Информационные процессы, Том 6, № 3, 2006, стр. 264–271.

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ

Некоторые свойства измерений аналитического пространства

А.В. Висков

Российский университет Дружбы Народов, Москва, Россия
Поступила в редколлегию 29.06.2006

Аннотация—В статье рассматриваются вопросы оперативного многомерного анализа данных. Вводится и формально определяется понятие аналитического измерения. В литературе могут быть найдены различные определения и концепции взаимосвязи между уровнями агрегации измерений. В этой статье утверждается, что между ними существует связь "часть-целое". Основываясь на введенном определении и особенностях связей между элементами аналитических измерений, определяются и доказываются их свойства.

1. ВВЕДЕНИЕ

Одним из наиболее ценных активов большинства организации в настоящее время является информация. Любая крупная компания рано или поздно сталкивается с задачей сбора согласованной информации необходимой для принятия тактических и стратегических решений. Возникает потребность в организации централизованного хранения информации с целью ее последующего анализа и получения отчетности. Одним из вариантов такого хранения информации является создание хранилища данных (как правило, информационно-аналитической системы масштаба предприятия) [1].

Для анализа содержащейся в хранилище данных информации, как правило, применяется технология оперативного анализа данных (OLAP, Online Analytical Processing). OLAP - это категория программных средств, которые дают возможность осуществлять управление, администрирование и анализ данных, с целью получения глубокого осмысления информации посредством быстрого, консолидированного, интерактивного доступа к широкому спектру всевозможных аспектов информации, полученной преобразованием сырых, необработанных данных, отражающих реальную многомерность предметной области в понимании пользователей. Функциональность OLAP приложений характеризуется динамическим многомерным анализом консолидированных данных предприятия, в процессе поддержки аналитической и навигационной деятельность конечных пользователей [2], [3].

Средства OLAP представляют данные так, как если бы они были помещены в n-мерное аналитическое пространство, предоставляя возможность изучать их в терминах фактов, являющихся предметов анализа, и измерений, показывающих различные аспекты, в соответствии с которыми можно проводить анализ показателей предметной области, определяемых рассматриваемыми фактами. Эта концепция послужила импульсом для развития схемы данных в виде звезды, при которой предмет анализа находится в центре, а окружают его аналитические измерения.

2. ОПРЕДЕЛЕНИЕ АНАЛИТИЧЕСКОГО ИЗМЕРЕНИЯ

Факты сами по себе почти бесполезны. Они приобретают значения только тогда, когда их определяют аналитические измерения. Разговаривать о продаже бесполезно, если не известно кто продает, когда, кому и т. д.

Измерения составляют множество объектов одного или нескольких типов, организованных в виде иерархической структуры и обеспечивающих информационный контекст показателей, описывающих анализируемые факты. Каждое измерение включает направления консолидации данных, состоящие из серии последовательных уровней обобщения, где каждый вышестоящий уровень соответствует большей степени агрегации данных по соответствующему измерению [4], [5], [6]. Объекты, совокупность которых образует измерение, называются членами измерений. В качестве примера аналитического измерения можно рассмотреть измерение времени с уровнями агрегации (обобщения): “Дни”, “Месяцы”, “Кварталы”, “Годы” - наиболее часто используемое в анализе, которое может содержать следующие члены: “17 января 1977 года”, “январь 1977 года”, “1-ый квартал 1977 года” и “1977 год”.

Как уже было сказано, объекты в измерениях могут быть различного типа, например “производители” - “марки автомобиля” или “годы” - “кварталы”. Эти объекты должны быть организованы в иерархическую структуру так, чтобы объекты одного типа принадлежали только одному уровню иерархии.

На основе любого измерения можно получить граф, показывающий, каким образом анализируемые данные могут быть агрегированы по аналитическому измерению. Таким образом, аналитическое измерение можно определить следующим образом.

