УДК 001.57:004.658.2

ФОРМАЛЬНОЕ ПРЕДСТАВЛЕНИЕ СТРУКТУРЫ СИСТЕМ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ, ОСНОВАННЫХ НА OLAP-ТЕХНОЛОГИИ

© 2010 г. ***С.Ю. Семченков***

Рязанский государственный радиотехнический университет

amfibius@mail.ru

Поступила в редакцию 12.04.2010

Предложена математическая модель систем аналитической обработки данных на основе концепции базового многомерного куба. Сформулированы условия корректного агрегирования показателей. Рассмотрены способы устранения аномалий в иерархиях измерений.

Ключевые слова: OLAP-системы, математическая модель многомерного куба, аномалии в иерархиях.

Введение

OLAP (On-Line Analytical Processing – интерактивная аналитическая обработка данных) [1, 2] – один из способов получения и анализа данных. Суть этой технологии заключается в том, что информация, собранная для автоматизированной обработки, представляется в виде многомерного куба с возможностью произвольного манипулирования ею.

OLAP-системы являются эффективным средством для анализа и представления данных, полученных из хранилищ данных [3]. Хранилище данных (ХД) – это база данных, обладающая следующими свойствами:

1) ХД содержит ретроспективные данные.

2) ХД содержит обобщенные данные различной степени агрегации.

3) ХД содержит в явном виде информационные поля, представленные в форме измерений многомерного куба, для обеспечения многомерности представления данных.

4) Хранилища оптимизированы для выполнения запросов, содержащих объединение таблиц и агрегирование, выполняемых на больших объемах данных.

Для реализации OLAP-системы необходимо наличие отдельной многомерной СУБД, интегрирующей данные из внешних источников и обрабатывающей аналитические запросы пользователей системы.

Основными понятиями многомерной модели данных являются многомерный куб (гиперкуб), измерение, уровень измерения, показатель [4]. Особенностью измерений является их иерархическая структура, которая используется для агрегации и детализации значений показателей.

В то же время существующие математические модели многомерных систем обладают следующими недостатками, касающимися формальной структуры этих моделей.

1) Моделирование запросов к многомерной БД с помощью преобразования их в эквивалентные реляционные выражения или логические формулы делает невозможным выполнение произвольной последовательности операций.

2) Множественное наследование в иерархиях затрудняет реорганизацию иерархической структуры, так как одна и та же вершина может принадлежать различным путям агрегации, имеющим различные агрегирующие функции.

3) В процессе эволюции многомерной базы данных возникают семантические конфликты вследствие различных отношений «родитель – потомок» для нескольких вершин, являющихся родительскими для одного и того же потомка.

4) Структура измерения допускает разбиение всего множества его значений на пересекающиеся подмножества, что может привести к некорректному агрегированию.

Формальное описание гиперкуба OLAP-систем

Для описания многомерной модели, устраняющей указанные недостатки, можно предложить следующий формализм, основанный на теории бинарных отношений [5].

Пусть Θ – конечное множество всех измерений многомерного пространства для конкретной предметной области, Ψ – множество всех уровней, соответствующих измерениям множества Θ, V – множество всех значений, соответствующих измерениям множества Θ, Y – мно-

жество возможных значений всех ячеек многомерного куба. Гиперкуб – это многомерная структура, состоящая из множества ячеек и хранящая взаимосвязанные данные, описывающие предметную область. Множество-носитель всех гиперкубов Λ представим в виде декартова произведения множеств: $\Lambda = \Theta \times \Psi \times V \times Y$.

Множество Θ состоит из следующих элементов: $\Theta = \{D_1, D_2, \ldots D_i, \ldots D_q\}$, где D_i – i-е измерение, q – количество измерений. Измерение – это множество объектов одного или нескольких типов, организованных в виде иерархической структуры. Эти объекты называются значениями измерения. Графически иерархическая структура может быть представлена в виде дерева. Под уровнем измерения будем понимать множество вершин иерархической структуры, имеющих одинаковый ранг. Обозначим через DL_i^k k-й уровень i-го измерения. Множество Ψ можно представить в виде объединения $\Psi = \Psi_1 \cup \Psi_2 \cup \ldots \cup \Psi_q$, где

$\Psi_1 = \{DL_1^1, DL_1^2, \ldots\}$ – множество уровней 1-го измерения,

$\Psi_2 = \{DL_2^1, DL_2^2, \ldots\}$ – множество уровней 2-го измерения,

……..

