ИНФОРМАТИКА

НАУЧНЫЙ ОТДЕЛ

УДК 519.72

ПОСТРОЕНИЕ ИЕРАРХИЙ В МНОГОМЕРНЫХ МОДЕЛЯХ ДАННЫХ

П.Г. Редреев

Омский филиал Института математики СО РАН,
лаборатория методов преобразования и представления информации
E-mail: redreev@mail.ru

В работе рассматривается способ автоматизированного определения иерархий в измерениях многомерного представления данных, сформированного из исходной реляционной базы данных. Построение иерархий осуществляется на основе зависимостей атрибутов исходной базы данных.

Ключевые слова: реляционная база данных, аналитическая обработка данных, многомерная модель данных.

Construction of Hierarchies in Multidimensional Data Models

P.G. Redreev

Omsk Branch Institute of Mathematics SB RAS,
Laboratory of Methods of Transformation and Representation of Information
E-mail: redreev@mail.ru

This article considers a method of automated determination of hierarchies in dimensions of multidimensional data model, constructed from the source relational database. Construction of hierarchies is realized on the basis of attribute dependencies of source database.

Key words: relational database, online analytical processing, multidimensional data model.

ВВЕДЕНИЕ

Системы оперативного анализа данных OLAP (Online Analytical Processing) предоставляют пользователю возможности для многомерного анализа данных и принятия решений. Для представления данных в OLAP-системах используются многомерные модели данных, являющиеся гиперкубами, то есть обобщением электронных таблиц на произвольное количество измерений (dimensions). В многомерных моделях данные рассматриваются либо как меры (measures), которые являются числовыми значениями, либо как текстовые измерения [1]. Меры — это величины, подвергаемые анализу по измерениям. Измерение включает в себя уровни измерения, позволяющие пользователю анализировать меры с различной степенью детализации. Из уровней измерения может формироваться иерархия. Наличие иерархий позволяет осуществлять выполнение таких часто используемых для анализа данных операций, как roll-up и drill-down [2, 3].

Существуют различные способы задания иерархий в измерениях многомерных моделей [1]. Расположение одного из уровней иерархии выше другого интуитивно объясняется тем, что значения более высокого уровня логически содержат значения более низкого уровня [1, 4]. Как правило, иерархии задаются разработчиком при формировании схемы многомерной модели данных.

Для реляционных баз данных, используемых в качестве исходных данных для гиперкубов, заданы функциональные и многозначные

зависимости. Эти зависимости могут быть использованы при создании иерархий, если в качестве уровней измерений многомерной модели используются атрибуты исходной базы данных. В данной работе рассмотрен способ автоматизированного определения иерархий для модели данных «композиционная таблица», формируемой из исходной реляционной базы данных.

1. МОДЕЛЬ ДАННЫХ «КОМПОЗИЦИОННАЯ ТАБЛИЦА»

Рассмотрим модель данных «композиционная таблица», которая является обобщением модели «семантическая трансформация» [5], на случай списка значений в одной ячейке. Обозначим R — исходное реляционное отношение, R^* — результирующее отношение. Пусть X, Y_i, Z_i — множества атрибутов из R $(i = 1, 2, \ldots, N)$. Атрибуты X остаются неизменными в R^*, значения атрибутов Y_i становятся именами атрибутов в R^*, домены атрибутов Z_i, дополненные пустым значением, распределяются между доменами атрибутов, введенных для Y_i. Атрибуты Y_i и Z_i в явном виде отсутствуют в R^* $(i = 1, 2, \ldots, N)$. W_i — дополнительное множество атрибутов, которые используются в логических формулах-ограничениях, но в R^* отсутствуют. Поскольку X и Y_i являются координатами для Z_i в R^*, естественными являются ограничения: $X \cap Y_i = \varnothing, X \cap Z_i = \varnothing, Y_i \cap Z_i = \varnothing$ $(i = 1, 2, \ldots, N)$. $W_i \in R \backslash (X \cup Y_1 \cup \ldots \cup Y_N \cup Z_1 \cup \ldots \cup Z_N)$, $|Dom(Y_i)| = L_i, |Z_i| = M_i$, $|\cdot|$ — мощность множества.

