Формирование Представления Данных со Списочными Компонентами для Работы с Реляционными Базами Данных по Технологии OLAP

П.Г. Редреев

Омский филиал института математики СО РАН, ул. Певцова, 13, Омск, 644099, Россия

redreev@mail.ru

Аннотация. *В работе рассматривается технология автоматизации построения гиперкубического представления данных, содержащего списочные компоненты в ячейках, из исходной реляционной базы данных. Построенное сложноструктурированное табличное представление данных реализует технологию OLAP. При построении формально определены промежуточная и целевая модели данных. Определено условие существования представления целевой модели данных.*

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

1 Введение

Технология OLAP (online analytical processing) [1,2] при работе с базами данных предоставляет наиболее удобный способ представления информации для её анализа. Работа пользователей с данными осуществляется с помощью выполнения операций над гиперкубом: slice-and-dice, drill-down, roll-up и др.

В данной работе предполагается, что основой аналитической работы пользователя является необходимость формирования новых гиперкубов, а не многократное формирование реализации одного и того же гиперкуба. Таким образом, следует уделить внимание сокращению времени формирования схемы нового гиперкуба, а формирование представления гиперкуба должно быть выполнено автоматически.

Автоматизация формирования представления гиперкуба основывается на использовании формального определения промежуточной и целевой моделей данных, задающих схемы представлений и способы их формирования.

В работе [3] при формировании гиперкуба используется следующая последовательность преобразований: $\boldsymbol{RRD} \Rightarrow \boldsymbol{TJ} \Rightarrow \boldsymbol{ST}$, где ***RRD*** – реляционное представление данных, ***TJ*** – таблица соединений, ***ST*** – гиперкуб «семантическая трансформация». Здесь ***RRD*** – представление исходной модели данных, ***ST*** – целевой, ***TJ*** является промежуточным представлением.

В качестве целевой модели данных рассмотрим «композиционную таблицу», являющуюся обобщением модели гиперкуба «семантическая трансформация» [3] на случай списка значений в одной ячейке, разделенных знаками препинания, то есть списочных компонентов. Для модели «семантическая трансформация» требуется выполнение условия существования, гарантирующего, что в одну ячейку гиперкуба будет помещено не более одного значения. Так как для представления «композиционной таблицы» это ограничение отсутствует, то необходимо сформулировать другое условие существования. В качестве промежуточной модели будем использовать «таблицу связанных соединений», являющуюся частным случаем модели «таблица соединений».

Примером «композиционной таблицы» является план учебной нагрузки Вуза (рис. 1).

Код	Наименование дисциплины	Трудоемкость										Число недель в семестре			
												17		18	
												Семестр			
												1	1	2	2
		Всего	лекции	лаб	практ	семинар	сам раб	Экзамен	Зачет	курс. пр	курс. раб	Л	Пр	Л	Пр
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ГСЭ	**Общегуманитарные и социально-экономические дисциплины**														
ГСЭ.Ф.00	Федеральный компонент	**1260**													
ГСЭ.Ф.01	Иностранный язык	340	0		176		164	4	1,2,3				3		2
ГСЭ.Ф.02	Физическая культура	408	0		140		268		1,2,3,4				2		2
ГСЭ.Ф.03	Отечественная история	108	34		34		40	1				2	2		
ГСЭ.Ф.06	Правоведение	70	18		36		16	4							
ГСЭ.Ф.07	Психология и педагогика	70	18		18		34	3							
ГСЭ.Ф.10	Философия	108	34		34		40	1				2	2		
ГСЭ.Ф.11	Экономика	156	36		36		84	2			2			2	2
ГСЭ.Р.00	Национально-региональный(кузовский) компонент	**270**													
ГСЭ.Р.01	Социология	90	18		18		54		3						
ГСЭ.Р.02	Политология	90	34		18		38		3						
ГСЭ.Р.03	Деловой иностранный язык	90	0		36		54		5,6						
ГСЭ.В.00	Дисциплины по выбору студентов	**270**													
ГСЭ.В.01	Русский язык и культура речи	70	18		18		34		2					1	1
ГСЭ.В.02	Культурология														
ГСЭ.В.03	Социология маркетинговых исследований	70	18		0		52		3						
ГСЭ.В.04	История Омского региона														
ГСЭ.В.05	Профессиональный ин-ый язык	130	0		64		66	8	7						
	Всего ГСЭ	**1800**													

Рис.1. Традиционная форма представления учебного плана.

