близько трьох годин робочого часу на тиждень. Забезпечення співробітників сучасного підприємства зручними засобами інформаційного пошуку є важливим способом підвищення ефективності праці і надає переваги над конкурентами, які подібними засобами не володіють.

Постановка задачі. Існуючі програмні та апаратні засоби не вирішують повністю задачу інформаційного пошуку в корпоративних сховищах. Широко поширені системи інформаційного пошуку в корпоративному сегменті засновані на методах DataMining, частотного пошуку і в слабкій мірі враховують зміст оброблюваних документів та взаємозв'язок їх змісту з семантикою предметної галузі промислового підприємства.

У сучасному світі інформаційних технологій спостерігається постійне зростання інтересу до методів інтелектуальної обробки даних. Ці тенденції визначаються, з одного боку, зростаючими обсягами інформації, що зберігається та інтеграцією сховищ даних, а з іншого - постійним зростанням попиту на інформаційні послуги, пов'язані з обробкою цих даних. Зазначені фактори проявляються як на рівні корпоративних інформаційних систем в галузі медицини, економіки, прогнозування і т.д., так і на глобальному рівні, де одним з наслідків є потреба в інтелектуальній обробці даних. Основою багатьох систем інтелектуального пошуку та обробки інформації є машини логічного висновку, що реалізують ті чи інші методи доказів логічних теорем. Однією з основних проблем побудови

ефективних машин висновку є швидкість обчислень, яка зазвичай визначається як кількість обчислювальних операцій, необхідних для обробки бази знань заданого обсягу. Слід зазначити, що швидкість обчислень також залежить від ступеня виразності мови опису предметної області. Так, існують сучасні реалізації машин виводу для логіки висловлювань (побудовані, наприклад, на базі табличного (tableaux) методу) які дозволяють досягти лінійних обчислювальних витрат на здійснення висновку щодо обсягу бази знань. Проте висновок на складних моделях, описаних на мові логіки предикатів або дескриптивної логіки часто призводить до поліноміального або навіть експоненціального зростання обчислювальних витрат в залежності від обсягу вихідної специфікації.

По суті, інтелектуальний аналіз даних - це обробка інформації і виявлення в ній моделей і тенденцій, які допомагають приймати рішення. Принципи інтелектуального аналізу даних відомі протягом багатьох років, але з появою великих об’ємів даних вони отримали ще більш широке поширення. Великі об’єми даних привели до вибухового зростання популярності більш широких методів інтелектуального аналізу даних, тому, що інформації стало набагато більше, і вона за своєю природою і змістом стає більш різноманітною і обширною. При роботі з великими наборами даних вже недостатньо відносно простої і прямолінійної статистика. Вимоги бізнесу призвели від простого пошуку і статистичного аналізу даних до складнішого інтелектуального аналізу даних. Для вирішення бізнес-завдань потрібно такий аналіз даних, який дозволяє побудувати модель для опису інформації і в кінцевому підсумку призводить до створення результуючого звіту. Цей процес ілюструє рис. 1.

Процес аналізу даних, пошуку та побудови моделі часто є ітеративним, так як потрібно розшукати і виявити різні відомості, які необхідно витягти. Необхідно також розуміти, як зв'язати, перетворити і об'єднати їх з іншими даними для отримання результату. Після виявлення нових елементів і аспектів даних підхід до виявлення джерел і форматів даних з наступним зіставленням цієї інформації з заданим результатом може змінитися.

Рис. 1. Модель опису інформації

Інструменти інтелектуального аналізу даних. Інтелектуальний аналіз даних - це не лише використовувані інструменти або програмне забезпечення баз даних. Інтелектуальний аналіз даних можна виконати з відносно скромними системами баз даних і простими інструментами, включаючи створення своїх власних, або з використанням готових пакетів

