Вестник УГАТУ
Уфа : УГАТУ, 2012 Т. 16, № 6 (51). С. 189–201

УДК 004.62

В. В. Антонов, Г. Г. Куликов, Д. В. Антонов

ТЕОРЕТИКО-МНОЖЕСТВЕННАЯ МОДЕЛЬ ИС ДЛЯ МНОГОМЕРНОГО АНАЛИТИЧЕСКОГО АНАЛИЗА, ОТВЕЧАЮЩАЯ ТРЕБОВАНИЯМ ХРАНИЛИЩ ДАННЫХ

В статье рассмотрены проблемы построения математической модели предметной области с позиций методов, учитывающих нечеткость описаний модели исследуемого объекта. Рассмотрена возможность представления бизнес-процессов в виде совокупности взаимодействующих семантически определенных объектов. *Предметная область*; *семантическая модель*; *многомерная модель*; *пространства данных*

Окружающий нас мир многомерен и обычно каждый объект характеризуется множеством параметров. При построении модели, как правило, приходится снижать размерность мира, очерчивая круг параметров, представляющих интерес для исследования. В процессе анализа данных, поиска решений часто возникает необходимость в построении зависимостей между различными параметрами, число которых может варьироваться в широких пределах. Традиционные средства анализа, оперирующие данными, которые представлены в виде таблиц реляционной БД, не могут в полной мере удовлетворять таким требованиям. Для анализа информации наиболее удобным способом ее представления является многомерная модель или гиперкуб, ребрами которого являются последовательности значений одного из анализируемых параметров – измерения, относящиеся к анализируемой предметной области. Это позволяет анализировать данные сразу по нескольким измерениям, т. е. выполнять многомерный анализ. Множественность измерений предполагает представление данных в виде многомерной модели. На пересечениях осей измерений располагаются данные, количественно характеризующие анализируемые факты – меры (рис.1).

Таким образом, имеем множество измерений D, любая точка в данном пространстве может быть представлена $(d_1,\ldots,d_n)$, $d_i \in D$, с мерой $(m_1,\ldots,m_k)$, $m_i \in M$, где M – множество мер.

Тогда многомерное пространство определено с одной стороны схемой R, атрибутами которой являются множество измерений и множество мер $R = <D, M>$, а с другой стороны множеством отношений r над этой схемой R.

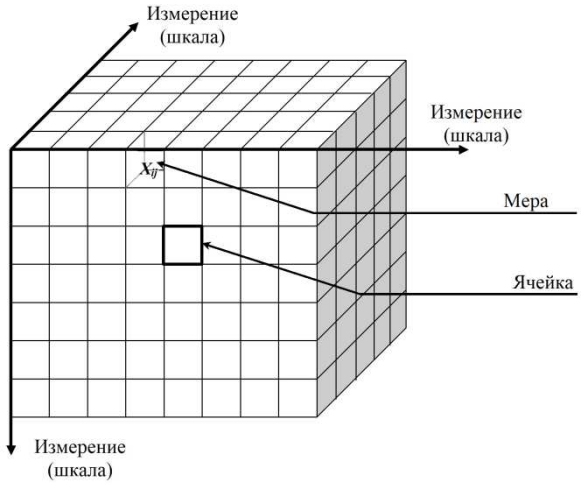

Рис. 1. Многомерное представление данных

Т. е. $\forall A \in D$, существует проекция $r[A]$ такая, что $\forall a \in r[A]$: $a \in$ ALL, где ALL – все возможные значения измерения. Само многомерное пространство может быть определено декартовым произведением

$$\text{Space}(r) = \{\otimes_{A \in D}(r[A] \cup \text{ALL}) \cup \{0,\ldots,0\}\}.$$

Основным способом исследования задач анализа данных является их отображение на формализованный язык и последующий анализ полученной модели. С увеличением размеров и сложности системы существенно усложняется ее моделирование с помощью известных математических выражений, так как увеличивается число переменных и параметров. В результате, создание адекватной модели становится практически невозможным. Вместо этого Л. Заде предложил лингвистическую модель, которая использует не математические выражения, а слова, отражающие качество.

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

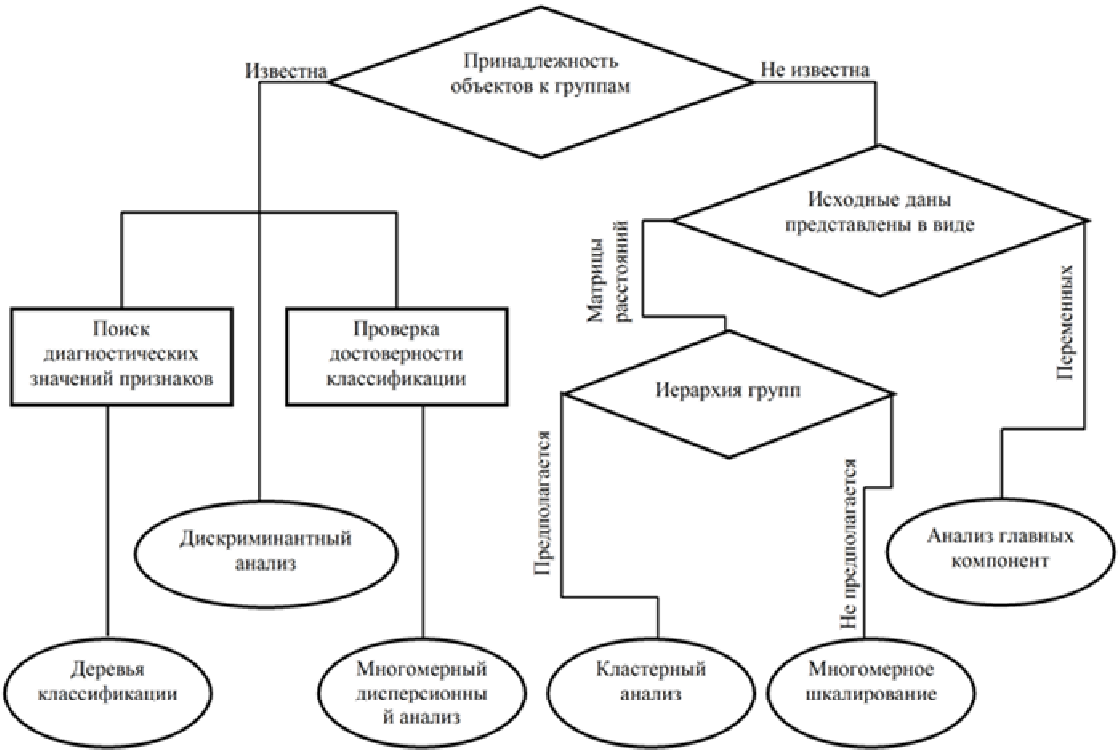

Рис. 2. Применимость методов многомерного анализа данных

Применение словесной модели не обеспечивает точность, аналогичную математическому моделированию, но позволяет создать достаточно качественную модель. В этом случае предметом обсуждения становится нечеткость слов языка описания системы.

Основная проблема анализа таких многомерных матриц данных заключается в том, что человеческий мозг не способен одновременно оперировать более чем тремя измерениями пространства. Для сведения многомерных данных к двум измерениям с минимальными потерями информации была разработана специальная группа методов статистического анализа данных – многомерный анализ данных. Существует большое количество подходов к анализу многомерных данных. В зависимости от того, какой именно ответ необходимо получить, производится и выбор метода (рис. 2).

