УДК 004.62:004.942

О.В. *Летуноеа, Е.В. Малахов,* ***Т.В. Филатова***
Одесский национальный политехнический университет, 65044, г. Одесса, проспект Шевченко, 1

МАНИПУЛИРОВАНИЕ ИНФОРМАЦИОННЫМИ ХРАНИЛИЩАМИ С ПЕРЕМЕННЫМ ВЕКТОРОМ АТРИБУТОВ

В настоящее время объемы информации, накапливаемые в базах данных и информационных хранилищах настолько велики, что сложность задачи анализа этой информации, ее выборки, поиска и обработки значительно возросла. Во многих случаях объекты реального мира являются «многослойными». Каждый такой слой описывает конкретный аспект деятельности или функционирования (существования) объекта. Информация о таких объектах обычно хранится в многомерных структурах данных (многомерных хранилищах). При анализе в системах поддержки принятия решения над этой информацией выполняется ряд уже стандартизованных операций. Это возможно, если «слои» объекта относятся к одной предметной области и имеют одну «размерность». Если каждый информационный «слой» описывает рассматриваемый объект в отдельной предметной области, и при этом даже количество свойств объекта в каждой предметной области отличается от других, то возникают задачи построения единой структуры для этого объекта, проверки ее адекватности и манипулирования хранимой информации. Соответственно необходимо разработать подходы к решению этих задач.
Ключевые слова *- Атрибуты - Переменный вектор атрибутов - Многомерное представление данных — Гиперкуб*

Now volumes of the information accrued in data bases and information storehouses are so great, that the complexity of a task of the analysis of this information, its sample, search and processing considerably has increased. In many cases the objects of the real -world are "«multilayer". Each such layer describes concrete aspect of activity or^functioning (existence) of object. The information on such objects is usually stored in mulivariete structures given (mulivariete storehouses). At the analysis in systems of support of acceptance of the decision above this information a number already of standardized operations is carried out. It is possible, if "«layers" of object concern to one subject domain and have one "«dimension". If each information "«layer" describes considered(examined) object in a separate subject domain, and thus even the quantity(amount) of properties of object in each subject domain differs from others, there are tasks of construction of uniform structure for this object, check of its adequacy and manipulation of the stored) information. Accordingly it is necessary to develop the approaches to the decision of these tasks.
Keywords *- Attributes - Variable vector of attributes - Multivariate data presentation - ffypercube*

I. ВВЕДЕНИЕ

Исследованиями в области многомерных структур данных (их представления, хранения и обработки) на данный момент занимаются многие передовые исследовательские организации. Достаточно большое количество материала посвящено исследованиям в области представления и хранения многомерных структур данных, и в этих направлениях уже определены критерии оценки и стандартизации. Однако возможности обработки обработки многомерных структур данных и посвящена представленная статья авторов.

На данном этапе развития передовых компьютерных технологий во всем мире проводятся более подробные исследования различных областей деятельности человека, причем практически все операции компьютеризированы и выполняются автоматически. Обязанностью человека является лишь отслеживание происходящих процессов, проведение анализа полученных результатов, выведение закономерностей, построение прогнозов и

обеспечения на данном этапе направлены на создание таких программных продуктов, которые могли бы хранить и обрабатывать весь необходимый объем информации (сколь угодно большой), при этом информация должна быть корректной, правильно храниться, чтобы быть правильно обработанной и верно истолкованной. Для осуществления этих мероприятий и применяются многомерные структуры данных, а именно гиперкубы.

Данные структуры позволяют производить анализ большого объема информации любого вида, получать данные различной степени детализации о деятельности всей организации в целом и, в частности, каждого ее сотрудника и производить анализ полученных данных.

II. ОСНОВНАЯ ЧАСТЬ

Для получения более детального представления о внутреннем строении и содержании многомерных структур рассмотрим математическое представление гиперкуба.

Пусть существует в рассматриваемой предметной области некоторый объект (событие) б, представленный на рисунке 1 в трехмерном виде, где измерения 1, 2 и 3 соответствуют названиям полей таблицы, а численные показатели измерений - данным, содержащимся в строках таблицы.