Определение. Измерение - это связный направленный граф, описывающий точку зрения на анализируемые данные. Каждая вершина в этом графе соответствует уровню агрегации, а ребро отражает тот факт, что каждая сущность на уровне, определяющем конец ребра, декомпозируется на множество сущностей нижележащего уровня, соответствующего началу ребра (то есть, ребро отражает суть отношения “часть-целое” между сущностями уровней).

3. СВЯЗИ МЕЖДУ ЭЛЕМЕНТАМИ ИЗМЕРЕНИЙ

Измерение включает уровни, которые образуют области различной детализацией. Эта детализация показывает, как элементы измерения группируются для применения агрегационных функций. Связи, определенные между элементами на разных уровнях агрегации, являются композицией.

Как показано на рисунке 1, можно выделить три различных, независящих от области применения типа отношений “часть-целое”, определяемых составной структурой целого.

а) Куча - в случае отсутствия определенной структуры целого, считающегося гомогенным (например, кучка риса).

б) Коллекция - в случае, когда рассматриваются различные элементы целого, имеющего однородную составную структуру (например, колонна грузовиков).

в) Комплекс - в случае, когда различные части целого выполняют различные роли, целое является комплексом с гетерогенной составной структурой (например, части некоторого механизма).

Куча, коллекция и комплекс представляют пограничные случаи во всем множестве различных вариантов целого, как с полным отсутствием структуры, так и со сложной внутренней организацией.

Главная цель определения связей между различными сущностями аналитического измерения состоит в том, чтобы показать, каким образом могут применяться агрегационные функции (такие как сумма, минимум, максимум, среднее, и т.д.). Так как эти функции рассматривают элементы измерения как идентичные (играющие в агрегации одну и ту же роль), то эти связи должны рассматриваться как связи между элементами коллекции. С этого момента, отноше-

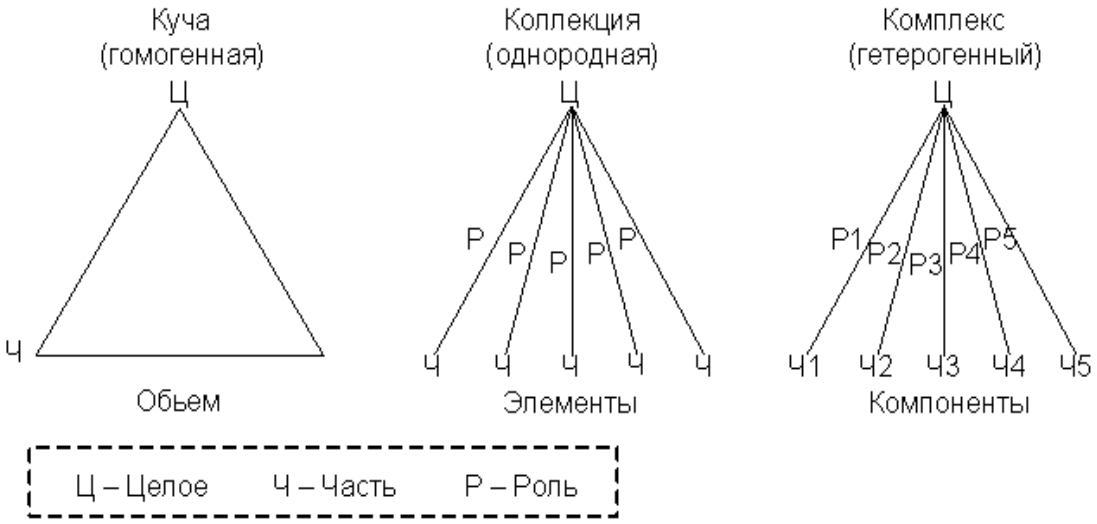

Рис. 1. Типы целого.

ния "часть-целое" между уровнями агрегации в аналитическом измерении должны пониматься как образующие коллекцию.

В случае рассмотрения коллекции, для описания и обоснования свойств целого, а также связей между составляющими элементами можно использовать аксиоматику классической экстенсиональной мереологии [7], [8].