Методы анализа, работающие только с одной переменной в определенный момент времени, получили название одномерных методов, позволяют довольно просто получить как первоначальную информацию для анализа, так и целые цепочки взаимно независимых результатов. Методы одномерного анализа, в частности методы функционально-стоимостного анализа, строятся в предположении существования и воспроизводства линейных связей, тогда как методы многомерного строятся в предположении существования и моделирования нелинейных связей. Таким образом, одномерный анализ представляет частный случай многомерного и его отправную точку. Практически все задачи одномерного анализа ставятся и решаются в предположении того, что в природе существует так называемый гауссовский закон распределения данных. То же самое происходит, когда речь идет о решении некоторого класса специфических многомерных задач, эмпирическое распределение данных в которых хорошо согласуется с гауссовским распределением.

Таким образом, многомерные данные при рассмотрении вопроса об анализе по одному измерению могут рассматриваться как одномерные. Последовательный анализ по каждому измерению также может быть рассмотрен как одномерный анализ одномерных данных. Полученные независимые результаты могут быть представлены в виде вышеприведенной схемы с агрегированием данных по каждому измерению с соответствующими мерами. При установлении соответствующего отношения в результате может быть получена новая многомерная модель в свойствах данных которой уже присутствуют зависимости между измерениями, и эти данные

уже не могут анализироваться только по одному измерению как одномерные данные методами одномерного анализа (рис. 3). Можем говорить о некотором многомерном информационном объекте.

В качестве примера многомерных данных рассмотрим данные ГИС, которые являются сочетанием обычных баз данных с атрибутивными данными с географически организованной информацией.

В ГИС форме представления координатных данных соответствуют два основных подкласса моделей – векторные и растровые.

В растровых моделях в качестве атомарной модели используют двумерный элемент – пиксель (ячейка). Упорядоченная совокупность атомарных моделей образует растр, который, в свою очередь, является моделью карты или геообьекта. Т. е., данные представляются в виде: X, Y – пространственные координаты, Z – зависящая от них переменная. В общем случае каждую точку сетки с координатами (x, y) характеризует некоторый вектор состояний (z_1,…, z_n).

Для всей сетки получаем набор векторов Z_1,…, Z_n – параметров в точках сетки. Очевидно, что часть параметров являются координатами, и пространственное положение может быть выражено через относительные единицы, например, как обратно пропорциональное квадрату расстояний между объектами [8]. Показатели состояния Z_1,…, Z_n разделяются на:

- входные переменные C_i (i = 1,…, p), полученные тем или иным способом;
- выходные D_j (j = 1,…, q) – те которые выражаются через входные, т. е. $D_j = F(C_1,\ldots, C_p)$, j = 1,…, q, $p + q = n$.

Перевод условий практической задачи на язык математических моделей всегда был трудным и зачастую приводил к потере трудноформализуемой качественной информации. Многие современные задачи управления просто не могут быть решены классическими методами из-за очень большой сложности математических моделей, их описывающих. До появления теории нечетких множеств многие характеристики с присущими им неопределенностями игнорировались при моделировании. На рис. 4 показаны области наиболее эффективного применения современных технологий управления. Для систем с неполной информацией и высокой сложностью объекта управления оптимальными являются нечеткие методы управления (нечеткие системы управления и нечеткие нейронные сети).

В условиях применения автоматизированных систем происходит трансформация функций человека, возникают новые связи между человеком и системой, некоторые функции полностью передаются системе. Компьютеризация способствует расширению возможностей субъекта, порою качественно меняя содержание его деятельности. И эта деятельность должна быть четко определена и закреплена за субъектом так же, как и за автоматизированной системой (АС).

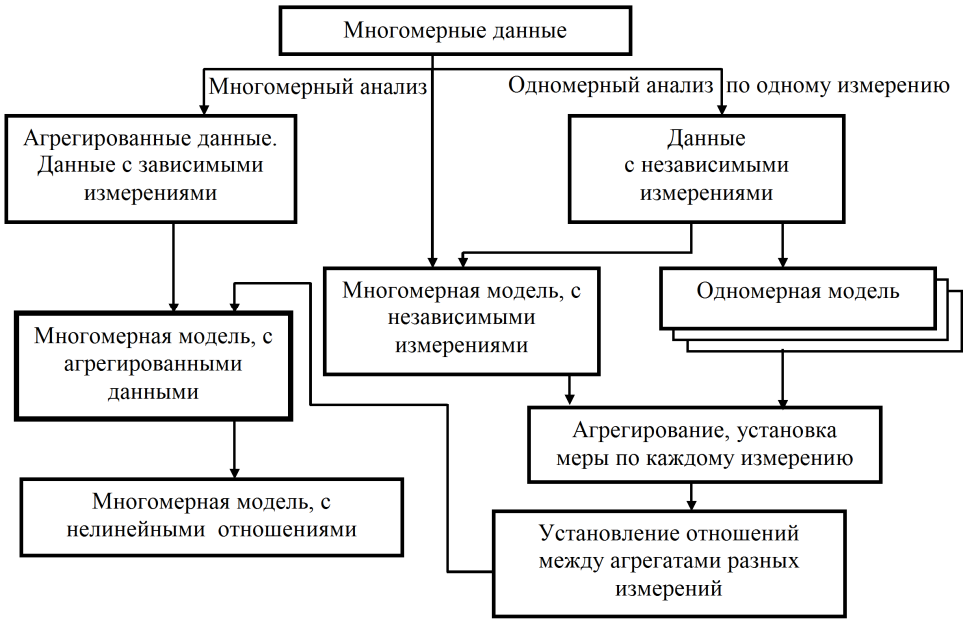

Рис. 3. Зависимость мерности данных от типа анализа

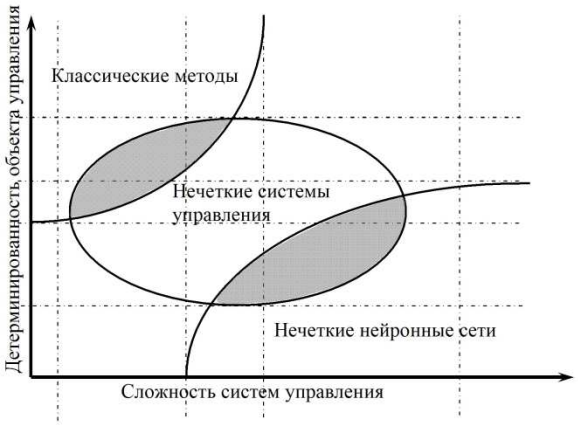

Рис. 4. Области эффективного применения технологий управления

Самая простая автоматизированная система предполагает взаимосвязь типа: человек – АС – человек, выделяя тем самым два элемента: человек – АС и АС – человек. Для сложных автоматизированных систем, распределенных во времени и пространстве, количество элементов увеличивается до четырех: человек – АС, АС – человек, АС – АС, человек – человек. Выполнение таких взаимосвязей многократно повторяется, что требует четкой регламентации всех выполняемых функций, разделения операций между человеком и машиной, введения инженерного термина – технологический процесс.

Любая система может рассматриваться с разных точек зрения – например, динамической, структурной, логической, физической. В результате мы получаем различные архитектурные представления как частные аспекты программной архитектуры, рассматривающие специфические свойства программной системы.

