Вестник УГАТУ
Уфа : УГАТУ, 2012 Т. 16, № 3 (48). С. 180–186

УДК 004.65

В. В. Миронов, Е. С. Макарова

АГРЕГАЦИЯ ПОКАЗАТЕЛЕЙ В OLAP-КУБЕ ПРИ СВЕДЕНИИ ПО ЗАВИСИМЫМ ИЗМЕРЕНИЯМ

Обсуждаются особенности агрегации данных в многомерном кубе при наличии измерений, между которыми имеются функциональные зависимости. Рассматриваются основные понятия и графическое представление многомерной модели данных. Обсуждается процесс формирования сводных ячеек. Выявляется эффект фильтрации идентифицирующих координат при подсчете сводных показателей, возникающий, когда привязанные к идентифицирующим неидентифицирующие факт-координаты не попадают в множество укрупнения. Приводятся примеры, иллюстрирующие процесс сведения показателей по зависимым измерениям. *OLAP-система*; *многомерная модель данных*; *гиперкуб*; *измерение*; *неаддитивная мера*; *показатель*; *агрегация*

В настоящее время аналитическая обработка данных (On-Line Analytical Processing – OLAP) [1] получила широкое распространение как средство анализа и информационной поддержки принятия управленческих решений. Аналитические модули появляются в составе самых разных приложений, предоставляя пользователям удобный интерфейс для организации данных в виде *гиперкубов* на основе многомерной модели (Multidimensional Model) и для оперативного анализа различных показателей с помощью OLAP-операций агрегации, детализации, вращения, среза и т. п.

Для адресации показателей в гиперкубе предусмотрены *измерения*, содержащие множества координат, которые могут быть организованы в виде иерархий. Во многих работах, например, [3, 8], на измерения накладываются ограничения взаимной независимости по аналогии с евклидовой системой координат. Однако многомерное моделирование многих проблемно-предметных областей требует использования измерений, находящиеся в функциональной зависимости друг от друга. Системы управления базами данных (СУБД), поддерживающие многомерную модель, вполне допускают наличие функционально зависимых измерений [4, 7]. Поэтому необходимо провести анализ особенностей агрегации многомерных данных для зависимых измерений.

Контактная информация: 8(347)273-78-23

Работа выполнена в рамках научной школы УГАТУ «Теория и практика разработки информационных систем» и поддержана грантом РФФИ 10-07-00167-а «Электронные документы со встроенной динамической моделью», а также грантом Президента Российской Федерации НШ-65497.2010.9 для ведущих научных школ.

Ключевую роль в OLAP-системах играют операции *агрегирования*, или *сведения*, т. е. процедуры автоматического формирования меньшего количества результирующих значений (агрегатов) из большего количества исходных значений. Стандартные агрегации, основанные на суммировании, подсчете количества значений, вычислении минимального и максимального значений показателей, достаточно просты для теоретического анализа благодаря свойству аддитивности сводного показателя (меры). Однако во многих случаях требуется нестандартная, неаддитивная агрегация. Современные OLAP-системы предусматривают возможность как аддитивного, так и неаддитивного агрегирования [5], что обуславливает необходимость исследований в общем виде.

В данной работе рассматривается агрегирование по функционально зависимым измерениям для общего случая неаддитивных мер.

ОСНОВНЫЕ ПОНЯТИЯ И ИСПОЛЬЗУЕМАЯ НОТАЦИЯ

В настоящее время отсутствует единство терминологического описания многомерных моделей данных, а также не существует и универсальной графической нотации для визуального представления OLAP-кубов. В данной статье используется графическая нотация для описания многомерных моделей данных, базирующаяся на работе [9].

Измерения (Dimensions). В общем случае куб данных имеет несколько измерений, задающих систему координат пространства данных. На рис. 1 показано графическое изображение куба с тремя осями – измерениями – *X*, *Y* и *Z*. Каждое измерение представляет собой конечное множество элементов (Members) или *координат* – значений данных, образующих

грань куба. Некоторый элемент измерения определяет координату по этому измерению. Совокупность таких координат по всем измерениям образует кортеж, который идентифицирует ячейку (Cell) куба. Ячейка – это часть данных, получаемая путем определения одного элемента в каждом измерении многомерного массива. На рис. 1 показана ячейка C_{ijk} с координатами (X_i, Y_j, Z_k), то есть соответствующая элементу i измерения X, элементу j измерения Y и элементу k измерения Z.