1) **Существование.** Если A является частью B, то и A и B существуют.

2) **Асимметричность.** Если A является частью B, то B не является частью A.

3) **Транзитивность.** Если A является частью B и B является частью C, то A является частью C.

4) **Дополнение.** Если A является некоторой частью B, тогда существует другой субъект C, который является оставшейся частью B.

5) **Экстенсиональность.** A и B состоят из одних и тех же частей тогда и только тогда когда A и B совпадают.

6) **Сумма.** Для любых двух субъектов всегда найдется составленный из них третий субъект.

Аксиома 5 в общем случае семантически не верна, так как одна и та же коллекция элементов может образовывать различные целые (например, два клуба, в одно и тоже время могут иметь одно и то же множество членов). Однако в случае применения агрегационных функций на элементах обоих коллекций результат будет получаться одинаковым. Кроме того, следует отметить, что может существовать более чем один способ декомпозиции целого, некоторые объекты могут рассматриваться как коллекции элементов различных типов (например, год является коллекцией триместров или кварталов) [5].

4. СВОЙСТВА АНАЛИТИЧЕСКИХ ИЗМЕРЕНИЙ

Исходя из определения аналитического измерения и аксиом мереологии, можно сформулировать некоторые свойства аналитических измерений.

Свойство 1: Граф измерения не содержит циклов.

Доказательство 1 (от противного): Предположим, что в графе измерения есть цикл. В цикле участвуют хотя бы два уровня. Тогда по аксиоме 3 (транзитивности) получаем, что элемент

A одного уровня есть часть элемента B другого уровня и наоборот. Получаем противоречие с аксиомой 2. Значит предположение не верно.

Свойство 2: Для каждого измерения существует единственный атомарный уровень агрегации, содержащий элементарные (которые не могут быть разделены на более мелкие) сущности.

Доказательство 2: Учитывая свойство 1, существует как минимум один уровень, элементы которого не содержат частей. Если существует более одного атомарного уровня, то так как граф измерения связный и, учитывая аксиому 3, будет существовать элемент E, который является композицией элементарных сущностей каждого из атомарных уровней. Согласно аксиоме 5, все коллекции элементарных сущностей составляющих E должны являться одной и той же коллекцией элементов. Таким образом, существует только один атомарный уровень.

Свойство 3: Для любого измерения, может существовать уровень "Все", содержащий элементы, составленные всеми элементарными сущностями измерения. Если такой уровень существует то:

а) его элементы не являются частями никакого другого уровня агрегации;

б) этот уровень агрегации содержит в точности один элемент;

в) этот уровень единственный в измерении.

Доказательство 3: Последовательно применяя аксиому 6, можно построить элемент E, состоящий из всех элементарных сущностей измерения.

а) Если E будет являться частью некоторого элемента E' то, по аксиоме 4 будет существовать элементарная сущность, не входящая в E, что противоречит предположению. Поэтому, E принадлежит уровню, элементы которого не являющегося частью никакого другого уровня в измерении.

б) Если этот уровень будет содержать два элемента, оба содержащие все элементарные сущности, то в соответствии с аксиомой 5 они будут являться одним и тем же элементом.

в) Этот уровень единственный, потому, что если бы существовал другой уровень, элементы которого включают все элементарные сущности, то они бы совпадали с элементом, имеющемся на уровне "Все" (по аксиоме 5).

Свойство 4: Те уровни элементы, которых не являются коллекцией элементов любого другого уровня (т.е. они не являются началом ребра в графе измерения) могут быть связанны некоторым ребром с уровнем "Все".

Доказательство 4: Элемент уровня "Все" может быть разложен на сущности любого уровня, покрывающие атомарный уровень. Если существует уровень, не покрывающий атомарный уровень, то согласно аксиоме 6, к его элементам могут быть добавлены новые элементы, составленные из коллекций недостающих элементов атомарного уровня. В результате получим сущности, покрывающие атомарный уровень, а соответственно элемент уровня "Все" может быть разложен на эти сущности, поэтому по определению измерения можно соединить этот уровень ребром с уровнем "Все".