Рассматривая цели и ограничения в виде симметричных элементов логической схемы можно достаточно просто сформировать на их основе решение, которое является по существу выбором одной или нескольких из имеющихся альтернатив. При этом нечеткое решение может рассматриваться как некоторая «инструкция», нечеткость которой является следствием неточности формулировки поставленных целей и ограничений, т. е. влияние нечеткой цели G и нечеткого ограничения C на выбор альтернатив характеризуется их пересечением, которое образует нечеткое множество решений D. Принципиальное отличие между случайностью и нечеткостью заключается в том, что функция принадлежности всегда является гипотезой (о характере целей и имеющихся ограничениях), что позволяет строить оценки для альтернатив посредством формального аппарата. Данный процесс может быть описан в форме алгоритма, соответственно которому конкретные значения параметров какого-либо сложного свойства объединяются, а затем на этой основе делается вывод об их обобщенном значении. Главный структурный элемент данного алгоритма – правило «если... то...», отражающее свойство причинно-следственной связи. Такая система, опирается не непосредственно на материальные факты или объекты, а на сведения о них. Модель, охватывающая информационную систему, может быть представлена в виде метабазы, в которой содержится информация по каждому виду объекта учета. С другой стороны информационная система представима в виде функциональной системы – т. е. в виде множества функций. Таким образом цели и ограничения задаются как нечеткие множества. Взаимосвязь между ними может быть определена отношением на декартовом произведении [6].

Практически все системы в той или иной степени связаны с функциями долговременного хранения и обработки информации. База данных – это совокупность структурированных и взаимосвязанных данных и методов, обеспечивающих добавление, выборку и отображение данных, является моделью некоторой предметной области, состоящей из связанных между собой данных об объектах, их свойствах и характеристиках. Системы, предоставляющие средства работы с БД, называются СУБД. Не решая непосредственно никаких прикладных задач, СУБД является инструментом для разработки прикладных программ, использующих БД. Хранилища данных строятся на основе многомерной модели данных, что подразумевает выделение отдельных измерений и фактов, которые анализируются по выбранным измерениям. Многомерная модель данных физически может быть реализована как в многомерных СУБД, так и в реляционных. В последнем случае она выполняется по схеме «звезда» или «снежинка». Данные схемы предполагают выделение таблиц фактов и таблиц измерений. Каждая таблица фактов содержит детальные данные и внешние ключи на таблицы измерений. Чтобы сохранять данные согласно какой-либо модели предметной области, структура БД должна максимально соответствовать этой модели. Для реализации хранилищ данных (ХД) используют современные СУБД и концепцию ХД. Концептуально модель ХД можно представить в виде схемы, показанной на рис. 5.

Хранилища можно рассматривать как набор моментальных снимков состояния данных: можно восстановить картинку на любой момент времени. Атрибут времени всегда явно присутствует в структурах данных хранилища.

Попав однажды в хранилище, данные уже никогда не изменяются, а только пополняются новыми данными из оперативных систем, где данные постоянно меняются. Использование технологии хранилищ данных предполагает наличие в системе следующих компонентов:

- оперативных источников данных;
- средств переноса и трансформации данных;
- метаданных – включают каталог хранилища и правила преобразования данных при загрузке их из оперативных баз данных;
- реляционного хранилища;
- OLAP-хранилища;
- средств доступа и анализа данных.

Архитектура ХД может быть представлена схемой приведенной на рис. 6.

Для работы с хранилищем данных используются СУБД, к которым предъявляются специальные требования, которые включают в себя поддержку интегрированного многомерного анализа. Хранилища данных не измеряются, а дополняют традиционные реляционные базы данных с первичной информацией. Когда требуется не только извлечь информацию, оперируя неточными или нечеткими понятиями, а определенным образом расположить ее по убыванию степени соответствия запросу, может быть применен нечеткий поиск в хранилищах данных. Вариант архитектуры хранилища данных с поддержкой нечетких срезов может быть представлен схемой, приведенной на рис. 7. Результатом выполнения нечеткого среза, помимо самого подмножества ячеек гиперкуба, удовлетворяющих заданным условиям, является индекс соответствия срезу, который в свою очередь представляет итоговую степень принадлежности к нечетким множествам измерений и фактов, участвующих в сечении куба, и рассчитывается для каждой записи набора данных.

В основе концепции OLAP также лежит принцип многомерного представления данных. OLAP – это способ представления данных в простом и понятном для конечного пользователя виде. Назначение систем класса OLAP – предоставить пользователям гибкий, интуитивно понятный и простой доступ к данным. Данные представляются в виде многомерного куба, причем пользователь может быстро свернуть или развернуть данные по любому измерению. Для построения систем OLAP используются специализированные многомерные базы данных, либо надстройки над обычными реляционными базами данных.

OLAP-серверы, или серверы многомерных БД, могут хранить свои многомерные данные по-разному, так как наряду с детальными данными, извлекаемыми из оперативных систем, хранятся и агрегированные показатели, с единственной целью – ускорить выполнение запросов. В результате за скорость обработки запросов к суммарным данным приходится платить увеличением объемов данных и времени на их загрузку. Причем увеличение объема может стать буквально катастрофическим. Для решения проблемы хранения агрегатов применяются подчас сложные схемы, позволяющие при вычислении далеко не всех возможных агрегатов достигать значительного повышения производительности выполнения запросов. Многомерное хранение позволяет обращаться с данными как с многомерным массивом, благодаря чему обеспечиваются одинаково быстрые вычисления суммарных показателей и различные многомерные преобразования по любому из измерений. Мир информационной системы состоит из объектов, которые различаются по уровню сложности, и им отвечают соответствующие уровни связей. Сложность характеризуется располагаемой внутренней информацией, которая количественно зависит от числа входящих в него объектов и числа установленных между ними связей различных уровней. Разделив систему на информационные объекты (функциональные модули) и описав все их интерфейсы взаимодействия, можно декларировать относительную полноту множества учитываемых отношений между элементами системы, которые определяют ее поведение и являются предметом анализа функциональной стабильности.

Формальная модель информационной системы адекватно отображает физическую систему, иными словами, является информативной, если расхождение между ее реакцией на входные воздействия и соответствующими событиями физической системы находятся в допустимых пределах. Можно говорить, что формальная модель информационной системы представлена совокупностью баз данных, а так как любая таблица базы данных обладает фиксированным набором атрибутов, сама формальная модель также определяется конечным набором атрибутов.