$\Psi_q = \{DL_q^1, DL_q^2, \ldots\}$ – множество уровней q-го измерения.

$\Psi_i (i = 1 \ldots q)$ – конечные множества, которые могут иметь различное количество элементов. Множества Ψ_1, Ψ_2,…, Ψ_q должны быть упорядоченными, то есть $\langle \Psi_1, \preceq \rangle$, $\langle \Psi_2, \preceq \rangle$, …, $\langle \Psi_q, \preceq \rangle$, где $\preceq$ – отношение линейного порядка, отражающее взаимосвязь между уровнями с агрегированными и детализированными значениями. Так как каждый уровень измерения может принадлежать только одному измерению, то существует бинарное отношение $r_l \subseteq \subseteq \Pi р_{(1,2)}(\Lambda)$. Первый аргумент этого отношения – измерение D_i, второй аргумент – уровень DL_i^k, соответствующий измерению D_i. Сечение бинарного отношения r_l посредством измерения D_i состоит из множества уровней измерений, принадлежащих измерению D_i, то есть

$$\text{Сеч}_{D_i}(r_l) = (DL_i^k \in \Psi \,|\, (D, DL_i^k) \in r_l \wedge D = D_i).$$

Для каждого уровня измерения существует множество принадлежащих ему значений. Пусть V_i – множество значений всех элементов измерения D_i: $V_i = \{v_i^1, v_i^2, \ldots, v_i^m\}$, где v_i^j – j-е значение i-го измерения, m – количество элементов измерения D_i. Тогда существует бинарное отношение $r_v \subseteq \Pi р_{(1,3)}(\Lambda)$. Первый аргумент этого отношения – уровень DL_i^j измерения D_i, второй аргумент – значение измерения. Сечение бинарного отношения r_v посредством уровня DL_i^k состоит из множества значений, принадлежащих уровню DL_i^k, то есть

$$\text{Сеч}_{DL_i^k}(r_v) = (v_i^j \in V_i \,|\, (DL, v_i^j) \in r_v \wedge DL = DL_i^k).$$

Вершины измерения могут иметь предков и потомков. Пусть вершина v принадлежит конкретному уровню DL_i^k измерения D_i, тогда существуют тернарные отношения $r_a \subseteq \Pi р_3(\Lambda) \times \Pi р_{(2,3)}(\Lambda)$ (отношение «предок») и $r_d \subseteq \Pi р_3(\Lambda) \times \Pi р_{(2,3)}(\Lambda)$ (отношение «потомки»). Первый и третий аргументы тернарных отношений r_a и r_d – значение измерения, второй аргумент – уровень измерения. Сечение тернарного отношения r_a посредством значения измерения v и уровня DL_i^k состоит из множества вершин, являющихся родительскими для вершины v, то есть

$$\text{Сеч}_{(v_i^j, DL_i^k)}(r_a) = (u_i^j \in V_i \,|\, (v, DL, u_i^j) \in \in r_a \wedge v = v_i^j \wedge DL = DL_i^k).$$

Сечение тернарного отношения r_d посредством значения измерения v и уровня DL_i^j состоит из множества вершин, являющихся потомками для вершины v, то есть

$$\text{Сеч}_{(v_i^j, DL_i^k)}(r_d) = (u_i^j \in V_i \,|\, (v, DL, u_i^j) \in \in r_d \wedge v = v_i^j \wedge DL = DL_i^k).$$

Основой многомерной модели является *базовый куб*, содержащий наиболее детализированные данные, соответствующие терминальным вершинам иерархии каждого измерения. Базовый куб C_b может быть представлен системой кортежей $\langle D_b, L_b, R_b \rangle$, где:

1) $D_b = \langle D_{b1}, D_{b2}, \ldots, D_{bq}, M_b' \rangle$ – кортеж измерений базового куба, $D_{bi} \subseteq \Pi р_1(\Lambda)$, $i = 1 \ldots q$, $M_{b'} \in \Pi р_1(\Lambda)$. $M_{b'}$ – измерение, представляющее показатель куба. Показатель куба – типизированная величина, являющаяся предметом анализа (например, количество проданного товара). Один базовый куб может содержать несколько показателей, организованных в иерархическую структуру. В этом случае этим показателям будут соответствовать несколько измерений показателей. Рассмотрим случай с

одним показателем, учитывая при этом, что все рассуждения могут быть обобщены на случай куба с произвольным числом показателей.

