# Оптимизация хранилища данных с представлением структуры в виде матроида средствами алгебры кортежей

## В. Н. Белов, П. П. Макарычев

Пензенский государственный университет

Пенза

oracool@gmail.com

Применение информационных технологий для управления вузом связано с хранением и анализом большого количества данных. Для решения таких задач целесообразно использовать хранилище данных. Однако построение хранилища данных – ресурсоемкий процесс, требующий привлечения высококвалифицированных специалистов. При проектировании хранилища данных разработчик большей частью полагается на свой опыт реализации подобных проектов и опыт известных специалистов в этой области, поскольку в настоящий момент проблема формализации процесса построения хранилища данных не решена. Большинство авторов описывают построение хранилища с одной таблицей фактов, что является наиболее простым вариантом. Следует учитывать, что наличие в хранилище данных нескольких таблиц фактов дополнительно усложняет выбор структуры хранилища данных. Зачастую ошибки проектирования обнаруживаются только на этапе разработки или тестирования, однако часть ошибок можно обнаружить и исправить еще на этапе проектирования. В данной работе решается задача оптимизации хранилища с использованием положений теории матроидов.

Предположим, что имеется некоторое хранилище данных с несколькими таблицами фактов, имеющее схему «снежинка». С помощью алгебры кортежей зададим C-систему , описывающую структуру хранилища данных [1]. В схеме отношения  значение первого атрибута соответствует таблице, внешний ключ которой ссылается на данные таблицы, соответствующие значению второго атрибута.

Построим транзитивное замыкание . C-система  избыточна, поскольку отношения, входящие в хранилище данных, структура которого соответствует , имеют все возможные связи, что приводит к росту требуемого объема памяти. Однако  позволяет определить структуры данных, подлежащих оптимизации, и выбрать наилучший вариант схемы хранилища данных. Такие структуры представляют собой C-системы, при этом множество таблиц хранилища данных задают домены атрибутов C-системы, а множество связей между таблицами хранилища данных задает множество элементарных кортежей.

Проведем предварительное исследование C-системы  для получения необходимых начальных данных. Определим множество таблиц, используемых как таблицы фактов, . Критерием принадлежности таблицы хранилища данных к множеству  является необходимость разбиения данных, содержащихся в ней, на основании признака, описываемого другой таблицей, при оперативном или интеллектуальном анализе данных. Также необходимо задать множество всех таблиц хранилища данных .

Наличие некоторых связей в хранилище данных обязательно. Они определяются согласно следующим критериям:

. (1)

При выполнении условий (1) элементарный кортеж  соответствует связи, которая не может быть удалена из структуры хранилища данных, поскольку в таком случае возможности оперативной обработки данных будут ограниченными, так как не будет возможности разбиения объектов, описываемых , по признаку, описываемому . Такие элементарные кортежи добавляются в C-систему .

Все структуры, подлежащие оптимизации, включают в себя одну или несколько таблиц фактов, и, соответственно, задают один или несколько матроидов.

Следующие критерии определяют вид C-системы , задающей матроид , где  и [2, с. 71]:

.

Матроид формируется в соответствии со следующими правилами:

 ранг матроида , задаваемого некоторой C-системой , равняется , где  – множество, являющееся подмножеством объединения доменов атрибутов отношения , в котором элементы соответствуют таблицам, входящим в структуру, описываемую C-системой ;

 множество  является множеством элементарных кортежей, входящих в C-систему ;

 множество независимых множеств  включает в себя пустое множество, максимальные независимые множества и их подмножества;

 максимальное независимое множество  удовлетворяет следующим условиям:

,  
где  – транзитивное замыкание .

Представление структуры хранилища данных в виде матроида позволяет решать множество задач с помощью жадного алгоритма. Для решения задачи с помощью жадного алгоритма необходимо задать вес для каждого элемента множества . В зависимости от требований, предъявляемых к хранилищу данных, и, соответственно, способа расчета веса элементов можно добиться различных вариантов построения хранилища данных, оптимальных для решения планируемых задач. При наличии нескольких критериев выбора структуры хранилища данных рационально использовать метод анализа иерархий для определения веса элемента носителя матроида.

После завершения работы жадного алгоритма все элементарные кортежи, вошедшие в решение, добавляются в C-систему . Это необходимо для того, чтобы при обработке других матроидов жадным алгоритмом учитывались достигнутые результаты.

Таким образом, алгоритм оптимизации структуры хранилища данных с помощью матроидов содержит следующие шаги:

1. Задается C-система, соответствующая схеме хранилища данных.
2. Строится транзитивное замыкание хранилища данных.
3. Определяется множество таблиц, используемых как таблицы фактов, и множество всех таблиц хранилища данных.
4. Задается C-система , описывающая связи, наличие которых обязательно.
5. Определяются все структуры, подлежащие оптимизации на основании соответствия выбранным критериям.
6. С помощью жадного алгоритма во всех матроидах в порядке уменьшения их ранга находится базис, сумма весов элементов которого максимальна.
7. Строится хранилище данных, схема которого задается C-системой .

В работе представлен вариант решения задачи оптимизации хранилища данных, позволяющий обнаруживать ряд ошибок, таких как высокая степень нормализации таблиц, допущенных при проектировании хранилища данных до начала этапов реализации и тестирования, что может значительно сократить объем ресурсов, требуемых для разработки хранилища данных. Предложенный алгоритм также способен оптимизировать схему хранилища данных в зависимости от выдвигаемых к нему требований. Поскольку все необходимые для работы алгоритма структуры данных и вычисления на каждом шаге алгоритма подробно описаны, возможна реализация программного средства, которое позволит решать задачу оптимизации структуры хранилища данных при минимальном участии пользователя, что позволит дополнительно сократить затраты времени на решение задачи.

Библиографический список

1. Кулик Б. А. Обобщенный подход к моделированию и анализу интеллектуальных систем на основе алгебры кортежей // Труды VI Международной конференции «Идентификация систем и задачи управления», Москва, 29 января – 1 февраля 2007 г. М.: ИПУ им. В. А. Трапезникова РАН, 2007. С. 679–15.
2. Новиков Ф. А. Дискретная математика для программистов. СПб.: Питер, 2000.