Схема результирующего представления строится из R по следующему правилу:

$$Sch(R) = \{X, Y_1, \ldots, Y_N, Z_1, \ldots, Z_N, W_1, \ldots, W_N\} \;\Rightarrow\; Sch(R^*) = \{X, \bigcup_{i=1}^{N} Dom(Y_i) \times \{Z_i\}\},$$

где Sch — схема описания отношения, Dom — множество допустимых значений атрибутов, $Dom(Y_i) = Dom(Y_{i1}) \times Dom(Y_{i2}) \times \ldots$, $Y_{ij} \in Y_i$. Символ $\cup$ обозначает, что «композиционная таблица» состоит из подтаблиц со схемами $\{X, Dom(Y_i) \times \{Z_i\}\}$ $(i = 1, 2, \ldots, N)$.

Пример 1. В качестве примера «композиционной таблицы» рассмотрим план учебной нагрузки вуза (рис. 1).

	Семестр					
	1		**2**			
	Лекции	**Практика**	**Лекции**	**Практика**	**Экзамен**	**Зачет**
Наименование дисциплины	*Кол-во часов*	*Кол-во часов*	*Кол-во часов*	*Кол-во часов*	*№ семестра*	*№ семестра*
Иностранный язык		3		2	4	1,2,3
Физическая культура		2		2		1,2,3,4
Отечественная история	2	2			1,2	
Философия	2	2			1,2,3	
Экономика			2	2	2	

Рис. 1. План учебной нагрузки

Здесь атрибуты множества X: *название дисциплины*, Y_1: *номер семестра, тип занятия*, Z_1: *количество часов данного вида занятий по дисциплине в неделю*, Y_2: *общесеместровые мероприятия по дисциплине*, Z_2: *номер семестра*.

Для модели «композиционная таблица» множества атрибутов X и Y_j $(j = 1, 2, \ldots, N)$ являются обобщенными координатами и могут рассматриваться как измерения. Иерархии атрибутов в X и Y_j $(j = 1, 2, \ldots, N)$ определяют порядок расположения значений атрибутов в заголовках строк и столбцов пользовательского представления в виде двумерной таблицы. Следовательно, иерархии должны быть определены таким образом, чтобы пользовательское представление было наглядным. Наиболее удобной для работы пользователя является древовидная структура заголовка таблицы, пример которой показан на рис. 2.

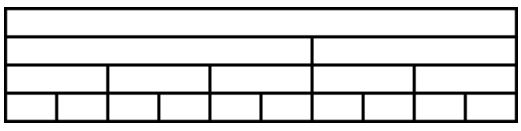

Рис. 2. Структура заголовка таблицы

2. ФОРМИРОВАНИЕ СХЕМЫ ИЕРАРХИИ

Рассмотрим способ определения иерархии в измерении. В качестве уровней измерения будем использовать атрибуты исходной базы данных. Пусть L — множество атрибутов X или Y_j «композиционной таблицы» со схемой $Sch(R^*) = \{X, \bigcup_{i=1}^{N} Dom(Y_i) \times \{Z_i\}\}$ $(i = 1, 2, \ldots, N)$.

Определение 1. *Схема иерархии* — это ориентированный ациклический и слабо связный граф $H = (A, E)$, где A — множество атрибутов, E — множество дуг.

Определение 2. Пусть C, D — атрибуты, H — схема иерархии. $C \prec D$, если в H существует путь из вершины C в D.

Определим способы задания частичного порядка на множестве атрибутов.

Для задания частичного порядка на множестве атрибутов, входящих в функциональные и многозначные зависимости, используем следующее эвристическое правило.

Атрибуты из множества атрибутов, принимающего меньшее количество значений, располагаются в иерархии выше, чем атрибуты из множества, принимающего большее количество значений.