програмного забезпечення. Складний інтелектуальний аналіз даних опирається на досвід і алгоритми, визначені за допомогою існуючого програмного забезпечення і пакетів, причому з різними методами асоціюються різні спеціалізовані інструменти. Наприклад, IBM SPSS, який йде корінням в статистичний аналіз та опитування, дозволяє будувати ефективні прогностичні моделі за тенденціями і давати точні прогнози. IBM InfoSphere Warehouse забезпечує в одному пакеті пошук джерел даних, попередню обробку і інтелектуальний аналіз, дозволяючи отримувати інформацію з вихідної бази прямо в підсумковий звіт. Останнім часом стала можлива робота з дуже великими наборами даних і кластерна/великомасштабна обробка даних, що дозволяє робити ще складніші узагальнення результатів інтелектуального аналізу даних за групами і зіставлення даних. Сьогодні доступний абсолютно новий спектр інструментів і систем, включаючи комбіновані системи зберігання та обробки даних. Можна аналізувати найрізноманітніші набори даних, включаючи традиційні бази даних SQL, необроблені текстові дані, набори "ключ/значення" і документальні бази. Кластерні бази даних, такі як Hadoop, Cassandra, CouchDB і Couchbase Server, зберігають і надають доступ до даних такими засобами, які не відповідають традиційній табличній структурі. Зокрема, більш гнучкий формат зберігання бази документів надає обробці інформації нову спрямованість і ускладнює її. Бази даних SQL суворо регламентують структуру і жорстко дотримуються схеми, що спрощує запити до них та аналіз даних з відомими форматом і структурою. Документальні бази даних, які відповідають стандартній структурі типу JSON, або файли з деякою машиночитаємою структурою теж легко обробляти, хоча справа може ускладнюватися різноманітною і мінливою структурою. Наприклад, в Hadoop, який обробляє абсолютно "сирі" дані, може бути важко виявити і витягти інформацію до початку її обробки та зіставлення.

Основні методи інтелектуального аналізу даних. Кілька основних методів, які використовуються для інтелектуального аналізу даних, описують тип аналізу і операцію з відновлення даних. На жаль, різні компанії і рішення не завжди використовують одні й ті ж терміни, що може посилити плутанину і складність. Розглянемо деякі ключові методи, як використовують ті чи інші інструменти для інтелектуального аналізу даних.

Асоціація. Асоціація (відношення), ймовірно, найбільш відомий, знайомий і простий метод інтелектуального аналізу даних. Для виявлення моделей робиться просте зіставлення двох або більше елементів, часто одного і того ж типу. Створити інструменти інтелектуального аналізу даних на базі асоціацій або відносин неважко. Наприклад, в InfoSphere Warehouse є майстер, який видає конфігурації інформаційних потоків для створення асоціацій, досліджуючи джерело вхідної інформації, базис прийняття рішень і вихідну інформацію.

Класифікація. Класифікацію можна використовувати для отримання представлення про тип покупців, товарів або об'єктів, описуючи кілька атрибутів для ідентифікації певного класу. Наприклад, автомобілі легко класифікувати за типом (седан, позашляховик, кабріолет), визначивши різні атрибути (кількість місць, форма кузова). Вивчаючи новий автомобіль, можна віднести його до певного класу, порівнюючи атрибути з відомим визначенням. Ті ж принципи можна застосувати і до покупців, наприклад, класифікуючи їх за віком і соціальної групи. Крім того, класифікацію можна використовувати в якості вхідних даних для інших методів. Наприклад, для визначення класифікації можна застосовувати дерева прийняття рішень. Кластеризація дозволяє використовувати загальні атрибути різних класифікацій з метою виявлення кластерів.

Кластеризація. Досліджуючи один або більше атрибутів або класів, можна згрупувати окремі елементи даних разом, отримуючи структурований вивід. На простому рівні при кластеризації використовується один або декілька атрибутів в якості основи для визначення кластера подібних результатів. Кластеризація корисна при визначенні різної інформації, тому що вона корелюється з іншими прикладами так, що можна побачити, як подібність і діапазони узгоджуються між собою. Метод кластеризації працює в обидві сторони. Можна

припустити, що в певній точці мається кластер, а потім використовувати свої критерії ідентифікації, щоб перевірити це.

Прогнозування. Прогнозування - це широка тема, яка розпочинається від передбачення відмов компонентів обладнання до виявлення шахрайства і навіть прогнозування прибутку компанії. У поєднанні з іншими методами інтелектуального аналізу даних прогнозування передбачає аналіз тенденцій, класифікацію, зіставлення з моделлю і відношення. Аналізуючи події або екземпляри, можна передбачати майбутнє. Наприклад, використовуючи дані по авторизації кредитних карт, можна об'єднати аналіз дерева рішень транзакцій людини з класифікацією і зіставленням з історичними моделями з метою виявлення шахрайських транзакцій. Послідовні моделі, які часто використовуються для аналізу довгострокових даних - метод виявлення тенденцій, або регулярних повторень подібних подій. Наприклад, за даними про покупців можна визначити, що в різний час року вони купують певні набори продуктів.

Дерева рішень. Дерево рішень, пов'язане з більшістю інших методів (головним чином, класифікації та прогнозування), можна використовувати або в рамках критеріїв відбору, або для підтримки вибору певних даних в рамках загальної структури. Дерево рішень починають з простого питання, яке має дві відповіді (іноді більше). Кожна відповідь призводить до наступного питання, допомагаючи класифікувати та ідентифікувати дані або робити прогнози. Дерева рішень часто використовуються із системами класифікації інформації про властивості та із системами прогнозування, де різні прогнози можуть грунтуватися на історичному досвіді, який допомагає побудувати структуру дерева рішень і отримати результат.