Меры (Measures). В каждой ячейке размещены в общем случае несколько мер – показателей, хранящихся в кубе [6]. На рис. 1 с ячейкой ассоциировано p показателей: $(f_1, f_2, \ldots, f_p)$. Обычно все меры одного q-го типа в кубе однородны, т. е. представляют собой различные значения некоторого показателя. Как правило, в роли показателей выступают некоторые числовые значения. Совокупность ячеек куба, соответствующих зафиксированным координатам из разных измерений, образуют срезы (Slices) куба по этим координатам. Результатом среза по координатам всех измерений куба будет отдельная ячейка.

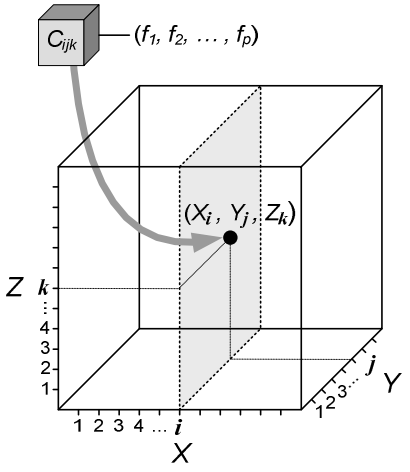

Рис. 1. Гиперкуб данных

Иерархия (Hierarchy) – способ введения в измерении укрупненных координат посредством группирования. В каждом измерении гиперкуба задается множество так называемых ключевых (Key Members) или факт-координат (Fact Members), т. е. однородных элементов измерения с наибольшей степенью детализации (гранулярности). Во множестве факт-координат может быть задано разбиение на подмножества, каждому подмножеству соответствует укрупненная координата измерения, совокупность которых составляет вышестоящий уровень иерархии. Множество укрупненных координат входит в состав измерения наряду с факт-координатами. В свою очередь, оно может быть разбито на подмножества, которым соответствуют укрупненные координаты следующего уровня, и т. д. На самом верхнем уровне иерархии вводится единственная корневая координата «Все» («All»), символизирующая совокупность всех элементов нижестоящего уровня. В общем случае для одного измерения может быть задано несколько иерархий, основанных на общем множестве факт-координат.

Графическая нотация для представления многомерной модели поясняется на рис. 2.

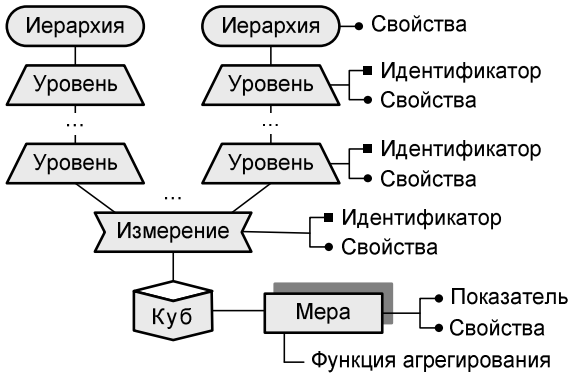

Рис. 2. Условные обозначения элементов OLAP-куба

Гиперкуб данных в целом обозначается символом куба. Выносными линиями с ним связаны принадлежащие ему измерения, которые обозначаются выпукло-вогнутыми шестиугольниками, а также меры, обозначаемые прямоугольниками. Иерархии измерений обозначаются овалами, а уровни иерархий – трапециями. Нижние уровни иерархий соединяются с символами своих измерений, верхние – с символами своих иерархий, а промежуточные – с вышестоящими и нижестоящими уровнями. Символы измерений означают множества факт-координат, символы уровней – укрупненных координат, а символы иерархий – корневые координаты типа «Все». С помощью выносных линий с темными квадратиками к измерениям и уровням прикрепляются имена атрибутов, являющихся идентификаторами координат, а линий с кружками – имена атрибутов-свойств координат. Аналогичным образом к мерам прикрепляются имена атрибутов-показателей фактов и атрибутов-свойств мер. Кроме того, к мерам прикрепляются имена агрегатных функций, используемых для получения сводных показателей для укрупненных координат.