2) $L_b = \langle DL_{b1}, DL_{b2}, ..., DL_{bq}, ML_{b'} \rangle$ – кортеж уровней измерений куба, $D_{bi} \subseteq \Pi p_2(\Lambda)$, $i = 1...q$, $ML_{b'} \in \Pi p_2(\Lambda)$. $ML_{b'}$ – это уровень измерения показателя куба. Необходимо, чтобы уровни всех измерений были представлены наиболее детализированными данными соответствующих измерений.

3) R_b – это множество значений ячеек куба, то есть множество кортежей вида $x = \langle x_1, x_2, ..., x_q, m_x \rangle$, где $x_i \subseteq \Pi p_3(\Lambda)$, $i = 1...q$, $m_x \in \Pi p_4(\Lambda)$.

Текущее состояние операций над базовым кубом отражается в *многомерном кубе*. Многомерный куб C может быть представлен системой $\langle C_b, D, L, R \rangle$, где:

1) C_b – базовый куб.

2) $D = \langle D_1, D_2, ..., D_n, M' \rangle$ $(n \le q,\ D \subseteq D_b)$ – кортеж измерений куба. M' – измерение, представляющее показатель куба.

3) $L = \langle DL_1, DL_2, ..., DL_n, ML' \rangle > (L \subseteq \Pi p_2(\Lambda))$ – кортеж уровней измерений.

4) R – множество значений ячеек куба в виде кортежей $x = \langle x_1, x_2, ..., x_n, m_x \rangle$, $R \subseteq \subseteq \Pi p_{(3,\,4)}(\Lambda)$.

Операции над многомерными кубами

Для получения информации из базы данных посредством гиперкуба используются соответствующие операции. Все операции над многомерными кубами можно разбить на простейшие – *повышение уровня*, *применение функции*, *выборка*, *проекция* и операции, основанные на базе простейших, – *срез* и *навигация*. Аргументами каждой операции являются исходный куб C и куб-шаблон C^σ. Результатом каждой операции, применяемой к существующему кубу, является новый куб C'. В дальнейшем будем полагать, что исходный куб C представлен системой множеств $\langle C_b, D, L, R \rangle$, где C_b – базовый куб, $D = \langle D_1, D_2, ..., D_n, M' \rangle$, $L = \langle DL_1, DL_2, ..., DL_n, ML' \rangle$, R – это множество значений ячеек куба, представленное в виде кортежей. Для вычисления агрегированных значений необходимо также задать агрегирующую функцию f, используемую по умолчанию. Функция f определяет способ объединения значений нескольких кортежей в одно значение. Куб-шаблон C^σ представлен системой множеств $\langle C_b, P^\sigma, PL^\sigma, R^\sigma \rangle$, где C_b – базовый куб (базовый куб для куба-шаблона и многомерного куба, к которому применяется этот шаблон, один и тот же для всех операций), P^σ – множество измерений, PL^σ – множество уровней измерений, $R^\sigma = \varnothing$ (для всех операций).

Операция *повышение уровня* $C' = \varphi(C, C^\sigma)$ заключается в том, что значения измерений, уровень которых необходимо повысить, заменяются значениями, соответствующими более высокому уровню этих измерений, значения остальных измерений не изменяются. *Применение функции* $C' = \theta(C_1, C^\sigma)$ – получение агрегированных значений на основе детализированных с помощью функции агрегации. Для выполнения этой операции (а также операций *срез* и *навигация*) необходимо задать агрегирующую функцию f^σ, которая будет перекрывать агрегирующую функцию f многомерного куба C. *Выборка* $C' = \rho(C, C^\sigma)$ – выделение подмножества из исходного многомерного куба.