Комбінації. На практиці дуже рідко використовується тільки один з цих методів. Класифікація і кластеризація - подібні методи. Використовуючи кластеризацію для визначення найближчих сусідів, можна додатково уточнити класифікацію. Дерева рішень часто використовуються для побудови і виявлення класифікацій, які можна простежувати на історичних періодах для визначення послідовностей і моделей.

Обробка із запам'ятовуванням. При всіх основних методах часто має сенс записувати і згодом вивчати отриману інформацію. Для деяких методів це абсолютно очевидно. Наприклад, при побудові послідовних моделей та навчанні з метою прогнозування аналізуються історичні дані з різних джерел і примірників інформації. В інших випадках цей процес може бути більш яскраво вираженим. Дерева рішень рідко будуються один раз і ніколи не забуваються. При виявленні нової інформації, подій і точок даних може знадобитися побудова додаткових гілок або навіть зовсім нових дерев. Деякі з цих процесів можна автоматизувати. Наприклад, побудова прогностичної моделі для виявлення шахрайства з кредитними картами зводиться до визначення ймовірностей, які можна використовувати для поточної транзакції, з подальшим оновленням цієї моделі при додаванні нових (підтверджених) транзакцій. Потім ця інформація реєструється, так що наступного разу рішення можна буде прийняти швидше.

Отримання і підготовка даних. Інтелектуальний аналіз даних опирається на побудову відповідної моделі і структури, які можна використовувати для обробки, виявлення та створення необхідної інформації. Незалежно від форми і структури джерела даних, інформація структурується і організовується відповідно до формату, який дозволяє виконувати інтелектуальний аналіз даних з максимально ефективною моделлю. Аналітичні змінні для даних, отриманих з безлічі різних джерел, можна скласти в єдину, певну структуру (наприклад, створити клас покупців певних рівнів і віків або клас помилок певного типу). Залежно від джерела даних важливо вибрати правильний спосіб побудови і перетворення цієї інформації, яким би не був метод остаточного аналізу даних. Цей крок також веде до більш складного процесу виявлення, збору, спрощення або розширення інформації відповідно до вхідних даних(рис. 2).

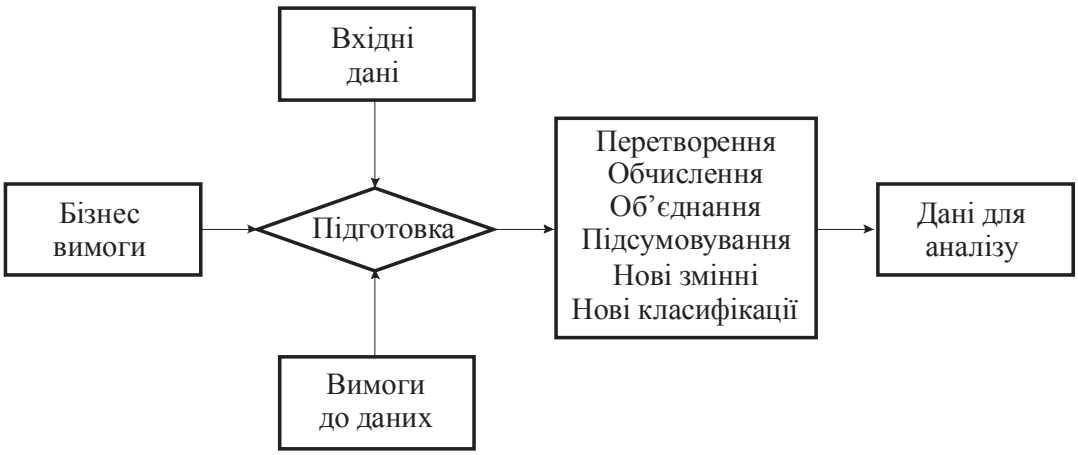

Рис. 2. Підготовка даних

Джерело даних, місце розташування і база даних впливають на те, як буде оброблятися і об'єднуватися інформація. Найбільш простим з усіх підходів часто служить опора на бази даних SQL. SQL (і відповідна структура таблиці) добре зрозуміла, але структуру і формат інформації не можна ігнорувати повністю. Наприклад, при вивченні поведінки користувачів з даними про продажі в моделі даних SQL (і інтелектуального аналізу даних в цілому) існують два основні формати, які можна використовувати: транзакційний і поведінково - демографічний. Основна проблема даних на основі документів - це неструктурований формат, який може вимагати додаткової обробки. Багато різних записів можуть містити аналогічні дані . Збір та узгодження цієї інформації з метою спрощення її обробки залежить від етапів підготовки та застосування інформації.