ФОРМИРОВАНИЕ СВОДНЫХ ЯЧЕЕК

Полноправность укрупненных координат как членов измерения – важная особенность многомерной модели, означающая, что эти элементы могут использоваться в качестве координат измерений для адресации ячеек гиперкуба точно так же, как обычные факт-координаты. Укрупненным координатам в гиперкубе соответствуют так называемые агрегированные или сводные ячейки (Aggregate / Pivot Cells), в отличие от ячеек фактов (Fact Cells), у которых все координаты являются факт-координатами.

Значения, содержащиеся в сводных ячейках, соответствующих укрупненным координатам некоторого уровня иерархии, должны формироваться на основе значений ячеек, соответствующих координатам нижестоящих уровней иерархии. Рассмотрим некоторое измерение X гиперкуба (рис. 3).

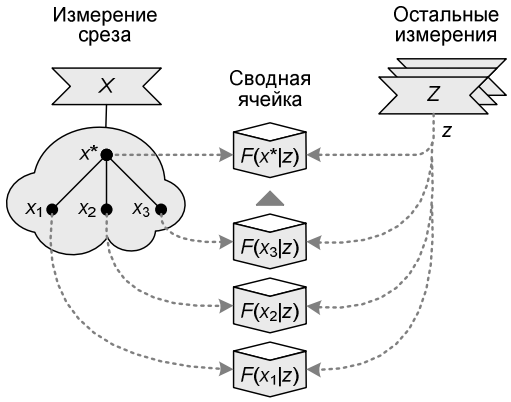

Рис. 3. Схема формирования сводных ячеек

Каждой укрупненной координате x^* некоторого уровня иерархии измерения X соответствует множество $\Omega(x^*) = \{x_1, x_2, x_3, \ldots\}$ дочерних координат нижестоящего уровня этой иерархии. Пусть $Z = \{Z_1, Z_{2,\ldots}\}$ – множество остальных измерений гиперкуба, а z – кортеж текущих координат этих измерений, то есть $z = \langle z_1, z_2, \ldots\rangle$, $z_1 \in Z_1$, $z_2 \in Z_2$, … . Для заданного кортежа z координатам $x_1, x_2, x_3, \ldots$ соответствуют ячейки с координатами $\langle x_1, z\rangle$, $\langle x_2, z\rangle$, $\langle x_3, z\rangle$, …, а укрупненной координате x^* – сводная ячейка с координатами $\langle x^*, z\rangle$. Указанные ячейки содержат значения некоторого показателя F:

$$\langle x_1, z\rangle \to F(x_1|z), \langle x_2, z\rangle \to F(x_2|z),$$
$$\langle x_3, z\rangle \to F(x_3|z), \ldots, \langle x^*, z\rangle \to F(x^*|z).$$

Тогда должна иметь место зависимость: $F(x^*|z) = \text{Aggr}\{F(x_1|z), F(x_2|z), \ldots\} = \text{Aggr}\, F(y|z)$, $y \in \Omega(x^*)$, где Aggr – некоторая функция агрегации (Aggregate Function), задающая правила вычисления агрегированных показателей сводных ячеек на основе исходных показателей укрупняемых ячеек. Суммирование часто применяется в качестве функции агрегации, в этом случае сводные ячейки содержат сумму показателей своих нижестоящих в иерархии ячеек, т. е. $F(x^*|z) = F(x_1|z) + F(x_2|z) + \ldots = \Sigma\, F(y|z)$, $y \in \Omega(x^*)$.

Сведение показателя – это вычисление сводного показателя (меры) для некоторой сводной ячейки путем агрегирования показателей ячеек нижестоящих уровней иерархии. Множество соответствующих локально дочерних ячеек формируется следующим образом. Берутся укрупненные координаты сводной ячейки по всем измерениям. Для каждой координаты через ее иерархию определяется множество соответствующих ей локально дочерних координат нижестоящего уровня иерархии (при этом если координата уже принадлежит нижнему уровню иерархии, т. е. является факт-координатой, то она и используется в качестве локально дочерней). Формируется искомое множество кортежей, адресующих локально дочерние ячейки, в виде декартова произведения множеств локально дочерних координат по всем измерениям.