Свойство 5: Каждый элемент уровня не являющегося атомарным содержит как минимум одну часть.

Доказательство 5: Сущности, не включающие частей, являются элементарными, а все элементарные сущности согласно свойству 2 располагаются на атомарном уровне.

Свойство 6: Каждый элементы уровня, не являющегося атомарным может иметь более чем одну часть.

Доказательство 6: В соответствии с аксиомой 4, если элемент включает часть, то найдется другой элемент, являющейся оставшейся частью исходного элемента.

Свойство 7: Элемент может являться частью нескольких коллекций в одно и тоже время.

Доказательство 7: Не существует мереологической аксиомы запрещающей вхождение элементов в несколько коллекций, наоборот обязательным условием является обеспечение суммируемости (объединения). В качестве доказательства можно рассмотреть пример на рисунке 2.

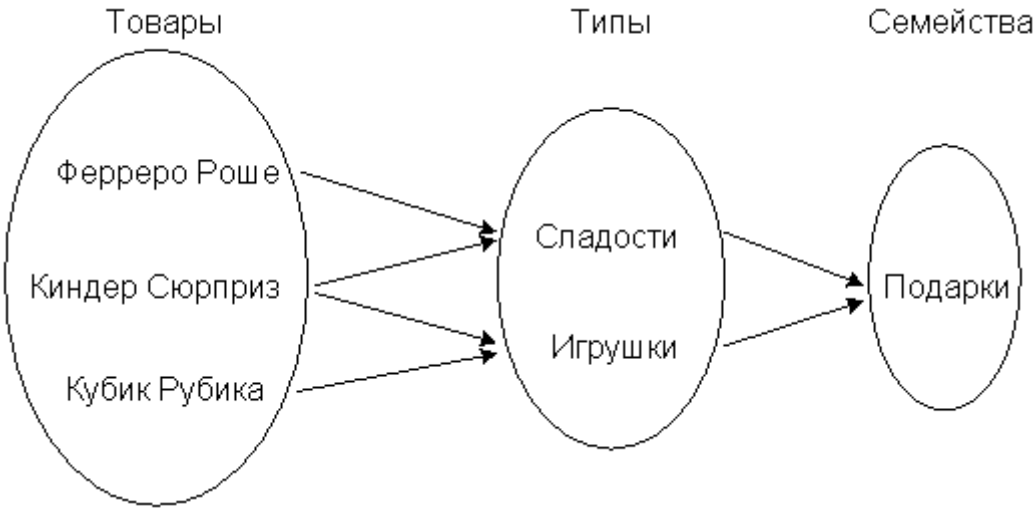

Рис. 2. Пример образования целого с пересечением.

Выше были рассмотрены свойства измерений, описывающие возможные связи, существующие между элементами и уровнями внутри измерения. Важно также проанализировать возможные отношения между элементами разных аналитических измерений как внутри одной звездной схемы, так и в различных схемах. Ниже будут рассмотрены два свойства измерений, касающиеся таких семантических связей как агрегация и специализация.

Свойство 8: Если элементарные сущности в измерении D являются частью элементарных сущностей в измерении D_A, тогда граф измерения D_A будет являться подграфом графа измерения D. Следует отметить, что атомарные элементы измерения D будут отличаться от атомарных элементов измерения D_A, и будут являться их частями.

Доказательство 8: Сгруппируем элементы атомарного уровня измерения D в коллекции таким образом, чтобы каждый элемент в этой коллекции являлся частью одного и того же элемента атомарного уровня в измерении D_A. Упорядочим получившиеся коллекции в соответствии с тем же критерием (иерархией), что используется для группировки элементов в измерении D_A. Тогда структура отношений "часть-целое" между уровнями построенная для измерения D, будет повторять структуру отношений графа D_A. Однако следует отметить, что элементарные сущности измерения D, могут быть упорядочены и по другим критериям, составляя другие отношения "часть-целое". Поэтому граф измерения D_A ,будет являться некоторым подграфом графа измерения D.