Для функциональной зависимости $C \to D$, где C и D — множества атрибутов, атрибуты из D располагаются в иерархии выше, чем атрибуты из C, так как различные значения множества атрибутов C могут определять одинаковое значение D. Таким образом, будем полагать, что для атрибутов $C_k \in C, D_l \in D$ $\forall k, l$ $C_k \prec D_l$.

Для многозначной зависимости $C \twoheadrightarrow D(E)$, где C, D, E — множества атрибутов, атрибуты из C располагаются в иерархии выше, чем атрибуты из $D \cup E$, так как по определению многозначной зависимости при существовании двух кортежей, совпадающих по C, существуют еще два кортежа с тем же значением C. Таким образом, будем полагать, что для атрибутов $C_k \in C, I_l \in D \cup E$ $\forall k, l$ $I_l \prec C_k$.

Некоторые последовательности уровней могут многократно использоваться в иерархиях измерений различных гиперкубов или задаваться в заголовках пользовательских представлений данных. Связь между этими атрибутами не всегда возможно установить с помощью зависимостей, заданных для исходной базы данных. Следовательно, для данных атрибутов задание отношения $\prec$ на множестве атрибутов необходимо предоставить пользователю. Заданные пользовательские иерархии будем использовать при формировании схемы иерархии.

Пример 2. Рассмотрим следующую схему БД:

R_1 =Студенты (№ студента, № группы, ФИО студента);
R_2 =Список групп (№ группы, Код группы, № специальности);
R_3 =Предметы (№ предмета, Предмет);
R_4 =Преподаватели (№ преподавателя, ФИО преподавателя);
R_5 =Неделя (№ дня недели, День недели);
R_6 =Начало занятий (№ пары, Время начала пары);
R_7 =Оценки (№ студента, № группы, № предмета, Оценка);
R_8 =Расписание (№ группы, № дня недели, № пары, № предмета, № преподавателя, № аудитории);
R_9 =Специальности (№ специальности, Специальность);
R_{10} =Нагрузка (№ предмета, № специальности, Количество часов).

При задании частичного порядка на множестве атрибутов для функциональной зависимости *№ группы → Код группы, № специальности* получаем *№ группы ≺ Код группы* и *№ группы ≺ № специальности*, для многозначной зависимости *№ группы ↠ № дня недели,№ пары* (*№ студента*) — *№ дня недели ≺ № группы*, *№ пары ≺ № группы*, *№ студента ≺ № группы*.

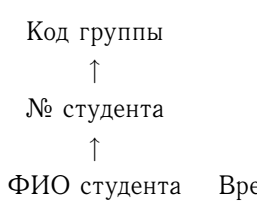

Рис. 3. Пользовательские иерархии атрибутов

Примеры пользовательских иерархий атрибутов для данной схемы БД изображены на рис. 3.

Приведем определение из теории графов [6], используемое в дальнейшем изложении.

Определение 3. Число дуг, которые имеют вершину x_i своей начальной вершиной, называется полустепенью исхода вершины x_i, и аналогично число дуг, которые имеют x_i своей конечной вершиной, называется полустепенью захода вершины x_i.

Рассмотрим *алгоритм построения схемы иерархии* H.

Шаг 1. Для каждой пользовательской иерархии $B_1 \prec \ldots \prec B_m$ добавляем в H дуги $B_i B_{i+1}$, где $B_i, B_{i+1} \in L$, $i = 1, \ldots, m-1$.

Шаг 2. Для каждой функциональной зависимости $C \to D$, где C, D — множества атрибутов, такой, что $C' \neq \emptyset$, $D' \neq \emptyset$, где $C' = L \cap C$, $D' = L \cap D$, добавляем в H дугу $C'_k D'_l$, $C'_k \in C', D'_l \in D'$ $\forall k, l$, если на шаге 1 не добавлена дуга $D'_l C'_k$.