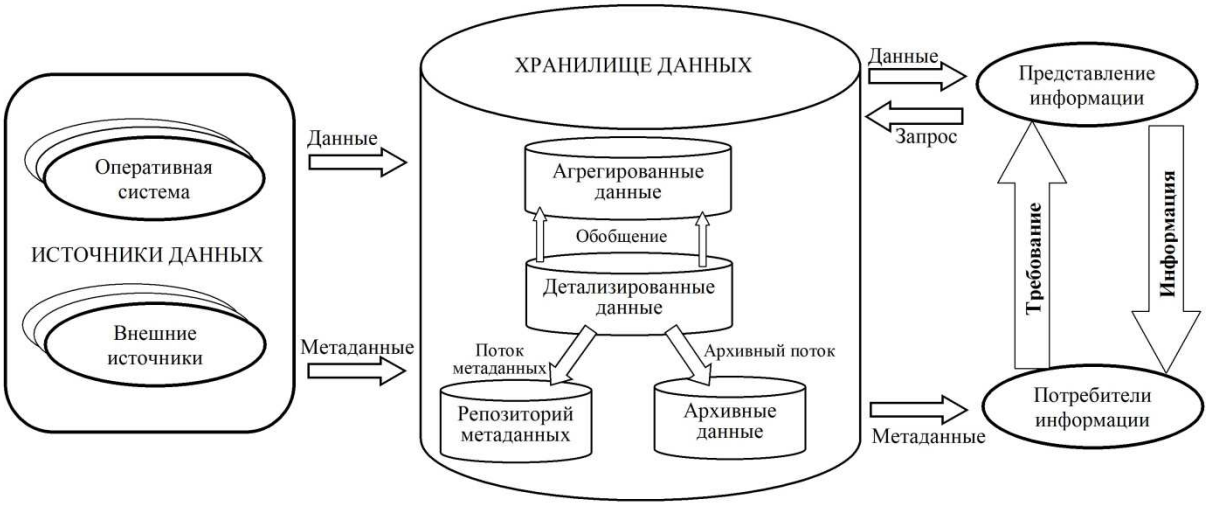

Рис. 5. Концептуальная модель хранилища данных

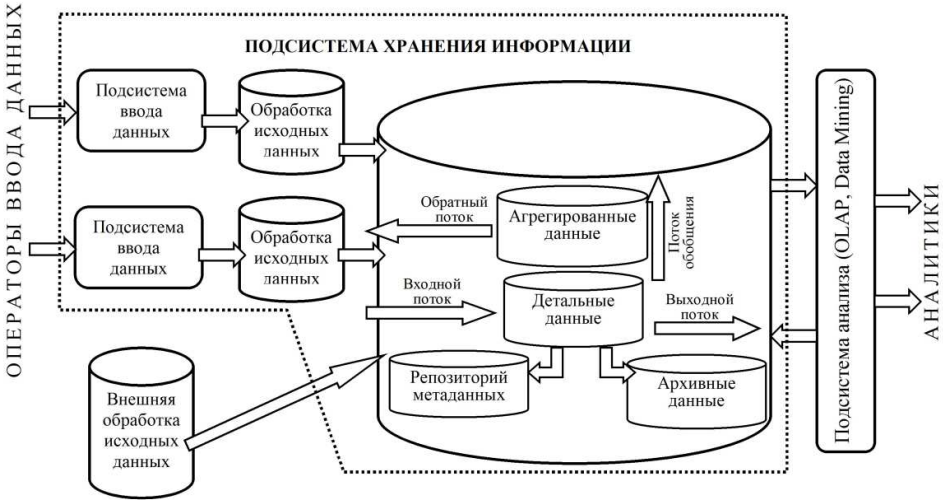

Рис. 6. Архитектура ХД

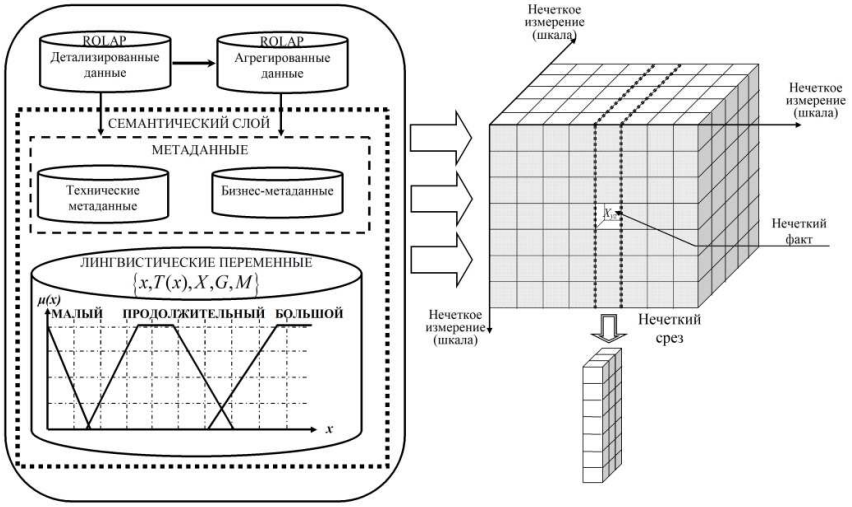

Рис. 7. Схема хранилища данных с поддержкой нечетких срезов

По мере уточнения знаний о физической системе уменьшается неопределенность формальной модели. В этом случае можно описать функцию, которая по известным параметрам объекта относит его к соответствующему отношению. В итоге система φ разбивается на совокупность подсистем φ_1, φ_2,…, φ_n, каждая из которых обладает меньшей неопределенностью, чем вся исходная.

В настоящее время широкое применение находит процессный подход для организации проектного и производственного менеджмента, основанный на формальных моделях жизненного цикла (ЖЦ) систем. Многие виды продукции представляют собой многосложные системы, основанные на взаимодействии совокупности управленческих и технических действий, в том числе технических средств, программного обеспечения и человеческого фактора. Их производство осуществляется с помощью процессов, имеющих разнообразные технические и управляющие «входы» и «выходы». При этом возрастает роль стандартов, используемых на всех стадиях менеджмента, прежде всего потому, что стандарты обеспечивают возможность взаимодействия различных компонент между собой. Можно определить независимые бизнес-процессы в качестве схем информационных объектов рассматриваемой модели. Очевидно, что правила взаимодействия экземпляров бизнес-процессов могут оцениваться двумя типами свойств: те, которые можно непосредственно измерить, и те, которые являются качественными и требуют по парного сравнения объектов, обладающих оцениваемым свойством, чтобы определить их место по отношению к рассматриваемому понятию. Модель, охватывающая информационную систему, может быть представлена в виде метабазы, в которой содержится информация по каждому виду объекта учета, в нашем случае о каждом экземпляре независимого бизнес-процесса (обозначим как множество G).

С другой стороны информационная система представима в виде функциональной системы – т. е. в виде множества функций (обозначим как множество F).

Введем обозначения:

Q_G – множество свойств, определяемых отношениями между элементами множества G.

Q_F – множество свойств, определяемых отношениями между элементами множества F.

Q_{FG} – множество свойств, определяемых связями между элементами множеств F и G.

Тогда взаимосвязи внутри множеств F и G могут быть определены отношениями на декартовом произведении

$$Q_F \times F = \begin{Bmatrix} z_i^F = \left(q_i^F, f_i\right) : q_i^F \in Q_F, \\ f_i \in F, i = 1,...,n \end{Bmatrix} \quad (1)$$

и

$$Q_G \times G = \begin{Bmatrix} z_i^G = \left(q_i^G, g_i\right) : q_i^G \in Q_G, \\ g_i \in G, i = 1,...,n \end{Bmatrix} \quad (2)$$

соответственно.

Взаимосвязь между ними может быть определена отношением на декартовом произведении

$$Q_{FG} \times G \times F = \begin{Bmatrix} z_i^{FG} = \left(q_i^{FG}, g_i, f_i\right) : q_i^{FG} \in Q_{FG}, \\ g_i \in G,\ f_i \in F, i = 1,...,n \end{Bmatrix} \quad (3)$$

Принадлежность элемента z_i^{FG} данному отношению может интерпретироваться следующим образом: «в элементе g_i информационной системы содержится информация по свойству q_i^{FG} функциональной части информационной системы f_i».