Проекция $C' = \pi(C, C^\sigma)$ – удаление измерения из многомерного куба при сохранении этого измерения в базовом кубе. Пусть куб-шаблон представлен следующим образом: $P^\sigma = \{d\}$, где d – это измерение многомерного куба C, на которое осуществляется проекция. $PL^\sigma = \{dl\}$, где dl – текущий уровень измерения d, на которое осуществляется проекция. Тогда операция проекции может быть задана следующим образом:

$$C'_b = C_b,$$

$$D' = D \setminus P^\sigma,$$

$$L' = L \setminus PL^\sigma,$$

$$R' = \{x \in R \mid \exists t \in R : \forall D_i \ne d\{t_i\} = \Pi p_i(\sigma_{t=x}(R))\}.$$

Срез $C' = \chi(C, C^\sigma) = \theta(\pi(C, C^{\sigma 1}), C^{\sigma 2})$ – это удаление выбранного измерения с последующей агрегацией измерений с использованием выбранной пользователем функции агрегации. Пусть куб-шаблон представлен следующим образом: $P^\sigma = \{d\}$, где d – это измерение, по которому производится срез. $PL^\sigma = \{dl\}$, где dl – текущий уровень измерения d, f^σ – применяемая функция агрегации, задаваемая пользователем. Тогда операцию *срез* можно представить через операции *применение функции* и *проекция* следующим образом:

$$C' = \chi(C, C^{\sigma}) = \theta(\pi(C, C^{\sigma 1}), C^{\sigma 2}),$$

где

$$P^{\sigma 1} = P^{\sigma} = \{d\},\ PL^{\sigma} = \varnothing,$$
$$P^{\sigma 2} = \varnothing,\ PL^{\sigma 2} = \varnothing,\ f^{\sigma 2} = f^{\sigma}.$$

Навигация $C' = \eta(C, C^{\sigma}) = \theta(\varphi(C, C^{\sigma 1}), C^{\sigma 2})$ – изменение уровня выбранного измерения с последующей генерацией нового куба с использованием операции «применение функции». Пусть куб-шаблон представлен следующим образом. $P^{\sigma} = \{d\}$, где d – это измерение, по которому мы осуществляем навигацию, $PL^{\sigma} = \{dl\}$, где dl – заданный уровень навигации, f^{σ} – применяемая функция агрегации, задаваемая пользователем. Функция f^{σ} перекрывает функцию f, определяющую метод агрегации по умолчанию. Предположим также, что $d – i$-е измерение множества измерений D многомерного куба C, т.е. $d = D_i$. Навигацию можно представить через уже рассмотренные операции *применение функции* и *повышение уровня*:

$$C' = \eta(C, C^{\sigma}) = \theta(\varphi(C, C^{\sigma 1}), C^{\sigma 2}),$$

где

$$P^{\sigma 1} = P^{\sigma} = \{d\},\ PL^{\sigma 1} = PL^{\sigma} = \{dl\},$$
$$P^{\sigma 2} = \varnothing,\ PL^{\sigma 2} = \varnothing,\ f^{\sigma 2} = f^{\sigma}.$$

Таким образом, алгебраическим описанием многомерного куба OLAP-систем и операций над многомерными кубами является алгебраическая система $\langle \{\Lambda\}, \{\varphi, \theta, \pi, \rho, \eta, \chi\}, \{r_l, r_v, r_a, r_d\} \rangle$.

Регулярная структура систем аналитической обработки данных

Агрегирование данных является механизмом построения многомерного куба для эффективного решения заданного класса задач. Опишем регулярные структуры, которые позволят использовать альтернативные классификации данных, оптимизировать доступ к ним, а также автоматизировать корректное внесение изменений в гиперкуб.

Для получения корректных результатов при агрегировании недостаточно введения понятия базового куба и определения многомерного куба через базовый. Необходимо, чтобы иерархическая структура каждого измерения многомерного куба удовлетворяла следующим требованиям:

1) все измерения многомерного куба должны быть попарно независимы, а показатель должен полностью определяться набором значений терминальных уровней иерархий измерений;

2) запрещается неполнота иерархий всех измерений;

3) в многомерном кубе не должно быть несбалансированных иерархий;

4) в многомерном кубе должно отсутствовать множественное наследование в иерархии измерения.