2 Модель данных «Таблица связанных соединений»

В качестве промежуточной модели при формировании «композиционной таблицы» предлагается использовать «таблицу связанных соединений». Совокупность свойств этой таблицы является достаточной для формирования большинства приложений.

Рассмотрим преобразование представления реляционной БД со схемой: R_1, R_2, ..., R_k в «таблицу связанных соединений» (C,l), где C – схема отношения, определенная на множестве атрибутов $A_1, A_2, ..., A_n$. Определим принцип формирования кортежей $t \in c$, где c – реализация схемы отношения C. Рассмотрим соединения, удовлетворяющие локальному свойству соединения без потерь информации (ССБПИ) [4], для всевозможных сочетаний из k по m реализаций r_j, $m=1,...,k$. Для каждого кортежа u, в котором атрибуты, являющиеся координатами ячейки «композиционной таблицы», имеют определенные значения, каждого соединения сформируем кортеж t по следующим правилам: $t[A_j]=u[A_j]$, если атрибут A_j принадлежит соединению, и $t[A_j]=\mathbf{emp}$ в противном случае. Каждому кортежу поставим в соответствие битовый вектор $l(t) = (l_1(t), l_2(t), ..., l_k(t))$, где $l_j(t)=1$, если реализация r_j схемы R_j участвует в текущем соединении, и $l_j(t)=0$ в противном случае.

Рассмотрим отношение частичного порядка над кортежами $t \in c$. Кортеж $t \in c$ является менее определенным или равным кортежу $t' \in c$, когда для любого атрибута A_i выполнено: если $t[A_i] \neq t'[A_i]$, то $t[A_i]=\mathbf{emp}$ и $l_j(t') \geq l_j(t)$, $j=1,\ldots,k$. В этом случае будем писать: $t \preccurlyeq t'$ и назовем кортеж t подчиненным кортежу t'. В представлении c достаточно хранить только кортеж t', который содержит в себе все менее определенные либо равные кортежи.

Пусть $X(J)=(R_{j(1)} \cup R_{j(2)} \cup ... \cup R_{j(m)})$, где $J=(j(1),j(2),\ldots,j(m))$. Определим операцию проекции на множестве c: $\pi_{X(J)}(c)$ есть совокупность кортежей $u[X(J)]$, определенных на множестве атрибутов $X(J)$, где для каждого $u[X(J)]$ существует кортеж $t \in c$ такой, что $u[X(J)]=t[X(J)]$ и $l_{j(i)}(t)=1$, $i=1,2,\ldots,m$.

Основываясь на способе формирования C-таблицы, сформулируем важные свойства.

Свойство 1. Для любого связанного соединения отношений $R_{j(1)}, R_{j(2)}, ..., R_{j(m)}$, где $m=1,2,\ldots,k$, выполнено: $R_{j(1)} \bowtie R_{j(2)} \bowtie ... \bowtie R_{j(m)} = \pi_{X(J)}(c)$, где $\bowtie$ – операция естественного соединения.

Это основное свойство «таблицы связанных соединений» показывает, что она содержит все связанные соединения отношений БД, в том числе и исходные отношения. Это позволяет считать «таблицу связанных соединений» обобщением универсального реляционного отношения [4]. Кроме того, операция проекции задает обратное преобразование из «таблицы связанных соединений» в БД.

Свойство 2. Реализация ***C***-таблицы ***c*** всегда существует и единственна для любой схемы реляционной БД.

Единственность «таблицы связанных соединений» доказывается аналогично единственности «таблицы соединений» [5].

3 Модель данных «Композиционная таблица»

Рассмотрим построение представления «композиционной таблицы». Обозначим R_1, R_2, ...,R_k – исходные реляционные отношения, C – соответствующая этим отношениям «таблица связанных соединений», R^* – результирующее отношение.

Пусть X, Y_i, Z_i – множества атрибутов из $R=(R_1, R_2, ...,R_k)$ $(i=1,2,\ldots, N)$. Атрибуты X остаются неизменными в R^* и являются наименованиями строк, значения атрибутов Y_i становятся именами столбцов в R^*, домены атрибутов Z_i, дополненные пустым значением, распределяются между доменами новых атрибутов, введенных для значений Y_i. W_i – дополнительное множество атрибутов, которые используются в логических формулах-ограничениях, но в R^* отсутствуют. Естественными являются ограничения: $X\cap Y_i=\varnothing$, $X\cap Z_i=\varnothing$, $Y_i\cap Z_i=\varnothing$ $(i = 1, 2, \ldots, N)$. $W_i\in R\backslash(X\cup Y_1\cup\ldots \cup Y_N\cup Z_1\cup\ldots \cup Z_N)$, $|Dom(Y_i)|=L_i$, $|Z_i|=M_i$.