Рис. 1 — Трехмерное представление события (объекта) О

где $A_t = \{a_n, a_{i2>} ..., o«. a_{iK}.,, ..., a_{in}\}$ - вектор атрибутов;

*он, <*a, —, a*.* а«+/. —. я/я - набор параметров, который характеризуют объект (событие) с определенной точки зрения.

Векторы атрибутов Л, могут быть разной размерности для одного и того же объекта (события), то есть иметь различное число параметров.

Пусть *Оц,* я/7, —, я/*- постоянный набор параметров, который остается неизменным в течение всего времени существования объекта (события). Выделим этот набор параметров всех векторов атрибутов в отдельный объект *U'* (рисунок 2), который будет являться постоянным объектом.

Пусть $a_{iK}4i, ..., a_{ln}$ - непостоянный набор параметров, который уникален для каждого вектора атрибутов. Выделим этот набор параметров всех векторов атрибутов в отдельный объект *О"* (рисунок 3), который будет иметь неоднородную структуру.

Рис, 2 - Трехмерное представление события (объекта) &

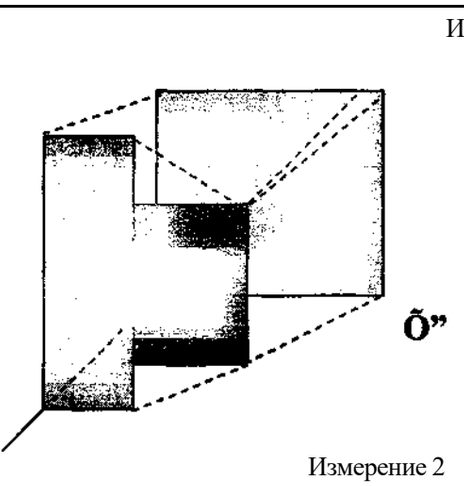


Рис. 3 ~ Трехмерное представление события (объекта) О"

Итак, было выделено два объекта *б'* и *О"* из первоначального *О,* причем параметры *О'* характеризуют неизменные параметры объекта *б.* Отсюда можно сделать вывод, что операции обработки будут производиться над объектом *б",* так как этот объект содержит уникальные значения каждой из плоскостей.

Однако, так как операции производятся над некоторым определенным набором данных, то есть вектором атрибутов, то необходимо ввести переменную величину, которая бы четко определяла необходимый набор параметров. Пусть такой величиной будет переменный вектор атрибутов *А= = 1,* .... I, величина которого равна номеру того набора параметров, то есть вектора атрибутов, который однозначно характеризует данный объект (событие) с определенной точки зрения.

Применение вектора атрибутов направлено на то, чтобы свести к минимуму получение пустых ответов на запросы пользователя. Технология применения вектора атрибутов состоит в следующем: задавая его точные параметры, тем самым задают измерения строящегося гиперкуба, что полностью не избавляет от образования пустот, однако, значительно уменьшает их объем, что, в свою очередь, положительно сказывается на объеме занимаемой гиперкубом информации и на быстродействии получения информации на запрос пользователя.

Рассмотрим следующий пример. Построим гиперкуб на основе события *б"* (рисунок 4). Из рисунка видно, что созданный гиперкуб, в данном случае трехмерный, содержит множество пустых значений. Например, если организовать запрос на получение информации, соответствующей точке А, то ответом будет пустое значение. Если задать запрос, соответствующий выбору из гиперкуба данных в точке В, то полученный ответ будет содержать информацию, содержащуюся в двух плоскостях, так как в первой плоскости нет данных, относящихся к данному запросу. При организации запроса, соответствующего выбору информации в точке С, получим данные, соответствующие только первой плоскости.

