УДК 004.912

Денис Костенко, Владислав Фрінцко, Вадим Гавриш, Ігор Коноваленко
Університет митної справи та фінансів, Україна

ОСНОВНІ ПРОБЛЕМИ ІНТЕГРАЦІЇ БАЗ ДАНИХ

Показані проблеми, які виникають під час інтеграції даних у інформаційні сховища. Розкрито проблеми рівня семантичної інтеграції даних. Запропоновано деякі способи вирішення зазначених проблем.

Ключові слова: інформаційні сховища даних, дані, семантичний підхід, онтології.

Denys Kostenko, Vladyslav Frintsko, Vadym Gavrish, Ihor Konovalenko

BASIC PROBLEMS OF DATA INTEGRATION

There are considered some problems which arise in data integrating process into information repositories. There are revealed the problems of data semantic integration level. Some ways to solve these problems are suggested.

Keywords: data repositories, data, semantic approach, ontologies.

На сьогодні у світі без зупинки з неймовірною швидкістю розвиваються інформаційні технології. З кожним днем інформації становиться все більше і релевантність запитів стає все менше. Отже, необхідно знайти зручний та не складний метод вирішення цієї проблеми.

Через технічний прогрес та розвиток пошукових технологій відбувається надзвичайне зростання обсягів доступної інформації, яка може бути корисна для вирішення важливих завдань в будь-якій сфері.

Необхідно зазначити, що внаслідок стрімкого розвитку інформаційних технологій та бізнес-процесів у великих компаніях, останнім часом, активно виникає питання не тільки про оптимізацію пошуку у базах даних, але і про проблеми інтегрування даних з однієї інформаційної системи в іншу. Через те що бізнес-середовище досить агресивне сьогодні, в сучасних умовах, все частіше зустрічаються випадки, коли більш великі корпорації поглинають дрібні, або коли фінансове становище однієї компанії погіршується настільки, що вона змушена поступатися більш успішній. Виникає ситуація, коли необхідно інтегрувати дані з декількох баз і помістити в одне єдине сховище для подальшого спільного використання. А це в свою чергу ускладнює процес пошуку.

Процес пошуку має свої особливості, один з яких – це велика ресурсність, вплив великої кількості факторів. Сьогодні існує багато методів вирішення цієї проблеми, але постійний розвиток у цій галузі вимагає постійного покращення. Через це проблема пошуку та обробки інформації завжди є актуальною.

Оптимізація інформаційного пошуку з'явилася ще в період розвитку пошукових систем. Структурні відмінності інформаційних сховищ (особливо при поєднанні декількох) можуть викликати такі проблеми при інтеграції даних:

– проблема неоднорідності, коли використовуються різні моделі даних для різних джерел;

– проблема назв, коли використовується різна термінологія, що призводить до омонімії і синонімії в іменуванні;

– семантичні проблеми, коли обрані різні рівні абстракції для моделювання подібних сутностей реального світу;

– структурні проблеми, коли однакові сутності представляються в різних джерелах з несхожими структурами даних.

Ці відмінності можуть також виражатися у використанні неоднакових типів даних для відображення однакових за змістом атрибутів (номер телефону як числове поле, або як строкове). У різних базах можуть зустрічатися випадки, коли в одній базі атрибут приймає тип даних «домен», а в іншій – «створена таблиця-довідник». Можуть існувати відмінності в одиницях виміру (температура за Цельсієм, Фаренгейтом або за Кельвіном).

Важливими у базах даних є відмінності «домен – група доменів» (наприклад, в одній базі даних адреса зберігається одним рядком, а в іншій існують окремі поля для індексу, міста, вулиці, номера будинку та квартири) і «дані – схема» (наприклад, в одній базі даних «доктор наук» – це значення атрибута «вчений ступінь» відносини «викладачі», а в іншій базі даних «доктор наук» – ставлення, яке містить дані про всіх викладачів з цим вченим ступенем). Для спрощення багатьох моментів існує досить ефективний підхід на семантичному рівні. Семантичний рівень інтеграції ґрунтується на змістовній спорідненості даних, які об'єднуються. Семантична інтеграція ґрунтується на знанні і обліку природи даних. Дані повинні зберігатися разом з метаданими. Це є складніше в реалізації, але значно збільшує комфортність роботи.

Є ще одна важлива проблема – інтеграційні програми не враховують семантику даних. А отже, дані повинні містити в собі описи власної семантики. Це в свою чергу ускладнює процес проектування сховища даних.

Саме в цьому випадку і повинна відбуватися семантична інтеграція, яка дозволить об'єднувати тільки ті дані, які відповідають, або найбільш близькі до одних і тих же сутностей в певній предметній області бази даних. Необхідно бачити не окремі факти, а «всю картину в цілому». Виникає проблема в інтеграції інформації, що зараз носить назву інтеграція інформації підприємства (Enterprise Information Integration, EII).

А вже на рівні семантичної інтеграції виникають такі проблеми:

– протиріччя у визначенні концептів;

– неоднозначність або різночитання імен;

– застосування несумісних метрик;

– протиріччя у визначенні відносин між даними;

– неоднозначність інтерпретації значень.

Онтологія частково вирішує ці проблеми і дає докладну концептуалізацію, яка описує семантику даних. Її функція аналогічна схемі бази даних. Однак мова визначення онтологій синтаксично і семантично багатша, ніж поширені підходи в базах даних. Онтологія надає теоретичний опис предметної області, а не структуру сховища

даних. Оскільки онтологія використовується для спільного використання та обміну інформацією, то вона повинна розділятися і містити узгоджену термінологію.

Отже, потрібно створювати єдину онтологію. Існує два методи формування єдиної онтології:

1) Шляхом розподілу – утворюється шляхом глобального опису концептів, відносин і функцій інтеграції з розподіленими словниками, для специфікації семантики кожного з наборів даних, які підлягають інтеграції;

2) Шляхом інтеграції – передбачає формування і поповнення глобальної онтології як результатів узгодженого об'єднання словникових ресурсів локальних онтологій, сформованих для наборів, даних, що підлягають інтеграції.

Також, для вирішення проблеми необхідно створити алгоритм перевірки отриманої інформації та визначення її змісту.

Література

1. Колисниченко Д.Н. Поисковые системы и продвижение сайтов / Д.Н. Колисниченко – М.: Диалектика, 2014. – 272 с.

2. Маннинг К. Введение в информационный поиск / К. Маннинг, П. Рагхаван, Х. Шютце – М.: Вильямс, 2011. – 600 с.

3. Wache H. Ontology-Based Integration of Information – A Survey of Existing Approaches / H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hubner // Proceedings of the IJCAI01 Workshop on Ontologies and Information Sharing, Seattle. – USA, 2001. – P. 108-118.