Поиск информации, соответствующей конкретному элементу f_i в g_i, сводится к определению отношения $R \subseteq G \times F$. Таким образом, о любой паре $(g_i, f_i) \in R$: $g_i \in G$, $f_i \in F$, $i = 1,\ldots, n$ можно сказать, что f_i является релевантным g_i, и решение задачи определения релевантности элементов множеств G и F, сводится к определению отношения $R \subseteq G \times F$. При этом $\forall g_i \in G$, $f_i \in F$, $g_j \in G$, $f_j \in F$, $i, j = 1,\ldots, n$ верно, что если $f_i \subseteq f_j$ и $g_i \subseteq g_j$, то есть все элементы g_i содержатся g_j и все элементы f_i содержатся в f_j и $(g_i, f_i) \in$ $\in R$, то выполняется $(g_j, f_j) \in R$. За исключением крайнего случая, когда отношение R есть само декартово произведение $G \times F$, отношение включает в себя не все возможные кортежи из декартового произведения. Это значит, что для каждого отношения имеется критерий, позволяющий определить, какие кортежи входят в отношение, а какие нет. Таким образом, каждому отношению R можно поставить в соответствие логическое выражение (предикат) Q_{FG}, зависящее от определенного числа параметров (n-местный предикат) и определяющее, будет ли кортеж (g_j, f_j) принадлежать отношению R. Таким образом, принадлежность кортежа отношению эквивалентна истинности предиката:

$$(g_j, f_j) \in R \Leftrightarrow \{Q_{FG}\} = \{G, F, R\} \quad (4)$$

Однако в любом случае, при информационном подходе для формализации предметной области, первичными являются категории объектов и отношения между ними, т. е. формально система отношений может быть представлена множеством объектов предметной области и множеством отношений между ними. Характеристику процесса можно представить в виде множества пар $\{<A_i, D_i>, i = 1,\ldots,n\}$, где A_i – непустое множество имен свойств (атрибутов), D_i – множество значений соответствующих атрибутов. Значения разбиваются на классы объектов, которые взаимодействуют друг с другом на основе правил. Пусть π – множество этих правил. На множестве атрибутов могут быть установлены отношения $G = \{\bar{G}, \tilde{G}\}$, которые делятся на количественные $\bar{G}$ и качественные $\tilde{G}$, для которых определено множество типов оценки, например T = {«проекты продвигаются в направлении достижения поставленных целей», «проекты ведутся согласно соответствующим директивам», «проекты реализуются в соответствии с планами», «проекты остаются жизнеспособными»}. Тогда любое правило оценки может быть представлено кортежем $\pi = <G, T>$.

Таким образом, совокупность информационных характеристик процесса $\{<A_i, D_i>, i = 1,\ldots,n\}$, установленных отношений $G = \{\bar{G}, \tilde{G}\}$ и правил установления отношений $\pi = <G, T>$ может быть использовано для формального определения процесса в виде следующего кортежа компонентов:

$$Z = \left\{\langle A_i, D_i\rangle, \{\bar{G}, \tilde{G}\}, \left\langle \{\bar{G}, \tilde{G}\}, T\right\rangle\right\}, i \in N \quad (5)$$

Значения атрибутов могут носить не числовой характер. В частности, в макроэкономических, социологических, маркетинговых, медицинских, правовых хранилищах данных широко используется лингвистическая форма представления данных. Для оценки характеристик, которые носят качественный характер, могут быть использованы порядковые шкалы, бальные элементы которых соответствуют градациям вербальных шкал. Уровням порядковых шкал можем поставить в соответствие значения лингвистических переменных и все дальнейшие операции производить с их функциями принадлежности. При этом их адекватность не может быть проверена средствами теории и в каждом существующем в настоящее время методе построения функции принадлежности формулируются свои требования и обоснования к выбору именно такого построения.

Рассмотрим N объектов, у которых оценивается интенсивность проявления характеристик атрибутов с наименованиями A_j, $j = 1,\ldots, k$, значения которых X_j, $j = 1,\ldots, k$ используются для оценки качественной характеристики Y. С позиции аппарата теории нечетких множеств моделями экспертного оценивания признаков служат полные ортогональные семантические пространства [9], где в качестве функций принадлежности используются нормальные треугольные числа и T-числа. Данный метод позволяет перейти от разноплановой качественной информации к единой абстрактной величине – значению функции принадлежности. Используем метод работы, с нечеткой информацией, рассмотренный в [9].

Пусть X_{ij}, $i = 1,\ldots, m_j$ – уровни вербальных шкал, применяемые для оценивания атрибутов с наименованиями A_j, $j = 1,\ldots, k$ и расположенные в порядке возрастания интенсивности их проявления.

Обозначим через a_i^j, $i = 1,\ldots,m_j$, $j = 1,\ldots,k$ – относительные числа объектов, отнесенных при оценивании атрибутов с наименованиями A_j, $j = 1,\ldots, k$ к уровню X_{ij}, $i = 1,\ldots, m_j$, $j = 1,\ldots, k$, $\sum_{i=1}^{m} a_i^j = 1$, $j = 1,\ldots,k$.

Можем построить k наборов нечетких чисел, соответствующих атрибутам с наименованиями A_j, $j = 1,\ldots, k$. Обозначим через $\mu_{ij}(x)$ функцию принадлежности нечеткого числа $\tilde{X}_{ij}$, соответствующего i-му уровню j-го атрибута $i = 1,\ldots, m_j$, $j = 1,\ldots, k$. В качестве оценки объекта могут быть взяты нечеткие числа $\tilde{X}_{ij}$, $i = 1,\ldots, m_j$, $j = 1,\ldots, k$ с соответствующими им функциями принадлежности $\mu_{ij}(x)$, $i = 1,\ldots, m_j$, $j = 1,\ldots, k$. Тогда оценка n-го объекта по атрибуту X_j, $j = 1,\ldots, k$ может быть представлена функцией принадлежности μ_j^n. Используя так называемую L-R функцию принадлежности, оценку каждого объекта можем записать в виде

$$\left\{\mu_j^n(x) = \left(a_{j1}^n, a_{j2}^n, a_{jL}^n, a_{jR}^n\right)\right\},$$
$$n = 1,\ldots, N,\ j = 1,\ldots, k$$

Конкретный вид функций принадлежности определяется с учетом специфики имеющейся неопределенности, реальной ситуации на объекте и числа степеней свободы в функциональной зависимости. Одним из наиболее удобных для

описания функции принадлежности являются монотонно возрастающие или убывающие сигмоиды, они удобны для задания лингвистических термов естественного языка, уравнения которых имеют вид: $\mu(x)=\frac{1}{1+e^{-a(x-b)}}$ для растущей функции и $\mu(x)=\frac{1}{1-(1+e^{-a(x-b)})}$ для убывающей, где a – крутизна сигмоиды, b – позволяет сдвигать точку центра по оси. При $a = 0$ сигмоида вырождается в прямую линию на отметке 0,5.

