Подход к интеграции интеллектуального анализа данных и реляционной СУБД на основе таблиц предвычислений

Т.В. Речкалов, М.Л. Цымблер

Южно-Уральский государственный университет (НИУ)

Интеллектуальный анализ данных (ИАД) представляет собой совокупность технологий, направленных на получение ранее неизвестных закономерностей в больших объемах данных. Интеграция алгоритмов ИАД в реляционные *системы управления базами данных (РСУБД)* представляет собой одну из актуальных задач аналитической обработки данных [1]. Интеграция позволяет избежать накладных расходов на экспорт информации из базы данных в систему интеллектуального анализа данных. В докладе будет представлен подход к интеграции ИАД и РСУБД на основе таблиц предвычислений.

Таблица предвычислений (ТПВ) представляет собой таблицу базы данных, содержащую предварительно вычисленные и повторно используемые данные для некоторого алгоритма ИАД. Столбцы ТПВ могут быть двух видов. В столбцах первого вида хранятся внешние ключи. Комбинация этих колонок представляет собой составной первичный ключ. В оставшихся столбцах второго вида хранятся предвычисленные результаты. Целью использования ТПВ является ускорение алгоритма ИАД за счет уменьшения количества вычислений. Примером возможного применения данного подхода является алгоритм кластеризации *K-Medoids* [2]. В данном алгоритме в ходе вычислений в качестве центра кластера выбирается один из объектов, подвергаемых кластеризации (называемый медоидом). В силу этого в качестве ТПВ может использоваться матрица расстояний (см. рис. 1).

	Id1	Id2	Distance
1	1	2	d_{12}
2	1	3	d_{13}
…	…	…	…
n^2-n	n	n-1	d_{nn-1}

Рис. 1. Таблица предвычислений алгоритма K-Medoids

Здесь столбец Distance предназначен для хранения предварительно вычисленных значений расстояний d_{ij} (расстояние между i-м и j-м объектами). Во время работы алгоритма выполняются SQL запросы на получение расстояний для объектов с указанными идентификаторами.

Наиболее трудоемкие вычисления алгоритм выполняет при нахождении новых центров кластеров. Каждый объект, не являющийся медоидом, необходимо рассмотреть в качестве кандидата на замену каждому медоиду. Временная сложность одной итерации алгоритма *без* учета сложности вычисления расстояния между объектами равна $O(k(n-k)^2)$ [2]. В предположении, что объект является точкой m-мерного пространства, а в качестве расстояния используется Евклидова метрика, уточненная оценка сложности итерации будет равна $O(mk(n-k)^2)$. В случае использования ТПВ сложность вычисления расстояния между двумя любыми объектами будет $O(1)$, а сложность одной итерации — $O(k(n-k)^2)$.

Применение ТПВ комбинируется с использованием многоядерных ускорителей, которые могут существенно увеличить эффективность обработки запросов [3]. На ускорителе выполняется вычисление ТПВ и ее сжатие. ТПВ хранится в сжатом виде в памяти ускорителя и алгоритм интеллектуального анализа данных обращается к ней во время вычислений.

Литература

1. Abadi D., Agrawal R., Ailamaki A., et al. The Beckman Report on Database Research // SIGMOD Record, 2014. Vol. 43, No. 3. P. 61–70.
2. Han J., Kamber K. Data Mining: Concepts and Techniques. Second edition. // Elsevier, 2006. 743 p.
3. Беседин К.Ю., Костенецкий П.С. Моделирование обработки запросов на гибридных вычислительных системах с многоядерными сопроцессорами и графическими ускорителями // Программные системы: теория и приложения. 2014. Т. 5. № 1-1 (19). С. 91-110.