На рисунке 3 приведен пример, в котором для одного случая продажи изучаются по измерению "Цветные товары", а для другого (в рамках другой схемы "звезда") продажи рассматриваются по измерениям "Товары" и "Цвета". Элементы в этих измерениях агрегированы, для того чтобы показать какого типа товары продавались, и какого цвета они были. На рисунке 3 показано, что граф составного измерения будет содержать как минимум граф каждого из измерений определяющих структуру отношений "часть-целое" для всех элементарных сущно-

стей. Следует, однако, заметить, что, например, элементы уровня “Цвет” измерения “Цвета” и элементы уровня “Цвет” измерения “Цветные товары” не совпадают. В то время как первые представляют собой цвета, вторые являются группами товаров объединенных по цветам.

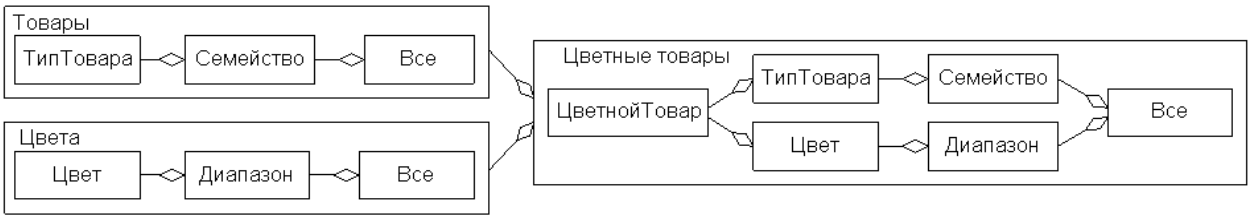

Рис. 3. Пример агрегации измерений.

Еще одним важным отношением между измерениями является отношение специализации.

Свойство 9: В общем случае, уровень и его специализация могут не принадлежать одному и тому же измерению.

Доказательство 9: Предположим, что и уровень L и его специализация L_s находятся в одном измерении. Для того чтобы определить сеть с уровнем “Все”, может потребоваться добавить на уровне L_s некоторые элементы, так как в таком случае L_s должен покрывать атомарный уровень. Эти элементы, могут не удовлетворять критерию специализации. Поэтому, не всегда возможно обеспечить принадлежность обоих уровней одному измерению.

На рисунке 4 показан пример, когда измерение “Люди” специализированно в виде измерения “Продавцы” на уровне “Профессия” (сплошная непрерывная стрелка). Это специализированное измерение содержит уровень, элементами которого являются все люди, являющиеся продавцами, и еще один уровень с единственным элементом (который является сущностью “Профессии”), представляющим собой множество всех продавцов. Разрывная линия показывает, что один уровень является специализацией другого (элемент уровня “Все” измерения “Продавцы” является сущностью “Профессии” в случае, когда выполняется критерий специализации “Профессия = ‘Продавец’”). Уровень “ВозрастнаяГруппа” для измерения “Продавцы” не рассматривается. Следует отметить, что если бы такой уровень в измерении “Продавцы” рассматривался, то он бы не являлся специализацией соответствующего уровня в измерении “Люди”, так как его элементы отличались бы от элементов на уровне “ВозростнаяГруппа” в измерении “Люди” (они включали бы меньше людей).

Обобщая пример, следует отметить что, если D_s это специализация измерения D на уровне L, то в этом случае D_s включает, по меньшей мере, уровень L_s (специализацию уровня L), и специализации каждого из уровней в D, содержащих часть элементов L_s. Эти специализированные уровни содержат в точности те элементы соответствующих уровней из D, которые являются любой частью элемента L_s. Кроме этих обязательных уровней, D_s также возможно включаются и другие уровни (которые не являются специализацией каких либо уровней в D), элементы которых не содержатся в D.