Рекурсивный характер локального сведения показателей выражается в том, что для вычисления сводного показателя некоторой ячейки необходимо предварительно вычислить показатели для локально дочерних ячеек, для вычисления которых, в свою очередь, необходимо вычислить показатели для их локально дочерних ячеек и т. д. до тех пор, пока не будут достигнуты факт-ячейки, показатели которых известны априори. Таким образом, задача последовательно сводится к серии аналогичных задач (рекурсия) для нижестоящих уровней вплоть до самых нижних в иерархиях измерений. Формально данную процедуру можно описать следующей рекурсивной формулой:

$$F_{\text{лок}}(\vec{t}) = \begin{cases} f(\vec{t}),\ \vec{t} - \text{кортеж факт-координат}, \\ \underset{\vec{t}^* \in \Omega^*(\vec{t})}{Aggr}\ F(\vec{t}^*) - \text{в противном случае}, \end{cases} \qquad (1)$$

где $\vec{t}^*$ – кортеж локально дочерних координат ячейки с координатами $\vec{t}$; $\Omega^*(\vec{t})$ – множество всех кортежей $\vec{t}^*$ ячейки с координатами $\vec{t}$.

СВЕДЕНИЕ ПО ИЗМЕРЕНИЯМ

Сведение по измерениям – процедура, предполагающая определенную последовательность

агрегирования локально дочерних показателей. А именно, сначала показатели агрегируются по координатам одного измерения, затем полученный агрегат, не зависящий от этого измерения, агрегируется по координатам другого измерения, далее – по координатам третьего измерения, и т. д. до последнего измерения, при агрегации по координатам которого получается искомый сводный показатель. Если требуемое агрегирование допускает сведение по измерениям, то это создает удобства на практике.

Формально эту процедуру можно представить следующим образом. Пусть необходимо выполнить локальное сведение показателей для ячейки с координатами $\vec{t}$. В соответствии с формулой (1) для этого необходимо выполнить агрегацию локально дочерних ячеек $\mathrm{Aggr}F(\vec{t}^{\,*})$ при $\vec{t}^{\,*} \in \Omega^*(\vec{t})$. Пусть гиперкуб имеет N измерений и множество измерений упорядочено. Запишем

$$\vec{t}^{\,*} = (t_1^{\,*}, t_2^{\,*}, ..., t_N^{\,*});$$
$$\Omega^*(\vec{t}) = \Omega_1^{\,*}(\vec{t}) \times \Omega_2^{\,*}(\vec{t}) \times ... \times \Omega_N^{\,*}(\vec{t}),$$

где $\Omega_i^{\,*}(\vec{t})$ – множество локально дочерних координат ячейки $\vec{t}$ по измерению i, $i = 1, 2, \ldots, N$, а «×» – символ операции декартова произведения множеств. Тогда сведение по измерению означает, что

$$\mathrm{Aggr}\, F(\vec{t}\,^*) =$$
$$= \mathrm{Aggr}^{(1)}\, \mathrm{Aggr}^{(2)} \ldots \mathrm{Aggr}^{(N)}\, F(t_1^{\,*}, t_2^{\,*}, ..., t_N^{\,*}).$$

Здесь $\mathrm{Aggr}^{(i)}$, $i = 1, \ldots, N$ – оператор агрегирования, применяемый при сведении по i-му измерению. Индекс (i) подчеркивает, что в самом общем случае неравномерной функции агрегирования этот оператор может изменяться от измерения к измерению; для равномерных агрегатных функций оператор агрегирования одинаков для всех измерений. Как видно, сначала сводится измерение N, затем $(N - 1)$ и т. д.

Рекуррентные соотношения формирования агрегированного значения при сведении по измерениям можно записать следующим образом. Пусть $\varphi_i(\vec{t}^{\,*}_{N-i})$ – промежуточный результат, когда сведение выполнено по i последним измерениям, $\vec{t}^{\,*}_{N-i} = (t_1^{\,*}, t_2^{\,*}, \ldots, t_{N-1}^{\,*})$ – список еще не сведенных измерений. Представим $\varphi_i(\vec{t}^{\,*}_{N-i})$ в виде $\varphi_i(t^*_{N-i} | \vec{t}^{\,*}_{N-i-1})$, где вертикальная черта («при условии») отделяет координаты измерения, по которому выполняется свертка на следующем шаге, от тех, которые не затрагиваются при агрегировании. Тогда

