Фрактальный поиск в базе данных с применением модели распределенных вычислений*

Т.Ю. Лымарь, Т.С. Мантрова

Южно-Уральский государственный университет (НИУ)

В теории фракталов под доменом понимаются самоподобные части системы, применив к которым определенную итерационную функцию, можно получить всю систему [1]. Применительно к реляционным базам данных доменом является комбинация атрибутов одного кортежа, с помощью которой можно разложить отношение на более мелкие части, сохранив при этом сведения о структуре данного отношения.

Целью данного исследования является фрактальный анализ реляционных баз данных с использованием модели распределенных вычислений MapReduce [2].

В ходе исследования разработан алгоритм поиска количества различных значений доменов для нахождения оптимального набора доменов, описывающих всю систему. Работа данного алгоритма состоит из следующих этапов:

1. Формируется множество изначально возможных структур доменов;
2. С использованием технологии распределенных вычислений MapReduce по списку возможных структур доменов вычисляется реальное количество доменов из таблицы для каждого сочетания. Каждый домен обрабатывается Map-элементами. Пары <ключ, значение>, произведенные всеми Map-элементами, группируются по ключу и направляются в виде пары <ключ, массив значений> в адрес Reduce-элементов. На шаге Reduce получаем количество различных значений доменов (Рис. 1);
3. Выбираются те домены, минимальное число которых позволит «собрать» всю таблицу.

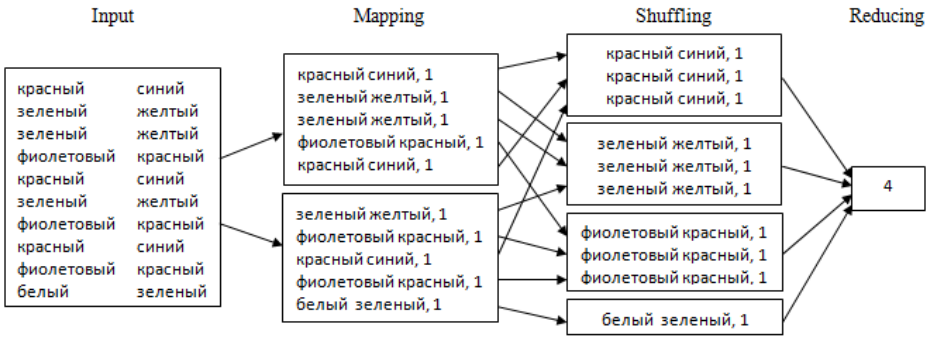

Рис. 1. Подсчет количества различных значений фрактальных доменов

Полученное количество различных значений домена отношения позволит выбрать оптимальное множество доменов, которое содержит в совокупности все атрибуты таблицы в единственном экземпляре, и сумма записей, содержащихся в доменах, будет минимальна.

Приведенный алгоритм реализован в прототипе системы, эксперименты проводятся на суперкомпьютере «СКИФ-Аврора ЮУрГУ».

Литература

1. Traina C. et al. Fast feature selection using fractal dimension. – 2000.
2. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters //Communications of the ACM. – 2008. – Т. 51. – №. 1. – С. 107-113.

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 12-07-00443-а.