УДК 519.237.8

ПОСТРОЕНИЕ ИЕРАРХИЙ В СИСТЕМЕ ОПЕРАТИВНОГО АНАЛИЗА ДАННЫХ

А. В. Меликов

Пензенский государственный университет

Поступила в редакцию 02.03.2012 г.

Аннотация. В данной работе рассматривается возможность автоматизированного определения иерархий и их способ построения в многомерной модели данных, которая сформирована из исходной реляционной базы данных информационной системы анкетирования. Построение иерархий осуществляется на основе зависимостей атрибутов исходной базы данных. Предложенный способ определения и построения иерархий позволяет сократить время формирования схемы новой многомерной модели данных, необходимой для создания многомерного хранилища данных.

Ключевые слова: информационная система анкетирования, реляционная база данных, статистическая обработка данных, многомерная модель данных.

Annotation. In this investigation they give consideration to computerized determination of hierarchies and the way of their construction in the multidimensional data model, that is shaped from the input relational database of information questionnaire system. Hierarchic construction is based on the dependence of input database attributes. The proposed method of definition and construction of hierarchies reduces the formation time of the new scheme of multidimensional data model, which required for creation of multidimensional data warehouse.

Keywords: information questionnaire system, relational database, statistical data processing, multidimensional data model.

ВВЕДЕНИЕ

Технология комплексного многомерного анализа данных получила название OLAP (Online Analytical Processing). Системы аналитической обработки в реальном времени предоставляют конкретному пользователю возможности для реализации многомерного анализа данных и принятия управленческих решений. Принять любое решение невозможно, не обладая достаточной для этого информацией, обычно количественной. Для этого необходимо создание хранилищ данных (Data Warehouse). OLAP – это ключевой компонент организации хранилищ данных [1].

В основе OLAP лежит понятие «гиперкуба» или «многомерного куба данных», в ячейках которого хранятся анализируемые данные. Ячейка (cell) – атомарная структура куба, соответствующая полному набору конкретны значений измерений (dimensions). Другими словами, для представления данных в системах OLAP используются многомерные модели данных – гиперкубы, являющиеся обобщением электронных таблиц на произвольное количество измерений. Иногда вместо термина «ячейка» используется термин «показатель», «мера» (measure). Показатель – это поле (обычно цифровое), значения которого однозначно определяются фиксированным набором измерений. Измерение – это множество однотипных данных, образующих одну из граней гиперкуба. Измерение включает в себя уровни измерения, которые позволяют конкретному пользователю анализировать меры с различной степенью детализации. Из уровней измерения в результате может формироваться иерархия. Иерархия – группировка объектов одного измерения в объекты более высокого уровня. Иерархии в измерениях необходимы для возможности агрегации и детализации значений показателей согласно их иерархической структуре. Наличие иерархий позволяет осуществлять выполнение следующих базовых операций, используемых для анализа данных:

- поворот;
- проекция. При проекции значения в ячейках, лежащих на оси проекции, суммируются по некоторому предопределенному закону;

• раскрытие (drill-down). Одно из значений измерения заменяется совокупностью значений из следующего уровня иерархии измерения;

соответственно заменяются значения в ячейках гиперкуба;

• свертка (roll-up/drill-up). Операция, обратная раскрытию;

• сечение (slice-and-dice) [2].

В измерениях многомерных моделей существуют различные варианты задания иерархий. Следует отметить, что иерархии могут быть сбалансированными (balanced), как например: иерархии, основанные на данных типа «дата/время», и несбалансированными (unbalanced). Типичный пример несбалансированной иерархии – иерархия типа «начальник-подчиненный». Существуют также иерархии, занимающие промежуточное положение между сбалансированными и несбалансированными (они обозначаются термином ragged – «неровный»). Обычно они содержат такие члены, логические «родители» которых находятся не на непосредственно вышестоящем уровне. Расположение одного из уровней иерархии выше другого интуитивно объясняется тем, что значения более высокого уровня логически содержат значения более низкого уровня [3]. Фактом является то, что определённые иерархии задаются разработчиком. Тем самым формируются схемы многомерной модели данных.