$$\begin{aligned}
&\varphi_0(\vec{t}^{\,*}_N) = F(\vec{t}^{\,*}_N) = F(\vec{t}^{\,*}),\\
&\varphi_1(\vec{t}^{\,*}_{N-1}) = \mathrm{Aggr}^{(N)}\, \varphi_0(t^*_N | \vec{t}^{\,*}_{N-1}),\\
&\varphi_2(\vec{t}^{\,*}_{N-2}) = \mathrm{Aggr}^{(N-1)}\, \varphi_1(t^*_{N-1} | \vec{t}^{\,*}_{N-2}),\\
&\varphi_i(\vec{t}^{\,*}_{N-i}) = \mathrm{Aggr}^{(N-i+1)}\, \varphi_{i-1}(t^*_{N-i+1} | \vec{t}^{\,*}_{N-i}),\\
&\ldots\\
&\varphi_N() = \mathrm{Aggr}^{(1)}\, \varphi_{N-1}(t_1^*) = \mathrm{Aggr}\, F(\vec{t}^{\,*}).
\end{aligned} \qquad (2)$$

Последнее выражение дает окончательный результат сведения по всем измерениям.

СВЕДЕНИЕ ПО НЕИДЕНТИФИЦИРУЮЩИМ ИЗМЕРЕНИЯМ

Сведение показателей по некоторым измерениям может иметь особенности, а именно, некоторые измерения могут не влиять на значение сводного показателя, и поэтому могут не учитываться в процессе агрегации. Этот эффект особенно проявляется при сведении корневых ячеек. Корневой по некоторому измерению называется сводная ячейка, которая по этому измерению имеет координату «Все». Сведение ячейки, корневой по некоторому измерению, в случае плотной иерархии означает, что требуется рассчитать показатель, соответствующий всей совокупности факт-координат этого измерения.

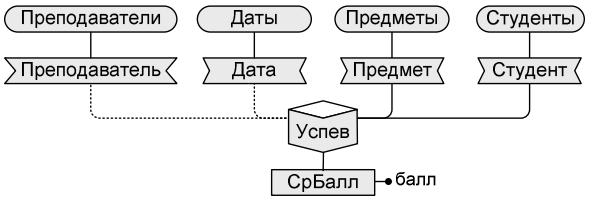

Рис. 4. К сведению по неидентифицирующим измерениям

На рис. 4 приведена простая модель гиперкуба, предназначенного для анализа успеваемости. Мера «Средний балл», основанная на факт-показателе «балл», базируется на 4 измерениях: «Студент», «Предмет», «Преподаватель», «Дата» (для простоты измерения даны без промежуточных уровней иерархий). Рассмотрим два случая сведения корневых ячеек:

1. Пусть необходимо вычислить сводный показатель, являющийся корневым по измерению «Студент» и (или) «Предмет», например, средний балл, соответствующий оценкам, проставленным определенным преподавателем

в определенный день, усредненный по всем студентам и всем предметам. Один преподаватель в один день может принять экзамены у нескольких студентов, поэтому для получения сводного показателя необходимо агрегировать (усреднять) показатель (балл) на множестве сдач этого рода.

2. Пусть теперь необходимо вычислить сводный показатель, являющийся корневым по другим двум измерениям: «Преподаватель» и (или) «Дата», например, средний балл, соответствующий оценкам определенного студента по определенному предмету, усредненный по всем преподавателям и датам. Хотя запрос сформулирован аналогично предыдущему, ситуация здесь иная. Конкретный студент по конкретному предмету может иметь только одну оценку, поставленную одним из преподавателей в один из учебных дней, поэтому множество сдач этого рода состоит из одного элемента, т. е. агрегация по измерениям «Преподаватель» и «Дата» фактически не требуется. При агрегировании по этим измерениям будет учтен единственный преподаватель и единственная дата (кем и когда принят экзамен) и результат будет тот же самый, как если бы этих измерений не было вовсе.

ФУНКЦИОНАЛЬНО ЗАВИСИМЫЕ ИЗМЕРЕНИЯ

В теории реляционных баз данных понятие функциональной зависимости используется для определения взаимосвязей между атрибутами отношения [2]. Формально функциональную зависимость множества B от множества A представляется в виде $A \rightarrow B$; это означает, что каждому элементу множества B соответствует один элемент множества A, а каждому элементу множества A – ноль, один или несколько элементов множества B. Множество A называется детерминантом, а множество B – зависимым.