Схема результирующего представления строится по следующему правилу:

$$Sch(C)=\{X, Y_1, \ldots, Y_N, Z_1, \ldots, Z_N, W_1, \ldots, W_N\} \Rightarrow Sch(R^*)=\{X, \cup Dom(Y_i)\times\{Z_i\}\ (i=1, 2, \ldots, N)\},$$

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 $\{X,Dom(Y_i)\times\{Z_i\}\}$ $(i=1,2,\ldots,N)$. Представление r^* со схемой $Sch(R^*)$ является плоской таблицей. Оно может быть представлено в виде двумерной таблицы (рис. 2).

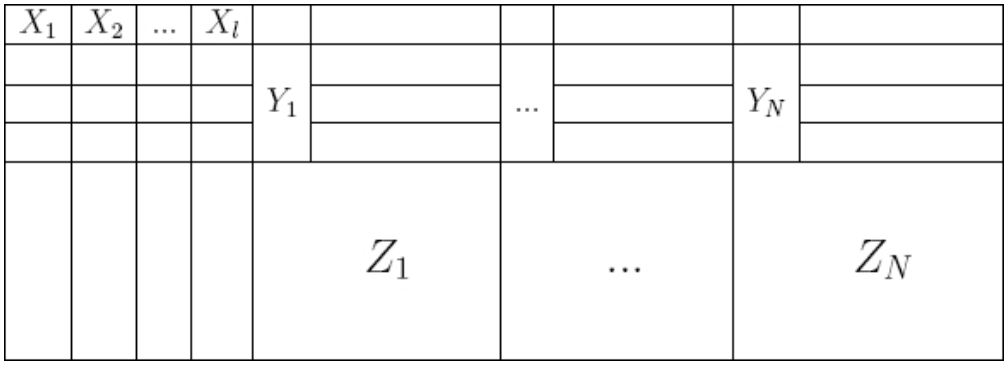

Рис.2. Общий вид представления «Композиционная таблица».

4 Условие существования представления *r**

Определение. Представление r^* сформировано корректно, если:

1. В ячейках r^* содержатся однородные значения.
2. В каждой строке $r_1^*\in r^*$ с определенными значениями атрибутов из X, в ячейке $r_1^*[\vec{y}_j.Z_{jp}]$, где $\vec{y}_j$ – определенные значения, $j=\overline{1, N}$, $p=\overline{1, L_j}$

а) содержатся все значения, соответствующие наборам $r_1^*[X]$, $\vec{y}_j$,

б) отсутствуют значения, для которых строка $(r_1^*[X],\vec{y}_j,z^1_{jp})$, $z^1_{jp}\in r_1^*[\vec{y}_j.Z_{jp}]$ не может быть получена при проекции связанного соединения некоторых отношений из набора R_1, R_2, ...,R_M по атрибутам X, Y_j, Z_{jp}.

Теорема. Представление r^* всегда корректно и единственно для совокупности R_1, R_2, ...,R_M, образующих связанные соединения.

Выполнение данного свойства для совокупности отношений гарантирует, что в ячейке «композиционной таблицы» не появится значения противоречащего прикладной области.

5 Заключение

«Композиционная таблица» является обобщением гиперкуба «семантическая трансформация» [3] на случай нескольких значений в одной ячейке, что позволяет реализовывать более широкий класс пользовательских приложений. Согласно условию существования «композиционная таблица» может быть построена для любой совокупности отношений, удовлетворяющих локальному ССБПИ.

На основе рассмотренного преобразования создан инструментарий для построения пользовательского представления, формирующий приложения без привлечения языка программирования.

Литература

[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] Lechtenborger J., Vossen G., Multidimensional normal forms for data warehouse design// Information Systems. – 2003 – Volume 28, Issue 5 – P. 415-434.
[3] Зыкин С.В. Формирование гиперкубического представления реляционной базы данных// Программирование. – 2006. – № 6. – С. 348-354.
[4] Ульман Дж. Основы систем баз данных// М. : Финансы и статистика, 1983. – 334 с.
[5] Зыкин С.В. Построение отображения реляционной базы данных в списковую модель данных// Управляющие машины и системы. – 2001. – №3. – С. 42-63.