Поясним более конкретно эти требования. Попарная независимость измерений многомерного куба означает, что не существует функциональной зависимости между любыми двумя атрибутами двух произвольно выбранных измерений многомерного куба [6]. В этом случае считается, что структура многомерного куба удовлетворяет общей многомерной нормальной форме [7]. В случае если измерения взаимозависимы друг от друга, то эта аномалия исключается путем удаления зависимых частей иерархии одного из измерений, что приводит к изменению структуры многомерного куба.

Запрет неполноты иерархии, несбалансированности иерархии и множественного наследования рассмотрим на примере (рис. 1).

В этом примере показана ситуация, когда в иерархии некоторого измерения D_s существуют вершины, не связанные ни с одной родитель-

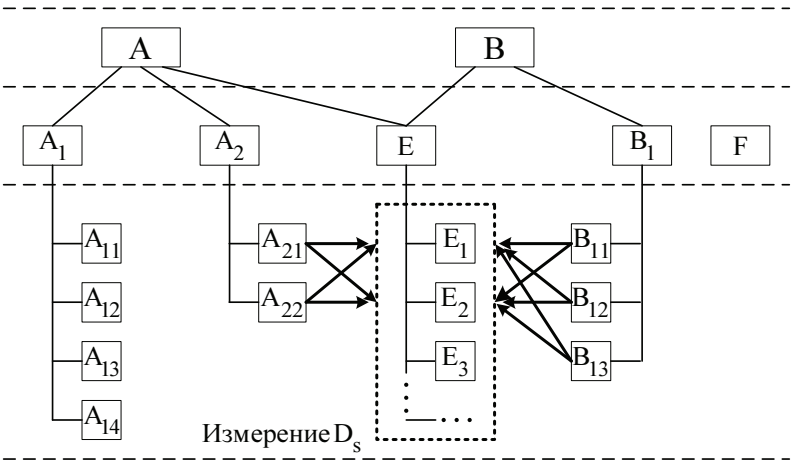

Рис. 1. Пример неполноты и множественного наследования в иерархии измерения

ской вершиной (например, фирма решает продавать новую группу товаров, однако реального наполнения склада нет). Более того, существуют вершины, которые могут быть отнесены к нескольким родительским вершинам (в случае анализа продаж это могут быть аксессуары, которые можно использовать с товарами из различных групп). Первый случай приводит к существованию значения измерения, не имеющего родительских вершин (неполнота иерархии).

Второй случай приводит к множественному наследованию. Однако и первый и второй случаи противоречат концепции корректного суммирования [8, 9, 10]. Согласно этой концепции, для выполнения корректного суммирования иерархии должны быть строгими и сбалансированными.

Вследствие сказанного для устранения неполноты иерархии введем дополнительную вершину *C*, которая будет являться родительской для *F*. Для устранения несбалансированности иерархии искусственно введем потомка вершины *F* – вершину *G*. Стоит отметить, что при появлении реальных потомков вершины F вершина *G* должна быть удалена из иерархии измерения. Результат изменений показан на рис. 2.

Аномалия множественного наследования может быть вызвана как концептуальными связями между родительскими вершинами и вершиной, располагающейся на более нижнем уровне иерархии, так и наличием альтернативных классификаций одного и того же понятия. Предложим способ устранения аномалии множественного наследования (рис. 3).

На верхнем уровне иерархии вводится дополнительная вершина *D*, которая будет родительской для *E*. Вместе с тем необходимо показать связь между вершинами *A*, *B* и *E*. Это можно сделать с помощью дополнительного трехмерного куба. Будем называть такой гиперкуб присоединенным. Идея присоединенного куба состоит в том, что ячейки этого куба содержат ссылки на значения измерения, в котором присутствует аномалия. Присоединенный куб состоит из трех измерений: разделяемое измерение, классификационное измерение и ссылочное измерение. Разделяемое измерение является общим для присоединенного куба и того многомерного куба, к которому относится данный многомерный куб. Классификационное измерение присоединенного куба позволяет классифицировать или ранжировать свойства сущностей, относящихся к разделяемому измерению. Ссылочное измерение определяет одно или несколько измерений, значения индексов которых будут находиться в ячейках присоединенного куба.