Измерение называется идентифицирующим для некоторой меры, если идентификатор факт-координат этого измерения входит в состав детерминанта функциональной зависимости факт-показателя этой меры, и неидентифицирующим в противном случае. Совокупность идентификаторов идентифицирующих измерений является идентификатором для показателя меры; показатель не зависит функционально от своих неидентифицирующих измерений. Вместе с тем, каждой факт-ячейке соответствуют единственные факт-координаты неидентифицирующих измерений, поэтому факт-координаты неидентифицирующих измерений функционально зависят от совокупности факт-координат идентифицирующих измерений как от детерминанта.

Формально функциональные зависимости между измерениями можно выразить следующим образом. Пусть $\vec{t}$ – кортеж факт-координат. Представим его как $\vec{t} = \langle \vec{p}, \vec{q} \rangle$ – где $\vec{p}$ – факт-координаты идентифицирующих измерений, а $\vec{q}$ – неидентифицирующих. Факт-показатели функционально зависят только от идентифицирующих измерений, поэтому $\vec{p}$ задает единственную факт-ячейку, и, следовательно, единственную соответствующую $\vec{p}$ совокупность $\vec{q}$, т. е. существует функция G такая, то $\vec{q} = G(\vec{p})$. Тогда для произвольных факт-координат $\vec{p}$ и $\vec{q}$ имеет место следующее соотношение:

$$f(\vec{p},\vec{q}) = \begin{cases} f(\vec{p}, G(\vec{p})) = \phi(\vec{p}), \text{ если } G(\vec{p}) = \vec{q}, \\ \text{Null в противном случае.} \end{cases}$$

Условие $G(\vec{p}) = \vec{q}$, означает, что факт-координаты неидентифицирующих измерений $\vec{q}$ соответствуют факт-координатам идентифицирующих измерений $\vec{p}$ (привязаны к ним), в этом случае показатель зависит только от $\vec{p}$. Нарушение этого условия означает, что координаты неидентифицирующих измерений не соответствуют координатам идентифицирующих, т. е. задана ячейка, которая заведомо пустая, что и отражается Null-значением показателя. Функция $\phi(\vec{p}) = f(\vec{p}, G(\vec{p}))$, зависящая только от идентифицирующих факт-координат, – это функция связанных показателей, т. е. показателей, относящихся к идентифицирующим факт-координатам и привязанным к ним идентифицирующим факт-координатам. Факт-ячейки, содержащие связанные показатели, будем называть связанными ячейками.

На рис. 5 поясняется ситуация функционально зависимых измерений на примере абстрактной двухмерной модели. Модель содержит идентифицирующее измерение P с шестью факт-координатами p_1, p_2, ...,p_6 и неидентифицирующее измерение Q тоже с шестью факт-координатами q_1, q_2,...,q_6.

Таким образом, факт-пространство содержит $6 \times 6 = 36$ факт-ячеек. Факт-координаты измерения Q функционально зависят от факт-координат измерения P: $(P) \rightarrow Q$, т. е. существует функция G такая, что $q = G(p)$, $p \in P$, $q \in Q$.

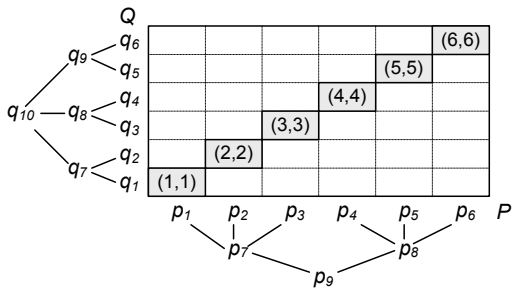

Рис. 5. К сведению по функционально зависимым измерениям

Каждой факт-координате измерения P соответствует одна факт-координата измерения Q; для примера на рис. 5 это $p_1 \to q_1$, $p_2 \to q_2$, … , $p_6 \to q_6$. Поэтому из 6 ячеек, адресуемых каждой из координат p_1, p_2, …p_6, заполненной (связанной) оказывается только одна, а остальные являются заведомо пустыми, содержащими Null-значения показателей. Функция связанных показателей имеет вид:

$$\phi(p_1) = f(p_1, q_1),$$
$$\phi(p_2) = f(p_2, q_2),$$
$$\ldots$$
$$\phi(p_6) = f(p_6, q_6).$$