Для реляционных баз данных, которые используются как исходные данные для гиперкубов, определены функциональные и многозначные зависимости. Данные зависимости можно использовать при создании иерархий в том случае, если в качестве уровней измерений многомерной модели используются атрибуты исходной базы данных. Возможность автоматизированного определения иерархий и их способ построения рассмотрены в данной работе на примере модели данных «композиционная таблица» (composite table), которая формируется из исходной реляционной базы данных.

1. МОДЕЛЬ ДАННЫХ «КОМПОЗИЦИОННАЯ ТАБЛИЦА»

Допустим, что X, Y_i, Z_i являются множеством атрибутов из исходного реляционного отношения R (i = 1,2,...,N). Обозначим R^* – результирующее отношение. Атрибуты X остаются неизменными в R^*, значения атрибутов Y_i становятся именами атрибутов в R^*, домены атрибутов Z_i распределяются между доменами атрибутов, введённых для Y_i. Стоит отметить, что атрибуты X и Y_i в явном виде отсутствуют в R^* (i=1,2,...,N). Но поскольку X и Y_i являются координатами для Z_i в R^*, то понятными станут следующие ограничения: $X \cap Y_i = \varnothing$, $X \cap Z_i = \varnothing$, $Y_i \cap Z_i = \varnothing$ (i = 1,2,...,N). Пусть W_i – дополнительное множество атрибутов, используемых в логических формулах-ограничениях, но отсутствующие в R^*. Тогда $W_i \in R \setminus (X \cup Y_1 \cup \cup ... \cup Y_N \cup Z_1 \cup ... \cup Z_N)$, $Dom(Y_i) = Dom(Y_{i1}) \times Dom(Y_{i2}) \times ..., Y_{ij} \in Y_i$, где Dom – множество допустимых значений атрибутов.

Схема результирующего представления строится из R по следующему правилу:

$$Sch(R^*) = \{X, \cup_{i=1}^{N} Dom(Y_i) \times \{Z_i\}\},$$

где Sch – схема описания отношения, символ $\cup$ обозначает, что «композиционная таблица» состоит из подтаблиц со схемами $\{X, Dom(Y_i) \times \times \{Z_i\}\}$ (i = 1,2,...,N) [4].

Пример 1. В качестве примера «композиционной таблицы» рассмотрим результаты информационной системы анкетирования (внутри вузовский опрос).

Общая статистика представляется по результатам теста. В таблице 1 показан фрагмент результатов анкетирования (общая статистика).

Таблица 1

Результаты анкетирования (общая статистика)

	Факультет_1			
	Кафедра_1		Кафедра_2	
	Группа_1	Группа_2	Группа_3	Группа_4
Вопрос_1	15	10	17	11
Вопрос_2	15	10	17	11
Вопрос_3	15	10	17	11

В данном случае атрибутами множества X являются вопросы, Y_1 – кафедры, Y_2 – группы, Y_3 – факультеты, Z_1 – количество ответивших.

Детальная статистика представляется по результатам вопросов в тесте. В таблице 2 показан фрагмент результатов анкетирования (детальная статистика).

Таблица 2

Результаты анкетирования (детальная статистика)

		Факультет_1			
		Кафедра_1		Кафедра_2	
		Группа_1	Группа_2	Группа_3	Группа_4
Вопрос_1	Вариант ответа_1	1	2	1	1
	Вариант ответа _2	2	1	6	3
	Вариант ответа _3	10	5	7	7
	Вариант ответа _4	2	2	3	0

В данном случае атрибутами множества X являются варианты ответов, Y_1 – кафедры, Y_2 – группы, Y_3 – факультеты, Z_1 – количество ответивших.

Для предложенной модели множества атрибутов X и Y_j (j = 1,2,...,N) могут рассматриваться в качестве измерений, так как они, по сути, являются обобщёнными координатами. Иерархии атрибутов в вышеописанных множествах определяют порядок расположения значений атрибутов в заголовках строк и столбцов таблицы. Из этого следует, что иерархии должны определяться так, чтобы пользовательское представление было наглядным.

