Параллельный алгоритм фрактального поиска в базе данных*

Т.Ю. Лымарь, Н.Ю. Староверова

Южно-Уральский государственный университет

Основной принцип применения теории фракталов – это поиск простых частей, подобных целому, применив к которым определенную итерационную функцию, можно получить всю систему [1]. Выделение так называемых самоподобных частей (доменов) называют фрактальным анализом.

Обычным применением теории фракталов в информационных технологиях является обработка графической информации, что вполне естественно и понятно, однако можно рассмотреть с точки зрения данной теории и системы баз данных. Совокупность большого количества записей таблиц, может стать источником дополнительной, гораздо более ценной информации, которую нельзя получить на основе одной конкретной записи. Вполне возможно, что анализируемая информация может быть самоподобна и может отображать определенную зависимость не только между записями таблиц, но и внутри самой записи [2].

Целью данного исследования является рассмотрение реляционных таблиц с точки зрения теории фракталов, определение подходящих методов фрактального анализа и разработка алгоритма поиска доменов в реляционных таблицах. Возможным применением полученных результатов является поиск функциональных зависимостей в таблицах и сжатие хранимых данных.

В ходе начатого исследования разработан последовательный алгоритм поиска доменов, работа которого состоит из следующих этапов:

1. Формируется множество изначально возможных структур доменов, выбирается минимальный набор, применив к которому некоторую функцию, можно восстановить всю таблицу;
2. По списку возможных структур доменов вычисляется реальное количество доменов из таблицы для каждого сочетания;
3. Выбираются те домены, минимальное число которых позволит «собрать» всю таблицу.

Алгоритм реализован для СУБД Oracle в качестве тестовой системы. Основными источниками данных для формирования минимального множества возможных доменов являются достаточно обширный в данной СУБД словарь данных и инструмент сбора статистики. В качестве основного критерия для выделения домена использовалось количество различных значений в атрибутах, входящих в домен.

Распараллеливание реализованного алгоритма проводится на основе использования фрактальной кластеризации [3] при выявлении доменов:

1. На выделенном узле формируется начальное множество D возможных доменов;
2. Элементы множества рассылаются остальным узлам;
3. Каждый узел-получатель пытается соответственно кластеризовать свой фрагмент анализируемой таблицы, в случае неприменимости полученного домена к имеющимся записям, создается новый элемент множества D, для которого процедура рекурсивно повторяется.

Литература

1. Мандельброт Б. Фрактальная геометрия природы. М.: Институт компьютерных исследований, 2002, 656 с.
2. Traina C. Jr., Traina A., Wu L., Faloutsos C. Fast feature selection using fractal dimension. URL: http://seer.lcc.ufmg.br/index.php/jidm/article/viewFile/4/1 (дата обращения: 27.01.2010).
3. Daniel B., Ping Ch. Using the Fractal Dimension to Cluster Data sets. URL: http://cms.dt.uh.edu/Faculty/ChenP/paper/KDD00.pdf (дата обращения: 27.01.2010).

* Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект 09-07-00241-а).