Вершины иерархии измерения необходимо пронумеровать. Вершины самого верхнего уровня, соответствующие наиболее агрегированным данным, будем обозначать одним индексом, причем нумерация начинается с нуля. Вершины, находящиеся на следующем уровне иерархии, обозначаются двумя индексами, разделенными точкой. При этом первый индекс совпадает с индексом родительской вершины, а нумерация второго индекса начинается с нуля. Вершины третьего уровня будут обозначаться тремя индексами, соответствующими предкам на первом и втором уровнях соответственно, и т.д.

В примере на рисунке 3 одним из измерений присоединенного куба будет исходное измерение D_s. В ячейках присоединенного многомерного куба будут содержаться ссылки в виде индексов на потомков *E* измерения D_s, которые связаны с другими значениями, содержащимися в измерении D_s присоединенного куба.

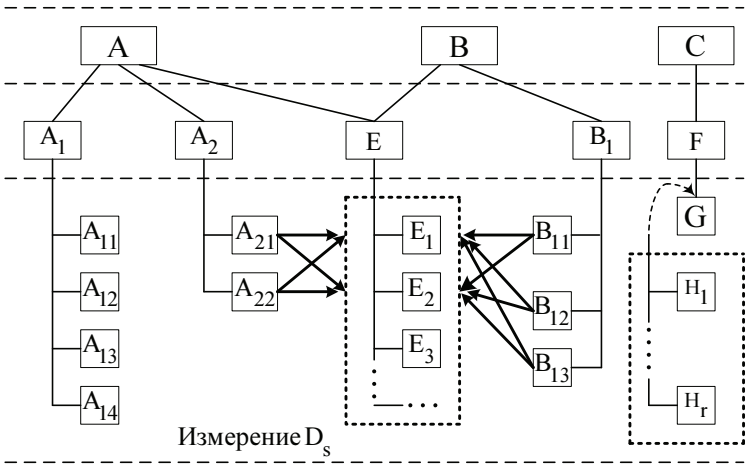

Рис. 2. Разрешение аномалии неполноты и несбалансированности иерархии измерения

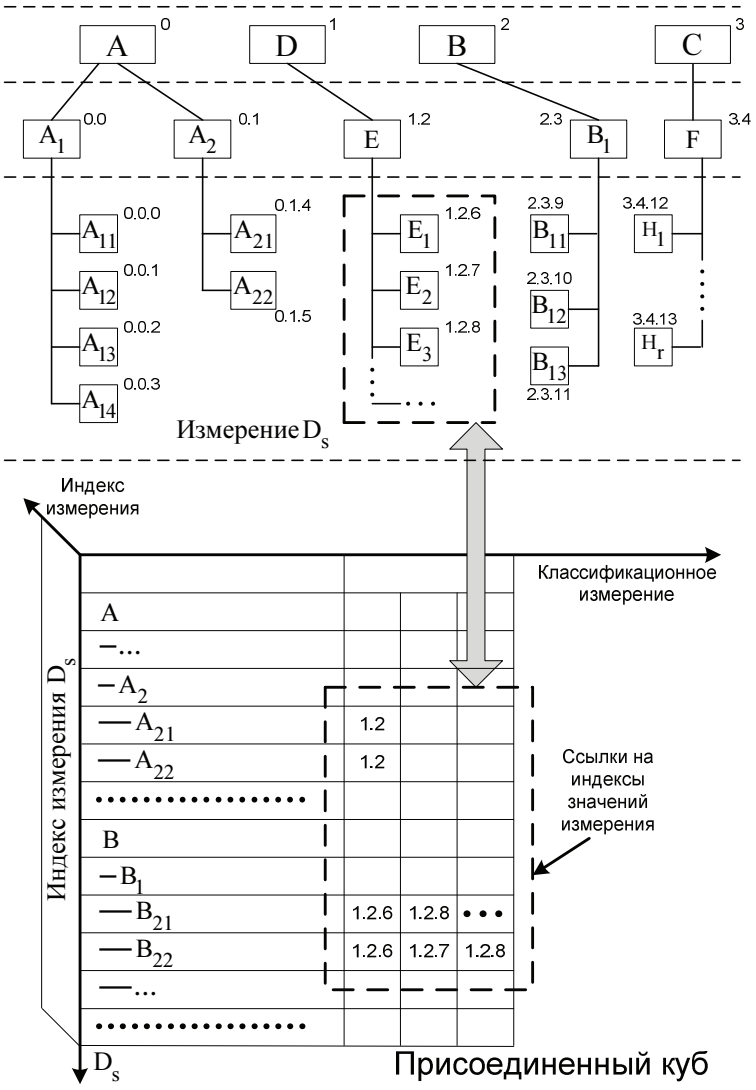

Рис. 3. Разрешение аномалии множественного наследования в иерархии измерения

Заключение

Разработанный подход на основе выделения наиболее детализированных данных обеспечивает следующие преимущества. 1. Произвольная последовательность выполнения операций без необходимости ресурсоемкой операции объединения с другими кубами. Соответствие значений между различными уровнями измерений гарантирует правильность результатов при выполнении запросов. 2. Сохранение результатов операции навигации для последующих навигаций. Это позволяет организовать кэш, в котором хранятся уже вычисленные кубы, и тем самым ускорить выполнение последующих запросов.

Предложенная регулярная структура систем аналитической обработки данных позволяет выполнять корректное вычисление агрегированных показателей, избегая множественного наследования. Стоит отметить, что в отличие от иерархий с множественным наследованием, связь между элементами основного куба и присоединенного не является жесткой, а иерархия измерения основного куба является строгой и сбалансированной. Таким образом, применение присоединенного куба позволяет как решить проблему множественного наследования, так и обеспечить альтернативную классификацию в иерархиях многомерного куба, не потеряв при этом регулярную структуру многомерного куба. Использование альтернативных классификаций, в свою очередь, позволяет решать аналитические задачи под различными углами зрения, соответствующим образом оптимизируя вычисления.