СВЕДЕНИЕ ПОКАЗАТЕЛЕЙ

Наличие функциональных зависимостей между измерениями приводит к тому, что сведение показателей выполняется не во всем пространстве факт-ячеек, а лишь в подпространстве связанных факт-ячеек. Действительно, факт-ячейки, не являющиеся связанными, игнорируются при агрегировании. Пусть для подсчета сводного показателя выполняется агрегация $\mathrm{Aggr}_{\vec{q}\in\Omega_Q} f(\vec{p},\vec{q})$ по некоторому подмножеству Ω_Q факт-координат $\vec{q}$ неидентифицирующих измерений для фиксированных факт-координат $\vec{p}$ идентифицирующих измерений. Для фиксированных координат $\vec{p}$ имеются соответствующие им координаты неидентифицирующих измерений $G(\vec{p})$, которым, в свою очередь, не более одной факт-ячейки. В соответствии с формулой (2) и с учетом того, что элементы, имеющие Null-значения, игнорируются при агрегировании, получаем

$$\underset{\vec{q}\in\Omega_Q}{\mathrm{Aggr}}\, f(\vec{p},\vec{q}) = \begin{cases} f(\vec{p}, G(\vec{p})) = \phi(\vec{p}), \text{если } G(\vec{p}) \in \Omega_Q, \\ \text{Null в противном случае.} \end{cases}$$

Первый случай соответствует ситуации, в которой совокупность координат неидентифицирующих измерений $G(\vec{p})$, соответствующая совокупности фиксированных координат идентифицирующих измерений $\vec{p}$, попадает в множество Ω_Q координат свертки. В этом случае значение агрегата – это значение факт-показателя $f(\vec{p},\vec{q}) = \phi(\vec{p})$, единственного из агрегируемых, имеющего не Null-значение. Второй случай соответствует ситуации, в которой координаты $G(\vec{p})$ лежат за пределами множества Ω_Q. В этом случае все агрегируемые показатели имеют Null-значения, поэтому и результат агрегирования – тоже Null-значение.

Возвращаясь к рис. 5, обратим внимание на укрупненные координаты измерений. В измерении P имеются две координаты второго уровня иерархии: p_7 и p_8, а в измерении Q – три: q_7, q_8 и q_9. При вычислении сводного показателя, скажем, $F(p_6, q_9)$, необходимо выполнить агрегирование факт-показателей по неидентифицирующим координатам q_5 и q_6, что дает

$$F(p_6, q_9) = \mathrm{Aggr}\,\{f(p_6, q_5), f(p_6, q_6)\} = f(p_6, q_6) = \phi(p_6).$$

Здесь учтено, что координата q_5 не соответствует координате p_6, поэтому показатель $f(p_6, q_5)$ имеет Null-значение.

Иной случай имеет место при вычислении сводного показателя, скажем, $F(p_6, q_8)$. Идентифицирующей координате p_6 соответствует неидентифицирующая координата q_6, которая, в свою очередь, принадлежит укрупненной координате q_9. Поэтому при сведении по укрупненной координате p_8 результат будет иметь Null-значение.

ЭФФЕКТ ФИЛЬТРАЦИИ

Итак, результатом промежуточного сведения по неидентифицирующим измерениям является либо значение показателя для соответствующих координат идентифицирующего измерения, либо Null-значение. Поскольку Null-значения не участвуют в дальнейшем агрегировании (игнорируются), то это явление можно рассматривать как особую форму фильтрации идентифицирующих координат при подсчете сводных показателей, а именно, идентифицирующие факт-координаты, для которых привязанные неидентифицирующие факт-координаты не попадают в множество укрупнения, фильтруются при агрегировании (множество укрупнения представляет собой некоторое подмножество всех факт-координат, соответствующее выбранным укрупненным координатам). Заме-

тим, что для корневой координаты типа «Все» в плотной (неразреженной) иерархии множество укрупнения совпадает с множеством всех факт-координат измерения, поэтому для каждой совокупности идентифицирующих факт-координат найдется совокупность привязанных неидентифицирующих координат, т. е. эффект фильтрации не возникнет.