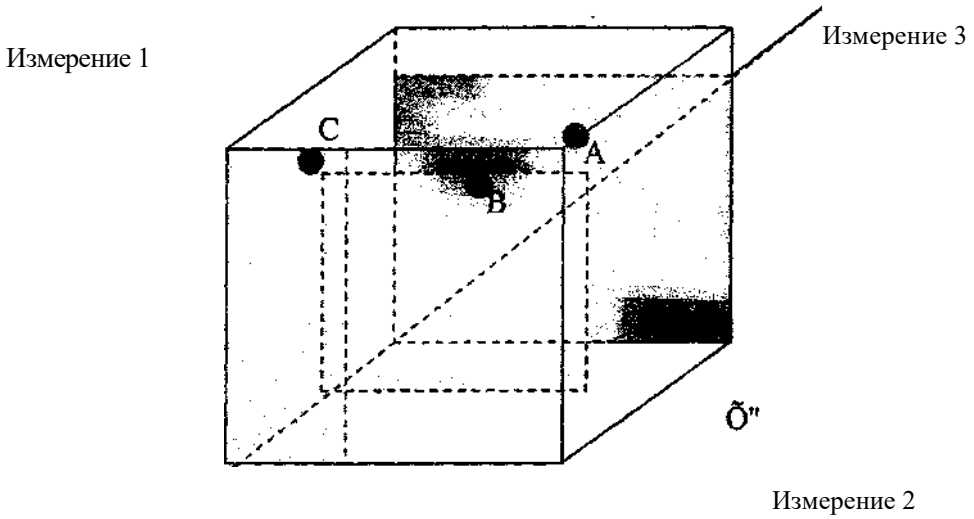
Применение приведенной математической модели возможно к любой предметной области, то есть к любым объектам реального мира — предприятиям, учреждениям, организациям и т.д. В данной статье все предлагаемые изыски будут рассмотрены на примере ВУЗа.

Для извлечения необходимой информации при анализе предметной области над многомерными структурами совершают определенные операции.

Существуют следующие виды операций над гиперкубами:

- сечение;
- вращение;
- консолидация;
- операция спуска;
- разбиение с поворотом [1].

Работа с гиперкубом сводится к различным его поворотам, группировкам и т.д. Можно менять количество измерений, способы группировки, но при этом необходимо учитывать, что гиперкуб очень быстро увеличивается в размерах, хотя при таком представлении данных работать с информацией легко и удобно. Поэтому для того, чтобы получить хороший результат, необходимо, чтобы на экран выводился не весь гиперкуб, а только нужная для текущего анализа часть. Для этого необходимо, во-первых, иметь возможность выбирать только те измерения, которые необходимы для анализа: если не имеет значения домашний адрес студента, а важен его средний балл, то нужно убрать измерение «домашний адрес». Во-вторых, иметь возможность выбрать/отсечь ненужные значения. Например, если из всех специальностей интересна «Экономическая кибернетика», то нужно строить куб только для этой специальности.



Рис, 4 — Трехмерное представление события (объекта)0"

Сложности и неопределенности возникают при необходимости использования переменного вектора атрибутов в проведении операций над данными гиперкуба.

Переменный вектор атрибутов - это определенный набор параметров, который описывает объект с определенной точки зрения и является неодинаковым для разных объектов, то есть имеет разную размерность. Как было отмечено ранее, гиперкуб имеет сильно разреженную структуру (содержит много пустот), причем, чем выше степень агрегирования (детализирования) измерений, тем больше разреженность данных и тем больше места занимает сам гиперкуб (рисунок 4).

Для большей наглядности и лучшего понимания использования переменного вектора атрибутов рассмотрим следующий пример.

Пусть имеются данные о двух группах ОИ-001 и ОИ-002, которые представлены в виде гиперкуба, имеющего следующие измерения: семестры (с 2000 по 2005 годы), ФИО студентов, предметы, баллы. Итак, получаем гиперкуб с четырьмя измерениями. Из полученного гиперкуба необходимо извлечь данные о студентах-магистрах.

Рассмотрим несколько способов решения поставленной задачи.

Способ 1. Соединение всех атрибутов в едином отношении.

Такое решение практикуется в подавляющем большинстве случаев. Однако в рассматриваемом случае оно неприемлемо, так как мы не можем объявить обязательными те атрибуты, которые являются характерными только для отдельных видов фактов, хотя это, безусловно, необходимо. Это может привести к нежелательным последствиям. Попробуем, наоборот, пусть атрибуты станут обязательными. Тогда потребуется ввод показателей атрибутов, которые характерны для одного факта, но абсолютно не свойственны и бессмысленны для другого (только определенные студенты являются магистрами). Возникает ситуация, когда пользователи в ответ на бессмысленный запрос могут ввести не менее бессмысленный ответ.