Висновок. Інтелектуальний аналіз даних - це не тільки виконання деяких складних запитів до даних, що зберігаються в базі даних. Незалежно від того, чи використовуєте ви SQL, бази даних на основі документів, такі як Hadoop, або прості неструктуровані файли, необхідно працювати з даними, форматувати або реструктурувати їх. Потрібно визначити формат інформації, на якому буде грунтуватися ваш метод і аналіз. Потім, коли інформація знаходиться в потрібному форматі, можна застосовувати різні методи (окремо або в сукупності), які не залежать від необхідної базової структури даних або набору даних.

Існуючі програмні та апаратні засоби не вирішують повністю задачу інформаційного пошуку в корпоративних сховищах. Широко поширені системи інформаційного пошуку в корпоративному сегменті засновані на методах DataMining, частотного пошуку і в слабкій мірі враховують сутність оброблюваних документів і взаємозв'язок їх змісту з семантикою предметної галузі промислового підприємства.

У сучасному світі інформаційних технологій спостерігається постійне зростання інтересу до методів інтелектуальної обробки даних. Ці тенденції визначаються, з одного боку, зростаючими обсягами інформації, що зберігається та інтеграцією сховищ даних, а з іншого - постійним зростанням попиту на інформаційні послуги, пов'язані з обробкою цих даних. Зазначені фактори проявляються як на рівні корпоративних інформаційних систем в галузі медицини, економіки, прогнозування і т.д., так і на глобальному рівні, де одним з наслідків потреб в інтелектуальній обробці даних. Основою багатьох систем інтелектуального пошуку та обробки інформації є машини логічного висновку, що реалізують ті чи інші методи доказів логічних теорем. Однією з основних проблем побудови ефективних машин висновку є швидкість обчислень, яка зазвичай визначається як кількість обчислювальних операцій, необхідних для обробки бази знань заданого об'єму. Для вирішення бізнес-завдань потрібно такий аналіз даних, який дозволяє побудувати модель для опису інформації і в кінцевому підсумку призводить до створення результуючого звіту.

ЛІТЕРАТУРА:

1. А.А. Барсегян Методы и модели анализа данных: OLAP и Data Mining. / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод -БХВ-Петербург; 2004г.

2. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. / Загоруйко Н.Г. - Новосибирск: Изд-во Ин-та математики, 1999. 270 с.

3. Корнеев В.В Базы данных. Интеллектуальная обработка информации. / Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В. -М.: Нолидж, 2001. 496 с.

4. Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, OLAP./ Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. - СПб: БХВ-Петербург, 2007. 275 с.

Рецензент: д.т.н., проф. Мясіщев О.А., Хмельницький національний університет

к.т.н., доц. Джулий В.Н., к.т.н., доц. Чешун В.Н.,
к.т.н., доц. Кривцун В.И., Солодеева Л.В.

АНАЛИЗ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ИНФОРМАЦИИ

В статье проведен анализ методов интеллектуальной обработки информации. Показано, что интеллектуальный анализ данных - это не только выполнение некоторых сложных запросов к данным, хранящимся в базе данных, независимо от того, используются базы данных на основе документов, или простые неструктурированные файлы. Необходимо работать с данными, форматировать или реструктурировать их, определить формат информации, на котором будет основываться метод и анализ. Когда информация находится в нужном формате, можно применять различные методы (отдельно или в совокупности), которые не зависят от необходимой базовой структуры данных или набора данных. Таким образом определено, что существующие программные и аппаратные средства не решают полностью задачу информационного поиска в корпоративных хранилищах. Широко распространены системы информационного поиска в корпоративном сегменте в слабой степени учитывают содержание обрабатываемых документов и взаимосвязь их содержания с семантикой предметной области промышленного предприятия.

Ключевые слова: интеллектуальная обработка информации, классификация, информационный источник, анализ данных, ассоциация, прогнозирования, дерево решений

V. Julie, V. Cheshun, V. Krivtsun, L. Solodeeva

ANALYSIS OF INTELLIGENT INFORMATION PROCESSING METHODS

The paper analyzed the predictive processing. It is shown that data mining - is not just doing some complex queries to the data stored in the database, regardless of whether you use the database on the basis of documents or simple unstructured files. It should work with the data format or restructure them to determine the format of information, which will be based method and analysis. When the information is in the right format, you can use a variety of methods (individually or collectively), that are independent of the required basic data structures or data set. So determined that the existing software and hardware do not fully solve the problem of information retrieval in enterprise storage. Widely distributed information retrieval system in the corporate sector in a weak weight on the content of the processed documents and their relationship with the semantics of the content of the subject area of industrial enterprises.

Keywords: intelligent information processing, classification, information source, data analysis, association, prediction, decision tree