Обозначим весовые коэффициенты оцениваемых атрибутов через ω_j, $j = 1,\ldots, k$, $\sum_{j=1}^{n}\omega_j=1$. Общая нечеткая оценка n-го объекта ($n = 1,\ldots, N$) может быть определена суммой оценок по всем атрибутам с учетом весовых коэффициентов $\tilde{A}^n=\omega_1\otimes\tilde{X}^n_{11}\oplus\ldots\oplus\omega_k\otimes\tilde{X}^n_{m_k k}$ и функцией принадлежности

$$\mu_n(x)=\begin{pmatrix}\sum_{j=1}^{k}\omega_j a^n_{j1},\sum_{j=1}^{k}\omega_j a^n_{j2},\\ \sum_{j=1}^{k}\omega_j a^n_{jL},\sum_{j=1}^{k}\omega_j a^n_{jR}\end{pmatrix},\ n=1,\ldots,N. \qquad (6)$$

Нечеткое число $\tilde{B}_1$, соответствующее наименьшей интенсивности, определяется $\tilde{B}_1=\omega_1\otimes\tilde{X}^n_{11}\oplus\ldots\oplus\omega_k\otimes\tilde{X}^n_{1k}$, а $\tilde{B}_m$, соответствующее наибольшей интенсивности $\tilde{B}_m=\omega_1\otimes\tilde{X}_{m_1 1}\oplus\ldots\oplus\omega_k\otimes\tilde{X}_{m_k k}$. Полученные нечеткие числа могут быть дефазифицированы, например по методу центра тяжести, полученные четкие числа обозначим соответственно Z_n, $n = 1,\ldots, N$, B_1, B_m. Эти числа можем использовать в качестве нормированной рейтинговой оценки n-го объекта, $n = 1,\ldots, N$, по формуле $E_n=\frac{Z_n-B_1}{B_m-B_1}$, $n = 1,\ldots, N$. В общем случае характеристика каждого объекта x_i может быть описана соответствующей лингвистической переменной $<A_j, T_j, D_j>$, где $T_j=\left\{T_1^j,T_2^j,\ldots,T_{m_j}^j\right\}$ – терм-множество лингвистической переменной A_j (набор лингвистических значений атрибута), m_j – число значений атрибута; D_j – (предметная шкала) базовое множество атрибута A_j. Для описания термов T_k^j, $k = 1,\ldots, m_j$ соответствующих значениям атрибута A_j, могут быть использованы нечеткие переменные $\left\langle T_k^j, D_j, \tilde{C}_k^j\right\rangle$, т. е. значение T_k^j – описывается нечетким множеством $\tilde{C}_k^j$ в базовом множестве D_j:

$$\tilde{C}_k^j=\left\{\left\langle\mu_{C_k^j}(d)\,|\,d\right\rangle\right\},\quad d\in D_j,\ k=1,\ldots,m_j. \qquad (7)$$

Тогда в качестве нечеткой характеристики объекта x_i может быть взято нечеткое множество второго уровня

$$\tilde{x}_i=\left\{\left\langle\mu_{x_i}(a_j)\,|\,a_j\right\rangle\right\},$$
$$\mu_{x_i}(a_j)=\bigcup_{k=1}^{m_j}\left\{\left\langle\mu_{\mu_{x_i}}(T_k^j)\,|\,T_k^j\right\rangle\right\},\ T_k^j\in T_j,\ a_j\in A_i. \qquad (8)$$

Все это позволяет перейти от разноплановой качественной информации к одной абстрактной величине – значению функции принадлежности. Для оценивания качественных характеристик объектов используем следующее представление их элементов [7, 9]:

- функция принадлежности для крайнего терм-множества, соответствующего минимальной интенсивности проявления признака может быть представлена в виде

$$\mu_{x_1}(x)=\begin{cases}1,\ 0\le x\le a_1-\frac{1}{2}\min(a_1,a_2)\\ 1-\frac{x-(a_1-\frac{1}{2}\min(a_1,a_2))}{\min(a_1,a_2)},\\ a_1-\frac{1}{2}\min(a_1,a_2)<x\le a_1+\frac{1}{2}\min(a_1,a_2)\\ 0,\ a_1+\frac{1}{2}\min(a_1,a_2)<x\le 1\end{cases}$$

- функция принадлежности для крайнего терм-множества, соответствующего максимальной интенсивности проявления признака может быть представлена в виде

$$\mu_{x_m}(x)=\begin{cases}0,\ 0\le x\le 1-a_m-\frac{1}{2}\min(a_{m-1},a_m)\\ 1+\frac{x-(1-a_m-\frac{1}{2}\min(a_{m-1},a_m))}{\min(a_{m-1},a_m)},\\ 1-a_m-\frac{1}{2}\min(a_{m-1},a_m)<x\le\\ \quad\le 1-a_m+\frac{1}{2}\min(a_{m-1},a_m)\\ 1,\ 1-a_m+\frac{1}{2}\min(a_{m-1},a_m)<x\le 1\end{cases}$$

- функция принадлежности для средних терм-множеств качественного признака может быть представлена в виде

$$\mu_{x_j}(x)=\begin{cases}0,\ 0\le x\le \sum\limits_{i=1}^{l-1}a_i-\frac{1}{2}\min(a_{l-2},a_{l-1},a_l)\\ 1+\dfrac{x-(\sum\limits_{i=1}^{l-1}a_i-\frac{1}{2}\min(a_{l-2},a_{l-1},a_l))}{\min(a_{l-2},a_{l-1},a_l)},\\ \sum\limits_{i=1}^{l-1}a_i-\frac{1}{2}\min(a_{l-2},a_{l-1},a_l)<x\le\\ \qquad\le\sum\limits_{i=1}^{l-1}a_i+\frac{1}{2}\min(a_{l-2},a_{l-1},a_l)\\ 1,\ \sum\limits_{i=1}^{l-1}a_i-\frac{1}{2}\min(a_{l-2},a_{l-1},a_l)<x\le\\ \qquad\le\sum\limits_{i=1}^{l}a_i+\frac{1}{2}\min(a_{l-1},a_l,a_{l+1})\\ 1+\dfrac{x-(\sum\limits_{i=1}^{l}a_i-\frac{1}{2}\min(a_{l-1},a_l,a_{l+1}))}{\min(a_{l-1},a_l,a_{l+1})},\\ \sum\limits_{i=1}^{l}a_i-\frac{1}{2}\min(a_{l-1},a_l,a_{l+1})<x\le\\ \qquad\le\sum\limits_{i=1}^{l}a_i+\frac{1}{2}\min(a_{l-1},a_l,a_{l+1})\\ 0,\ \sum\limits_{i=1}^{l}a_i+\frac{1}{2}\min(a_{l-1},a_l,a_{l+1})<x\le 1.\end{cases}$$

Исходя из приведенного, предметную область можно представить в виде многоуровневой среды, состоящей из множества элементов предметной области, множества функций и методов, работающих на этих элементах и множества свойств элементов и отношений между элементами, т. е. в виде онтологии, которая включает в себя описание свойств предметной области и взаимодействия объектов на некотором формальном языке, имеющем логическую семантику. Если система сложная, число факторов велико, то учет всех ее характеристик (компонент) приводит к чрезвычайной сложности. Поэтому в модель приходится вводить лишь ограниченное число, а оставшиеся компоненты учитывать, явно не вводя в модель, но учитывая их влияние как нечеткую реакцию модели на тот или иной выбор альтернативы. Очевидно, что алгебраическое сравнение компонент невозможно и может быть выполнено с применением методов нечеткой логики.