Кроме того, отношение, объединяющее несколько фактов по общему свойству (например, отношение «группа» объединяет разных студентов, обладающих каждый своим набором атрибутов, свойственных только ему, хотя имеются и общие показатели), будет обладать тенденцией к бесконечному росту при увеличении свершившихся фактов. Однако только незначительная часть атрибутов будет полезной в каждом конкретном случае.

Способ 2. Группировка общих атрибутов в едином отношении.

Такой способ решения позволяет избежать тех проблем,' которые свойственны первому варианту. На самом деле, поскольку атрибуты свойственные каждой конкретной сущности разнесены по разным отношениям, то на них можно накладывать любые ограничения и проверки. С другой стороны, не происходит бесконтрольного увеличения общего отношения, так как там хранятся только те атрибуты, которые являются общими для всех видов товаров. Но является ли это решение хорошим?

Одно из незыблемых правил реляционной теории гласит, что поле (группа полей) одного отношения не может быть другим отношением или ссылкой (прямой или косвенной) на другое отношение [4]. Нам же придется в отношение «группа» ввести атрибут «номер группы» и по значению, сохраненному в этом поле, переходить либо к отношению «группа ОИ-002», либо к отношению «группа ОИ-001». Такое решение проблемы приводит к нарушению принципов реляционной модели и лишает нас возможности полноценно использовать SQL, поскольку в нем не предусмотрена обработка подобных ситуаций.

Способ 3. Использование отдельных отношений.

Теперь сущности «группа ОИ-002» и «группа ОИ-001» хранят свои атрибуты в отдельных отношениях. Это решение лучше первого варианта, поскольку не имеет его недостатков, и лучше второго, так как не приводит к нарушению правил использования реляционной модели. Однако теперь утрачена такая сущность, как «группы». Если у нас есть отношения, поддерживающие ссылочную целостность с отношением «группы», то куда они будут ссылаться?

Итак, каждый из рассмотренных способов имеет свои недостатки, указанные выше, однако мы считаем возможным использовать комбинацию двух последних способов. То есть уникальные атрибуты групп хранить, как отдельные отношения, а общие, - как единое отношение, в котором содержится ссылка на уникальные атрибуты данного объекта сущности. Например, ФИО, данные о сессиях, баллах и предметах хранятся в одной сущности - «Группы ОИ» (рисунок 5 - трехмерное представление, таблица 1 - табличное представление), и здесь же стоит ссылка на сущность «Звание» (рисунок 6 — трехмерное представление, таблица 2 — табличное представление), в которой хранится информация о том специалист или магистр данный студент.