Разработанные математические формализмы легли в основу программно реализованной системы CuDBIS v. 1.02 [11]. Эта система позволяет выполнять основные операции администрирования над многомерными кубами и иерархиями измерений, а также служит для оптимизации структуры многомерного куба.

Список литературы

1. Codd E.F. Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. E.F. Codd and Associates, 1993.

2. Shoshani A. OLAP and statistical databases: similarities and differences // 16th ACM SIGACT SIGMOD SYGART Symp. on Principles of Database Systems, 1997. P. 185–196.

3. Inmon W.H. Building the Data Warehouse. Wiley, 2005. 543 p.

4. Laker K. OLAP Workshop 1: Basic OLAP Concepts [Электронный ресурс]. URL: http:// oracleolap.blogspot.com/2007/12/olap-workshop-1-basic-olap-concepts.html (дата обращения: 15.03.2009).

5. Woronowicz E. Relations and their basic properties // Journal of Formalized Mathematics. 1989. V. 1. URL: http://mizar.org/JFM/Vol1/relat_1.html (дата обращения: 16.03.2009).

6. Lehner W., Albrecht J., Wedekind H. Normal forms for multidimensional databases // Proceedings of the 10th International Conference on Scientific and Statistical Data Management (SSDBM'98), 1998. P. 63–72.

7. Lechtenbörger J., Vossen G. Multidimensional normal forms for data warehouse design // Information Systems. 2003. V. 28. № 5. P. 415–434.

8. Rafanelli M. and Shoshani A. STORM: A statistical object representation model // Proceedings of 5th International Conference on Statistical and Scientific Database Management, 1990. P. 14–29.

9. Lenz H.-J., Shoshani A. Summarizability in OLAP and statistical databases // Proceedings of 9th International Conference on Scientific and Statistical Database Management, 1997. P. 132–143.

10. Hurtado C.A., Mendelzon A.O. Reasoning about summarizability in heterogeneous multidimensional schemas // Proceedings of the 8th International Conference on Database Theory, 2001. P. 375–389.

11. Семченков С.Ю. CuDBIS v. 1.02. Свидетельство о регистрации программы для ЭВМ № 2009613357 от 26 июня 2009 г. (ФГУ ФИПС).

FORMALIZED STRUCTURE OF ANALYTICAL DATA PROCESSING SYSTEMS BASED ON OLAP TECHNOLOGY

S.Yu. Semchenkov

A mathematical model is proposed of analytical data processing systems built on the basis of the multidimensional cube concept. Conditions for correct aggregation of indicators are formulated. Techniques to eliminate hierarchy anomalies are considered.

Keywords: OLAP systems, mathematical model of multidimensional cube, hierarchy anomalies.