УДК 004.65

СПОСОБЫ КЛАСТЕРИЗАЦИИ ДАННЫХ СРЕДСТАМИ СУБД

Наумов Р.К. **(**федеральное государственное автономное образовательное учреждение высшего образования Университет ИТМО**)**

Научный руководитель – д.т.н., доцент Басов О.О.

(федеральное государственное автономное образовательное учреждение высшего образования Университет ИТМО)

Одним из перспективных направлений развития реляционных СУБД является внедрение в них средств интеллектуального анализа данных. Одной из задач интеллектуального анализа является задача кластеризация. В работе описаны алгоритмы кластеризации данных, а также способы ее реализации внутри СУБД: кластеризация с помощью системы ATLAS, с помощью библиотеки MADlib и с помощью расширения языка SQL – добавления оператора CLUSTER BY.

Введение. На сегодняшний день, несмотря на растущую популярность NoSQL СУБД, реляционные СУБД занимают лидирующую позицию среди инструментов управления базами данных. Основным из направлений развития РСУБД является внедрение в них средств интеллектуального анализа данных. Интеллектуальный анализ данных ориентирован на извлечение знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Интеграция позволяет как избежать накладных расходов по экспорту анализируемых данных из хранилища и импорту результатов анализа обратно в хранилище, так и использовать при анализе данных системные сервисы, заложенные в архитектуре СУБД. Одной из задач анализа данных является задача кластеризации. Задача кластеризации заключается в разбиении множества объектов сходной структуры на заранее неизвестные группы (кластеры) в зависимости от похожести свойств объектов. Кластеризация применяется в широком спектре приложений: сегментирование медицинских и спутниковых изображений, анализ ДНК-микрочипов и текстов и др.

Основная часть. Основным способом интеллектуального анализа данных в СУБД, в том числе решения задачи кластеризации, является разработка библиотеки хранимых процедур. Подключая библиотеку к своему приложению базы данных, прикладной программист получает возможность выполнять интеллектуальный анализ данных, не выходя за рамки СУБД.

Кластеризация данных в СУБД Oracle может производится с помощью системы интеллектуального анализа данных ATLAS, которая поддерживает одноименный язык запросов, являющийся надстройкой над SQL. Язык ATLAS добавляет в SQL поддержку пользовательских функций и функций, возвращающих в качестве значения реляционную таблицу. На языке ATLAS реализованы алгоритм кластеризации DBSCAN, а также алгоритмы поиска шаблонов и классификации.

Одним из методов анализа данных в реляционных СУБД PostgreSQL и Greenplum является библиотека MADlib, с помощью которой можно произвести операцию кластеризации, классификации, регрессия и др. Работа с библиотекой не требует экспорта и импорта данных внешних аналитических приложений. В реализации MADlib используются пользовательские функции, написанные разработчиками на язык программирования Python, которые обеспечивают обращения к словарю базы данных и формирование корректной структуры таблиц с выходными данными для заданных таблиц с входными данными.

Также начинает набирать популярность способ кластеризации непосредственно самим языком работы с данными внутри СУБД – SQL. Так, например, для кластеризации данных предлагается расширить язык SQL оператором CLUSTER BY. Данная конструкция подразумевает выполнение группировки строк результата запроса в соответствии со специфицированным алгоритмом кластеризации, в отличие от стандартного оператора

GROUP BY, который осуществляет группировку по точному совпадению значений в полях записей. Такой подход задействует PostGIS - расширение СУБД PosgreSQL, обеспечивающее поддержку географических объектов.

Выводы. В работе предлагается производить несложные задачи анализа данных, такие как кластеризация данных, непосредственно в месте хранения данных, в СУБД. Таким образом, разработчик сможет не только избежать расходов по экспорту анализируемых данных из хранилища и импорту результатов анализа обратно в хранилище, но и воспользоваться вшитыми в архитектуру СУБД средствами. Были представлены различные способы кластеризации данных: с помощью хранимых процедур, библиотек, а также с помощью расширения (надстройки) языка SQL.