2. СПОСОБ ПОСТРОЕНИЯ СХЕМЫ ИЕРАРХИЙ

Понятным является тот факт, что в качестве уровней измерения используются атрибуты исходной базы данных. Обозначим через L – множество атрибутов X или Y_j «композиционной таблицы» с вышеуказанной схемой описания отношения, причём $|\mathrm{Dom}(Y_i)| = L$.

Используя определение «схема иерархии» – это ориентированный ациклический и слабо связанный граф H = (A, E), где A – множество атрибутов, E – множество дуг [5], предполагается, что C, D – атрибуты, а H – схема иерархии. Следовательно, C D, если в H существует путь из вершины C в D. Принимая во внимание известное эвристическое правило: *атрибуты из множества атрибутов, принимающего меньшее количество значений, располагаются в иерархии выше, чем атрибуты из множества, принимающего большее количество значений,* и возможно определить способ задания частичного порядка на множестве атрибутов, входящих в функциональные и многозначные зависимости. Для функциональной зависимости C D, атрибуты из множества D располагаются в иерархии выше, чем атрибуты из множества C [так как различные значения C могут определять одинаковое значение D]. Таким образом, для атрибутов $C_k \in C$, $D_l \in D$ $\forall$ k, l C_k D_l. Для многозначной зависимости C D(E), где атрибуты из C располагаются в иерархии выше, чем атрибуты из D $\cup$ E [так как при существовании двух кортежей, совпадающих по C, существует ещё два кортежа с тем же значением C (по определению многозначной зависимости)]. Таким образом, для атрибутов $C_k \in C$, $I_l \in D \in E \in k$, l I_l C_k, где I – атрибут, подобный C.

Следует отметить, что существуют некоторые последовательности уровней, которые могут многократно использоваться в иерархиях измерений гиперкубов. Связь между такими атрибутами установить не всегда получается посредством зависимостей, которые задаются для исходной базы данных. Отсюда следует, что для этих атрибутов задание отношения на множестве атрибутов необходимо предоставить пользователю. Иерархии, которые были заданы пользователем, будут использоваться при формировании схемы иерархии.

Пример 2. Рассмотрим следующую схему БД:

R_1 = список групп (№ группы, № кафедры, название группы);

R_2 = список кафедр (№ кафедры, № факультета, название кафедры);

R_3 = список факультетов (№ факультета, название факультета);

R_4 = тесты (№ теста, название теста);

R_5 = вопросы (№ вопроса, № теста, текст вопроса);

R_6 = ответы (№ вопроса, № варианта ответа, ответ);

R_7 = анкетирование/назначение анкеты группе (№ группы, № теста).

Задавая частичный порядок на множестве атрибутов для функциональной зависимости, получаем, что:

№ группы → № кафедры, название группы, в результате:

№ группы ≺ № кафедры, № группы ≺ название группы.

Задавая частичный порядок на множестве атрибутов для многозначной зависимости, получаем, что:

№ группы → № теста, № вопроса, № варианта ответа, в результате:

№ теста ≺ № группы, № вопроса ≺ № группы, № варианта ответа ≺ № группы.

На рисунке 1 изображён пример пользовательской иерархии атрибутов, актуальной для данной схемы БД.

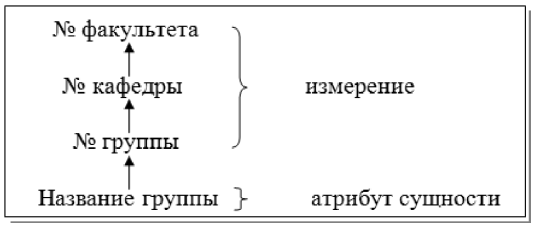

Рис. 1. Пользовательская иерархия атрибутов

Для дальнейшего изложения необходимо применить следующее определение: число дуг, которые имеют вершину x_i своей начальной вершиной, называется полустепенью исхода вершины x_i, аналогично, число дуг, которые имеют вершину x_i своей конечной вершиной, называется полустепенью захода вершины x_i [6].