Вновь возвратимся к рис. 5. Пусть требуется сводный показатель $F(p_8, q_9)$ ячейки, адресуемой двумя укрупненными координатами p_8 и q_9. Координата p_8 имеет 3 дочерние факт-координаты p_4, p_5 и p_6, для которых, в свою очередь, имеется 3 привязанных факт-координаты q_4, q_5 и q_6. Координата q_9 имеет 2 дочерние факт-координаты q_5 и q_6. Таким образом, координата q_4, привязанная к координате p_4, не является дочерней для укрупненной координаты q_9, и в процессе агрегирования соответствующая ячейка $f(p_4, q_4)$ будет проигнорирована, т. е.

$F(p_8, q_9) = \text{Aggr}\{f(p_5, q_5), f(p_6, q_6)\} = \text{Aggr}\{\varphi(p_5), \varphi(p_6)\}$.

Это эквивалентно фильтру с условием $P \neq \neq p_4$, установленному на измерении P.

Рассмотрим теперь сведение корневого показателя, скажем, $F(p_8, q_{10})$. Координата типа «Все» q_{10} содержит в качестве глобально дочерних все факт-координаты измерения Q, поэтому все 3 связанные ячейки, адресуемые факт-координатами p_4, p_5 и p_6, будут участвовать в агрегировании, т. е.

$$F(p_8, q_{10}) = \text{Aggr}\{f(p_4, q_4), f(p_5, q_5), f(p_6, q_6)\} = \text{Aggr}\{\phi(p_4), \phi(p_5), \phi(p_6)\}.$$

Как видим, эффект фильтрации отсутствует.

ЗАКЛЮЧЕНИЕ

Итак, при выполнении операции сведения показателей по измерениям в многомерной модели данных следует учитывать функциональные зависимости между измерениями. Предложено формализованное описание операции сведения по измерениям как процедуры последовательного агрегирования локально дочерних показателей для агрегатных функций общего вида. Введены понятия идентифицирующего и неидентифицирующего измерений, а также выявлена функциональная зависимость факт-координат неидентифицирующих измерений от факт-координат идентифицирующих измерений. Предложенное формализованное описание операции сведения по измерениям распространено на случай функционально зависимых измерений. Показано, что операция сведения по функционально-зависимым измерениям является особой формой фильтрации и сопровождается игнорированием идентифицирующих факт-координат, для которых привязанные неидентифицирующие факт-координаты не попадают в множество укрупнения.

СПИСОК ЛИТЕРАТУРЫ

1. **Codd E. F.** Providing OLAP for end-user analysis: An IT mandate. ComputerWorld, 1993.

2. **Date C. J.** An Introduction to Database Systems. 2004. ссылку на первое издание!

3. Emerging cubes for trends analysis in OLAP databases / S. Nedjar [et al.] // DaWaK. Lecture Notes in Computer Science. Springer, 2007. Vol. 4654. P. 135–144. DOI: 10.1007/978-3-540-74553-2_13.

4. **Harinath S., Quinn S.** Professional SQL Server Analysis Services 2005 with MDX. N. Y.: Wiley, 2007. 848 p.

5. **Spofford G., Harinath S.** MDX Solutions: With Microsoft SQL Server Analysis Services 2005 and Hyperion Essbase. N.Y.: Wiley, 2006. 744 p.

6. Методы и модели анализа данных: OLAP и Data Mining / А. А. Барсегян [и др.]. СПб.: БХВ-Петербург, 2004. 336 с.

7. Microsoft SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных / А. Б. Бергер [и др.]. СПб.: БХВ-Петербург, 2007. 928 с.

8. **Кузнецов С. Д., Кудрявцев Ю. А.** Математическая модель OLAP-кубов // Программирование. 2009. Т. 35, № 5. С. 26–36.

9. **Миронов В. В., Юсупова Н. И.** Концептуальные модели баз данных. Многомерные модели. Уфа: УГАТУ, 2010. 83 с.

ОБ АВТОРАХ

Миронов Валерий Викторович, проф. каф. АСУ. Дипл. радиофизик (Воронежский гос. ун-т, 1975). Д-р техн. наук по управл. в техн. системах (УГАТУ, 1995). Иссл. в обл. иерархических моделей и ситуационного управления.

Макарова Екатерина Сергеевна, асп. той же кафедры. Дипл. магистр по информатике и вычислительной технике (УГАТУ, 2010). Готовит дис. в обл. OLAP-технологий.