Базируясь на предложенном методе формирования структуры информационной системы в виде совокупности взаимодействующих семантически определенных и формализованных объектов, связанных друг с другом иерархическими отношениями классов атрибутов определяющих их бизнес-процессов, наиболее целесообразно использовать в качестве измерений при построении куба множеств, полученных согласно правил перекрытия взаимодействующих объектов учета. Таблица фактов при этом будет содержать целочисленные колонки, дающие числовую характеристику каждого, определенного таким образом измерения, и несколько целочисленных колонок – ключей для доступа к таблицам измерений, которые их расшифровывают.

Для каждого измерения составляем список уникальных значений из элементов, хранящихся в столбцах и производим предварительное агрегирование фактов для записей, имеющих одинаковые значения размерностей. Используя промежуточные таблицы (так называемые кросс-таблицы) можно связать элементы разных таблиц между собой, для чего каждой записи в таблицах измерений поставить в соответствие список, элементами которого будут номера фактов, при формировании которых использовались эти измерения. Для фактов соответственно каждой записи поставим в соответствие значения координат, по которым она расположена в гиперкубе. Измерения имеют иерархическую структуру, состоящую из одного или нескольких уровней, на основании которой осуществляются операции свертки или детализации.

Для каждого свободного бизнес-процесса может создаваться отдельная OLAP-таблица, логическое связывание которых и построение OLAP-таблиц следующего уровня возможно уже только на уровне агрегированных аналитических показателей, приведенных к сопоставимой форме.

Речь в данном случае идет о переходе к управлению данными, которые распределены по многим репозиториям с одновременным обеспечением базового набора функций над всеми источниками данных – т. е. к управлению пространствами данных. В некоторых случаях границы между пространствами данных могут быть плавающими, поэтому все виды связей между участниками должны быть формализованы, а семантическая интеграция развиваться во времени только там, где требуется. При этом

обеспечивается высокий уровень автономности компонентов.

В результате можем говорить об операциях над конечномерными векторными пространствами, представленными многомерными кубами, где в качестве базисов используются вектора измерений. Композиция таких кубов может быть представлена в виде линейной комбинации базисных векторов – тензорном произведении, результатом которого является наиболее общее пространство, в которое можно билинейно отобразить исходные пространства.

Рассмотрим вариант построения измерений гиперкуба, для чего произведем необходимые преобразования данных, хранящихся в таблицах базы данных. Так, в целях повышения производительности при построении гиперкуба найдем уникальные элементы, хранящиеся в столбцах, которые будут являться измерениями гиперкуба. Для записей, имеющих одинаковые значения размерностей, произведем предварительное агрегирование. Как уже было сказано выше, для нас важны уникальные значения, имеющиеся в полях измерений. Для построения срезов гиперкуба нам необходимы следующие возможности – определение координат (фактически значения измерений) для записей таблицы, а также определение записей, имеющих конкретные значения. Действия разбиты на два этапа – согласно приведенной классификации атрибутов. Для формализованных атрибутов можно сразу составить таблицу уникальных значений, на основании которой будет производиться построение измерения. В зависимости от типа подлежащих интеграции бизнес-процессов, определяем правило интеграции. Определяем ключевые слова (фразы), составляем таблицу совокупности вариантов их комбинаций, удовлетворяющих семантическому правилу (по функции принадлежности), и, как и в случае с формализованными атрибутами, строим таблицу уникальных значений (рис. 8). То есть вместо одной таблицы мы получили аналог нормализованной базы данных. Нас интересуют только координаты в нашем гиперкубе, поэтому определим координаты для значений измерений. Самым простым будет перенумеровать значения элементов. Для того чтобы в пределах одного измерения нумерация была однозначной, предварительно отсортируем списки значений измерений (словари, выражаясь терминами БД) в алфавитном порядке.

На практике данная концепция находит свое применение в Web-ориентированной среде, к которой предъявляется требование высокой производительности при обработке больших объемов данных, например Web-портале (кафедры), который, в свою очередь, является совокупностью связанных в единое целое системой каналов передачи информации баз данных, информационных хранилищ, баз знаний, а также информационных технологий, которые поддерживают процессы обработки, анализа и передачи информации различного уровня интеграции.

При этом при решении задач одной системы используются знания других систем. Аналогично одни и те же объекты в разных системах могут быть описаны различными свойствами и, соответственно, иметь отличную структуру. Одной из проблем является определение способа интеграции знаний различных разделов внутри одного Web-портала. Основной операцией для большего числа пользователей является операция поиска и получения информации. Причем сами данные, однажды сформированные, уже не подвергаются модификации. Динамические базы данных в такой ситуации являются неэффективными, и в мировой практике на сегодняшний день для решения подобных задач осуществляется переход к информационным хранилищам. Пусть D_j – коллекция документов, M_j – множество запросов, $d_j : M_j \to 2^{D_j}$ – отображение, сопоставляющее каждому запросу множество документов. Каждая информационная подсистема портала может быть представлена кортежем $S_j = (D_j, M_j, d_j)$, где $j = 1,\ldots, m$ – информационные подсистемы, использующиеся в портале. Тогда информационная система портала может быть представлена в виде распределенной информационной системы, базирующейся на глобальном тезаурусе T и определяться кортежем $S = (T, D, M, d)$ через локальные составляющие $D = \bigcup_j D_j$ и $d = \bigcup_j d_j$. Любую базу данных можно рассматривать как объектный архив, предназначенный для хранения атрибутов: пассивных объектов, не имеющих моделей состояний, активных объектов, подчиненных жизненному циклу.

Любую базу данных можно рассматривать как объектный архив, предназначенный для хранения атрибутов: пассивных объектов, не имеющих моделей состояний, активных объектов, подчиненных жизненному циклу.