В примере 3 приведена схема иерархий в многомерной модели данных информационной системы анкетирования, отвечающая примеру «композиционной таблицы», которая реализована посредством применения вышеописанного определения из теории графов.

Пример 3. Необходимо такое множество атрибутов L, по которому будет сформирована корректная схема иерархии. Пусть L = {№ группы, № теста, № варианта ответа, ответ, № кафедры, № факультета, название факультета}. По множеству L следует, что схема иерархии выглядит следующим образом (см. рис. 2):

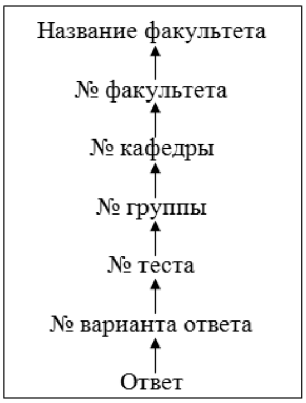

Рис. 2. Схема иерархии

ЗАКЛЮЧЕНИЕ

В данной работе были описаны возможность автоматизированного определения иерархий и их способ построения, который формирует иерархии в измерениях гиперкуба, при этом используются различные зависимости исходной базы данных с иерархиями атрибутов, которые изначально определены конкретным пользователем. Данный способ получил применение в системе многомерного анализа данных, полученных из системы анкетирования университета. С его помощью осуществляется обработка данных непосредственно при импортировании их из первоначальной системы сбора данных. Преимуществом данного подхода, реализуемого в системах OLAP, является сокращение интервала времени, необходимого для формирования схемы новой многомерной модели данных. Кроме того, иерархии в измерениях на примере описанной выше модели данных необходимы как для реализации операций анализа данных, так и для структурирования заголовков пользовательского представления. Формирование иерархий осуществляется таким образом, чтобы представление самой модели в виде двумерной таблицы было удобным для работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Туманов В. Е.* Проектирование хранилищ данных для приложений систем деловой осведомлённости (Business Intelligence Systems). / В. Е. Туманов, – М.: Интернет-университет информационных технологий – ИНТУИТ.ру, 2004.

2. *Елманова Н. З., Фёдоров, А. Г.* Введение в OLAP-технологии Microsoft. / Н. З. Елманова, А. Г. Фёдоров, – М.: Диалог-МИФИ, 2002. – С. 13–15.

3. *Бергер А. Б., Горбач И. В., Меломед Э. Л., Щербинин В. А., Степаненко В. П.* Microsoft SQL Server 2005. Analysis Services. OLAP и многомерный анализ данных. / А. Б. Бергер, И. В. Горбач, Э. Л. Меломед, В. А. Щербинин, В. П. Степаненко, – СПб.: БХВ-Петербург, 2007. – С. 128–135.

4. *Мельников Б. С.* Поисковое проектирование: Учебное пособие. / Б. С. Мельников, – М.: МЭИ, 2005. – С. 7–11.

5. *Мельников О. И.* Занимательные задачи по теории графов: Учеб.-метод. пособие. Изд-е 2-е. / О. И. Мельников, – Мн.: ТетраСистемс, 2001. – С. 34–41.

6. *Зыкин С. В.* Формирование гиперкубического представления реляционной базы данных. // С. В. Зыкин. Программирование, изд-е №6, – М.: Изд-во Наука, 2006. – С. 348–354.

Меликов А. В. – вед. программист УСК ФГБОУ ВПО «ПГУ», Тел.: +7 (937) 404-5858, e-mail: AlexeyV.Melikov@yandex.ru

Meliko Alexey V. – chief programmer of informational support division for quality management system, Quality system management department, Penza state university. Tel.: +7 (937) 404-5858, e-mail: AlexeyV.Melikov@yandex.ru