Шаг 3. Для каждой многозначной зависимости $C \twoheadrightarrow D(E)$, где C, D, E — множества атрибутов, такой, что $C' \neq \emptyset$, $D' \neq \emptyset$, где $C' = L \cap C$, $D' = L \cap (D \cup E)$, добавляем в H дугу $D'_l C'_k$, $C'_k \in C', D'_l \in D'$ $\forall k, l$, если на шаге 1 или на шаге 2 не добавлена дуга $C'_k D'_l$.

Шаг 4. Пока в графе H содержатся циклы, выполняем следующее. Определяем количество различных значений для каждого атрибута, соответствующего каждой вершине цикла. В цикле находим вершины A_i, соответствующие атрибутам, принимающим минимальное количество значений. Выбираем из вершин смежных из вершин A_i вершины B_j такие, что соответствующие атрибуты принимают максимальное количество значений. Удаляем из цикла одну из выбранных дуг $A_i B_j$.

Шаг 5. Дополняем в граф H вершины для атрибутов из L, отсутствующих в H в качестве вершин. Если граф H является несвязным, выполняем следующее.

Для каждой компоненты связности графа находим вершины с полустепенью исхода, равной 0. Для этих атрибутов определяем минимальную величину m_k количества различных значений.

Упорядочиваем компоненты связности графа по возрастанию m_k. Для компонент связности, у которых m_k одинакова, находим вершины компоненты с полустепенью захода, равной 0. Для этих атрибутов определяем максимальную величину n_l количества различных значений.

Упорядочиваем эти компоненты по возрастанию n_l.

Последовательно просматриваем компоненты связности графа H в соответствии с полученным упорядочением. Дополняем в граф дуги, идущие в каждую вершину с полустепенью захода, равной 0, текущей компоненты связности из каждой вершины с полустепенью исхода, равной 0, следующей компоненты.

Пример 3. Пусть L={*№ группы*, *№ студента*, *ФИО студента*, *№ специальности*, *Специальность*}. По множеству атрибутов L будет сформирована схема иерархии, изображенная на рис. 4.

ЗАКЛЮЧЕНИЕ

Рассмотренный в данной работе алгоритм формирует иерархии в измерениях гиперкуба, используя функциональные, многозначные зависимости исходной базы данных и иерархии атрибутов, заданные пользователем. Применение этого алгоритма в OLAP-системах позволит сократить время на формирование схемы новой многомерной модели данных.

Для рассмотренной модели данных «композиционная таблица» иерархии в измерениях необходимы не только для реализации операций анализа данных, но и для структурирования заголовков пользовательского представления. Предложенный алгоритм формирует иерархии таким образом, чтобы представление модели в виде двумерной таблицы было наиболее удобным для работы пользователя.

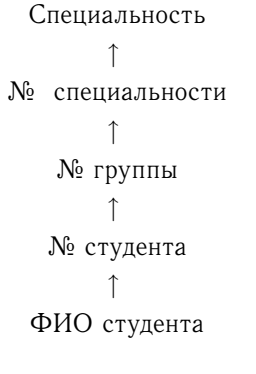

Рис. 4. Схема иерархии

Работа выполнена при финансовой поддержке РФФИ (*проект* 09-07-00059-а).

Библиографический список

1. *Pedersen T.B., Jensen C.S., Dyreson C.E.* A foundation for capturing and querying complex multidimensional data // Information Systems. 2001. № 26(5). P. 383–423.

2. *Педерсен Т.Б., Йенсен К.С.* Технология многомерных баз данных // Открытые системы. 2002. № 1. С. 45–50.

3. *Щавелев Л.В.* Способы аналитической обработки данных для поддержки принятия решений // СУБД. 1998. № 4–5. С. 51–60.

4. *Lechtenborger J., Vossen G.* Multidimensional normal forms for data warehouse design // Information Systems. 2003. № 28(5). P. 415–434.

5. *Кристофидес Н.* Теория графов. М.: Мир, 1978. 215 с.

6. *Зыкин С.В.* Формирование гиперкубического представления реляционной базы данных // Программирование. 2006. № 6. С. 348–354.