УДК 681.3.06

МЕТОД АВТОМАТИЧНОЇ КЛАСТЕРИЗАЦІЇ ТРАНЗАКЦІЙНИХ ДАНИХ НА ОСНОВІ B+ ДЕРЕВ

Роль В.І.

Тернопільський національний економічний університет, магістрант

І. Постановка проблеми

У наш час все більша кількість компаній, прагнучи до підвищення ефективності і прибутковості бізнесу користуються цифровими (автоматизованими) способами обробки даних і запису їх в БД. Це несе в собі як величезні переваги, так і породжує певні проблеми, пов'язані з обсягом отриманих даних, а саме: при колосальному збільшенні обсягу отриманої інформації ускладнюється її обробка і аналіз, робити висновки за отриманими даними стає все складніше, і ймовірність того, що деякі деталі можуть бути втрачені невблаганно зростає. Дана проблема з'явилася причиною розвитку різних підходів і методів, що дозволяють проводити автоматичний аналіз даних. Для вирішення даних питань існують математичні методи, які і утворюють напрям Data Mining.

Термін "транзакція" відноситься до підмножини предметів із загальної сукупності з змінним числом предметів (потужністю підмножини). Транзакціями є записи в таблиці, що містять категоріальні атрибути і мають різну довжину, на відміну від категоріальних БД, де довжина всіх записів однакова. У зв'язку з цим транзакційні БД вважаються менш впорядкованими, в порівнянні з категоріальними БД, що ускладнює їх кластеризацію.

ІІ. Мета роботи

Метою дослідження є удосконалення класичного алгоритму LargeItem в частині зручності і швидкості пошуку кластерних наборів в транзакційних БД великого обсягу (порядку 100000 транзакцій і більше). Використати для цього рекурсивні структури даних у вигляді збалансованого бінарного дерева B+.

ІІІ. Особливості реалізації удосконаленого алгоритму кластеризації

Основними етапами роботи класичного алгоритму є:

- первинний прохід по базі транзакцій і їх об'єднання в кластери.
- вторинний прохід по базі кластерів і розрахунок можливого поліпшення кластеризації.

Модернізація класичного алгоритму кластеризації полягає у використанні для зберігання кластерів замість звичайних бінарних дерев збалансованих бінарних дерев (B+tree).

Дана модернізація дозволяє відмовитися від використання Hash-таблиць. Це можливо за рахунок того, що в B+ деревах немає необхідності використовувати зовнішні індексні ключі, а також навігація і пошук по дереву зроблені простіше (за рахунок можливість послідовного доступу до ключів).

Висновки

У даній роботі була порушена актуальна область методів аналізу даних, що інтенсивно розвивається. Було розглянуто новий підхід до кластеризації. Розглянуті принципи роботи алгоритмів кластеризації, вивчена структура подання інформації у вигляді таблиць транзакційних даних, розглянуті принципи складання бінарних дерев, а також B+ дерев, модернізовано існуючий алгоритм кластеризації з урахуванням застосування в ньому B+ дерев.

Список використаних джерел

1. Барсегян и др. Методы и модели анализа данных: OLAP и Data Mining. – СПб., 2004.
2. Стаття Ke Wang, ChuXu, Bing Liu_Clustering Transactions Using Large Items 2003
3. В. Ганти. Добыча данных в сверхбольших базах данных / В. Ганти, Й. Герке, Р. Рамакришнан // Открытые системы, №9-10, 1999.