5. ЗАКЛЮЧЕНИЕ

Рассмотренные свойства измерений позволяют сделать некоторые выводы относительно структуры графа измерения, а также специфики связей между уровнями измерения и их элементами. Исходя из определения измерения, и учитывая свойства 1 и 2, можно сделать вывод, что в общем случае граф уровней агрегации измерения образует полусеть. Кроме того, свойство 3 и 4 показывает, что в графе измерений, для того, чтобы получить сеть, всегда

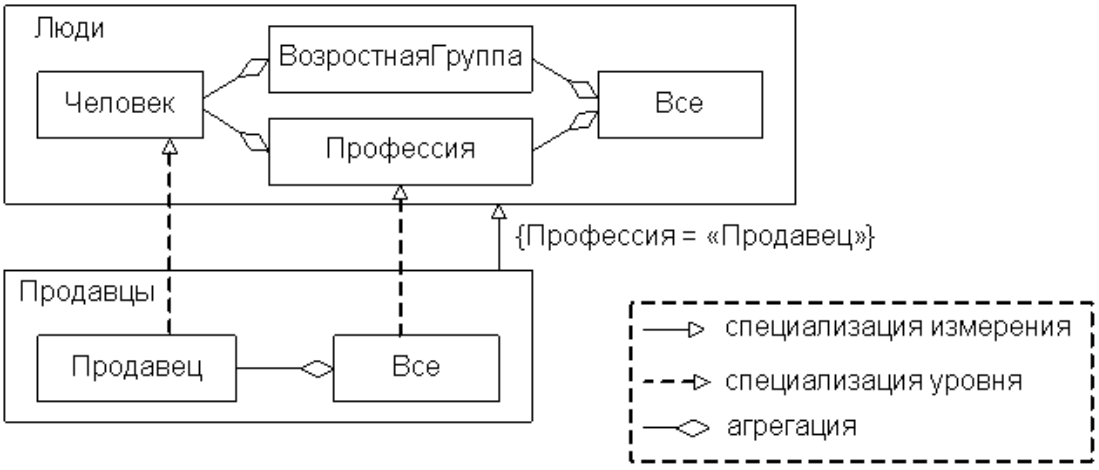

Рис. 4. Пример специализации измерений.

можно определить уровень “Все”. Свойство 5 и 6 свидетельствуют, что отношение между двумя уровнями будет включать 1..N частей для каждого целого. То есть для каждого элемента на вышележащем уровне всегда можно определить его разложение на нижележащем уровне в соответствующей иерархии. Свойство 7 поясняет, что часть может входить в состав как одного целого, так и нескольких. Следует отметить, что если элемент нижележащего уровня может являться частью только одного элемента вышележащего уровня, то в таком случае он обязательно входит в состав некоторого элемента вышележащего уровня, если же множества разложений элементов могут пересекаться, то, вообще говоря, элемент нижележащего уровня может и не входить в элемент вышележащего уровня. Однако, учитывая аксиому 6, которая свидетельствует о том, что для любой части всегда можно определить целое, получаем, что в случае необходимости всегда можно доопределить элемент вышележащего уровня, который будет включать рассматриваемый элемент нижележащего уровня. Это означает, что элемент нижележащего уровня всегда включается как минимум в один элемент вышележащего уровня в соответствующей иерархии.

На рисунке 5 показаны допустимые кардинальные отношения между элементами уровней в агрегационных иерархиях. Можно выделить два вида отношений, оба с как минимум одной частью для каждого целого. В наиболее распространенном случае оказывается, что существует в точности одно целое для каждой части. Однако также возможно, что определенная часть принадлежит нескольким целым. В таком случае, если найдется часть, которая не участвует ни в каком целом, всегда могут быть построены целые, так что каждая часть будет входить как минимум в одно целое.