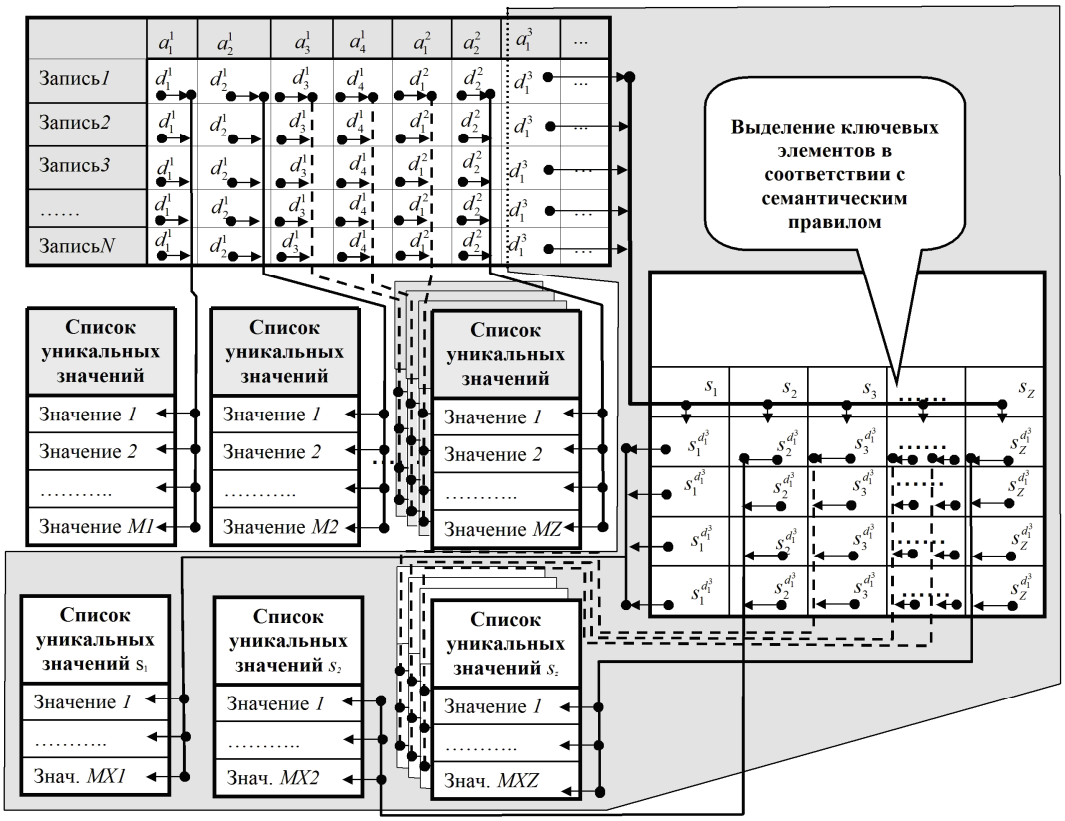

Рис. 8. Пример формирования OLAP куба

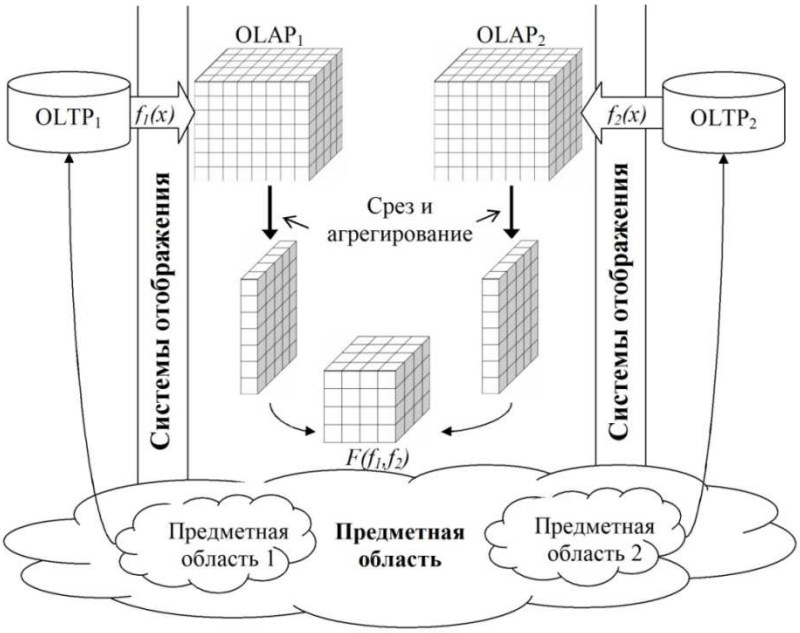

Рис. 9. Построение нового куба на основе отношений

Присутствует разнотемповость изменения атрибутов. Например, идентифицирующие данные о студенте не изменяются, а ряд специфических атрибутов, таких как факультет, группа, текущая успеваемость, являются динамическими и модифицируются. При окончании обучения данные становятся постоянными и никогда не могут быть изменены.

Таким образом, всю рассмотренную информацию целесообразно разделить на различные

уровни, как по времени существования, так и по доступу, а развитие базы данных заключается в добавлении и удалении кортежей, соответствующих экземплярам объекта. В этом случае база данных автоматически переходит в разряд информационных хранилищ полностью при выполнении следующих условий: периодичность удаления кортежей из хранилища соизмерима со временем его существования, а OLAP-куб дополняется новым измерением – фрагментированным временем (моментом переноса информации из базы данных в хранилище, в соответствии с произошедшим изменением значения атрибутов).

Если же информация в динамических базах данных после модификации не представляет интереса ни с одной из точек зрения, то организация информационного хранилища на основе таких баз не имеет смысла. Сами OLAP-кубы, соответствующие OLTP-системам, можно представить в виде многомерных информационных объектов, обладающих соответствующими свойствами.

Так как каждая OLTP-система, являясь реализацией модели предметной области, отображается в многомерную матрицу (OLAP-куб), существуют функции этого отображения. Рассматривая предметные области как часть расширенной предметной области, можно сделать вывод о возможности установления функции отношения между функциями отображения, и, как следствие, построения новой системы или OLAP-куба на основе этого отношения. Необходимо отметить, что эффективность Web-портала может быть существенно улучшена, если при моделировании более точно учитывать особенности шкал измерения.

СПИСОК ЛИТЕРАТУРЫ

1. Поддержка принятия решений в слабоструктурированных предметных областях: анализ ситуаций и оценка альтернатив / А. Н. Аверкин [и др.] // М.: Теория и системы управления. Известия РАН, № 3. 2006. С. 139–149.

2. **Беллман Р., Заде Л.** Принятие решений в расплывчатых условиях. М.: Мир, 1976. С. 172–215.

3. **Волкова В. Н., Денисов Ф. Ф.** Основы теории систем и системного анализа. СПб.: СПбГТУ, 2001. 512 с.

4. **Буч Г.** Объектно-ориентированный анализ и проектирование с примерами приложений. М.: Вильямс, 2008.

5. **Заболеева-Зотова А. В., Камаев В. А.** Лингвистическое обеспечение автоматизированных систем. М.: Высшая школа, 2008. 244 с.

6. Совместное использование учетных систем и технологии OLAP [Электронный ресурс] (citforum.oldbank.com/database/articles/olap_oltp.html).

7. **Антонов В. В., Куликов Г. Г., Антонов Д. В.** Теоретические и прикладные аспекты построения моделей информационных систем. LAP LAMBERT Academic Publishing GmbH & Co.KG, Germany, 2011. 134 с.

8. **Артемьев С. А., Замай С. С., Питенко А. А.** ГИС конструктор со средствами анализа данных для создания информационно-аналитических систем // Вестник КазНУ. 2004. Т. 9, № 3(42). С. 188–192.

9. **Полещук О. М., Полещук И. А.** Нечеткая кластеризация элементов множества полных ортогональных семантических пространств // Вестник Московск. гос. ун-та леса. 2003. № 1 (26). С. 117–127.

ОБ АВТОРАХ

Антонов Вячеслав Викторович, доц. каф. автоматизированных систем управления. Дипл. математик (БГУ, 1979). Канд. техн. наук по управл. в соц. и экон. системах (УГАТУ, 2007). Иссл. в обл. автоматиз. информ. систем.

Куликов Геннадий Григорьевич, проф., зав. той же каф. Дипл. инженер по автоматиз. машиностроения (УАИ, 1971). Д-р техн. наук по системн. анализу, автоматич. упр. и тепловым двигателям (УАИ, 1989). Иссл. в обл. АСУ и упр. силовыми установками ЛА.

Антонов Дмитрий Вячеславович, асп. той же каф. Дипл. инженер по автоматизированным системам обработки информации и управления (УГАТУ, 2010). Готовит дис. в обл. построения информационных систем.