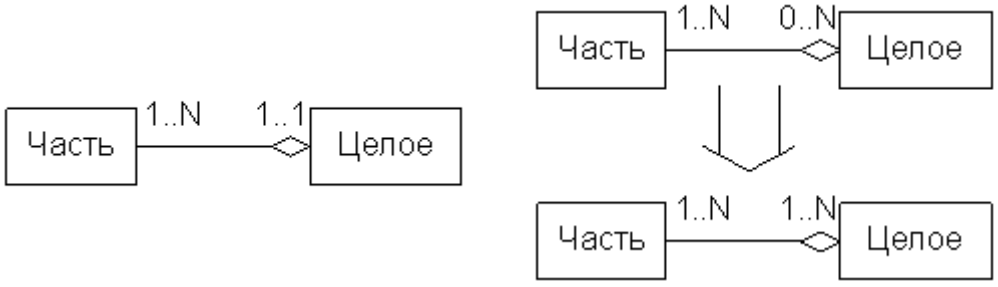

Рис. 5. Возможные кардинальные отношения между элементами уровней в аналитической иерархии

Специализации измерения предоставляет возможность показать специфику отношения "часть-целое" для некоторого подмножества элементов измерения, а также выделить общие для этого подмножества атрибуты. Все сущности некоторого уровня имеют общие свойства, так как уровень определяет заданный класс объектов, выполняющих одну и ту же роль в коллекции. Отношение ассоциация, также как и агрегации наследует в рамках специализации. Таким образом, отношение специализации имеет большое значение в вопросе повторного использования аналитических измерений. Существует возможность осуществить переход от схемы "звезда" S_1 к схеме "звезда" S_2, не только в том случае, когда обе они используют одни и те же измерения, но так же и в том случае, когда измерение из S_1 является специализацией измерений из S_2.

Рассматриваемые семантические особенности отношения между элементами измерений важны не только для конечных пользователей - аналитиков, они позволяют также оптимизировать запросы и повышать их производительность, позволяя, например, осуществлять параллельные вычисления по разным специализациям измерений.

Как показали исследования, наиболее серьезные проблемы, препятствующие в настоящее время внедрению средств многомерного анализа, связаны, прежде всего, с качеством данных предоставляемых для анализа, а также с производительностью аналитических запросов. При чем проблемы связаны в первую очередь не с объемом обрабатываемых данных, а со схемой их представления. Лучшее понимание структуры аналитических измерений как сущностей, определяющих аналитическое пространство, дает возможность улучшить схему построения многомерной модели, тем самым, повышая качество самих анализируемых данных, а также оптимизировать запросы, выполняющиеся вдоль аналитических измерений.

СПИСОК ЛИТЕРАТУРЫ

1. Inmon W. H. Building the Data Warehouse. *New York: John Wiley & Sons Inc.*, second edition, 1996.

2. Pendse N. The OLAP Report - What is OLAP? *Available at the URL http://www.olapreport.com/fasmi.html - Business Intelligence Ltd.* 2001.

3. OLAP and OLAP Server Definitions. *OLAP Council. - Available at the URL http://www.olapcouncil.org/research/glossaryly.htm*, 1997.

4. Codd E. F., Codd S. B., Salley C. T. Providing OLAP to user-analysts: An IT mandate. *Technical report - E. F. Codd & Associates.* 1993.

5. Abello A., Samos J., Saltor F. Understanding Analysis Dimensions in a Multidimensional Object-Oriented Model. *In 3rd International Workshop on Design and Management of Data Warehouses (DMDW).* SwissLife, 2001.

6. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. *Методы и модели анализа данных: OLAP и Data Mining.* СПб.:БХВ-Петербург, 2004.

7. Gerstl P., Pribbenow S. Midwinters, end games, and body parts: A classification of part-whole relations. *International Journal of Human-Computer Studies.* 43(5, 6):865-889, 1995.

8. Artale A., Franconi E., Guarino N., Pazzi L. Part-Whole relations in Object-centered systems: an overview. *Data and Knowledge Engineering (DKE).* 20, 1996.

Статью представил к публикации член редколлегии П.П. Бочаров