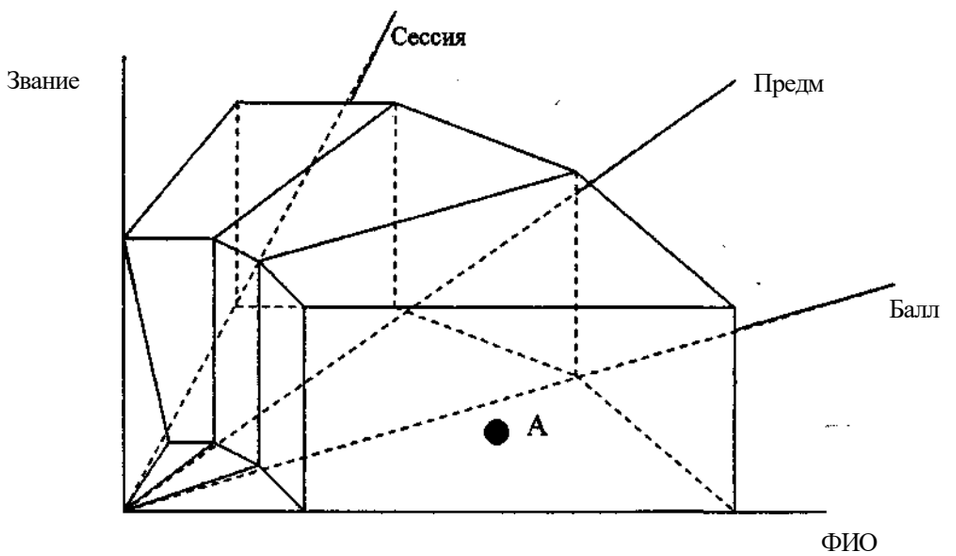


Рис. 5 - Трехмерное представление гиперкуба с пятью измерениями сущности «Группы ОИ».

Таблица 1

ФИО	Сессия	Предмет	Балл	Звание
Иванов И.И.	Зима 2005	Микроэконом ика	500	2
Иванов И.И.	Зима 2005	Системные технологии	480	2
Петров П.П.	Лето 2005	Макроэкономика	350	1
Петров П.П.	Лето 2005	Информационные системы	410	1
...		...		

Таблица 2

ФИО	Специалист	Магистр
Иванов И.И.	нет	да
Петров П.П.	да	нет

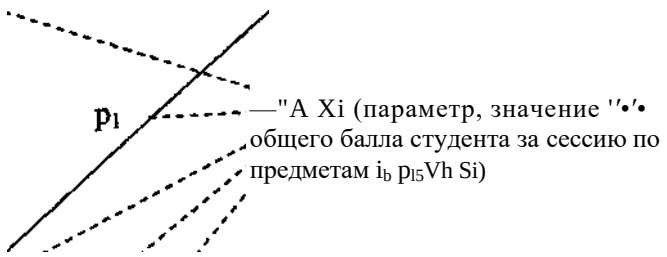
Таким образом, выбранный вариант наилучшим образом решает поставленную задачу и максимально исключает появление недостатков всех остальных способов решения.

Для реализации поставленной задачи будут использованы OLAP-средства, которые предоставляют удобные быстродействующие средства доступа, просмотра и анализа информации. Пользователь получает естественную, интуитивно понятную модель данных, организуя данные в виде многомерных кубов. Осями многомерной системы координат служат основные атрибуты анализируемого процесса (то, по чему ведется анализ) - измерения. Например, для кафедры как исследуемой предметной области такими измерениями могут быть предмет, учебный семестр, преподаватель и группа. Всегда в качестве одного из измерений используется время. Каждое измерение включает направления консолидации данных, состоящие из серии последовательных уровней обобщения (уровней иерархии), где каждый вышестоящий уровень соответствует большей степени агрегации данных по соответствующему измерению (различные уровни их детализации). В этом случае становится возможным произвольный выбор желаемого уровня детализации информации по каждому из измерений.

На пересечениях осей - измерений - находятся данные, количественно характеризующие процесс - параметры (стипендия, средний балл за сессию, количество часов на предмет и т.д.) (рис. 7).

Табличное представление многомерного представления данных о полученных баллах студента за сессию показано в таблице 3.

Благодаря такой модели данных пользователи могут формулировать сложные запросы, генерировать отчеты, получать подмножества данных.



Рис, 7 - Схема многомерного представления данных о полученных баллах студента за сессию

Таблица 3.

	Предметы сессии				
ФИО	(	*P*	*t*	и	5
Иванов И.И.	/i	*Pi*	*t*\	L>I	*Si*
Общая сумма баллов					X,

Однако использование OLAP-средств накладывает ряд следующих ограничений:

1) работа с OLAP предполагает четкое знание того, какую информацию вы хотите получить из базы данных, так как OLAP - это прежде всего инструмент добычи информации; получить ответы на нечетко поставленные вопросы невозможно, так как нельзя сформировать запрос к базе данных;

2) использование многомерного представления в виде гиперкубов может пагубно отразиться на быстродействии процесса обработки данных, так как, если гиперкуб содержит большое количество измерений, а, следовательно, и данных, то для их анализа ему потребуется много времени, поэтому необходимо выбирать именно те измерения, которые требуются для ответа на поставленный вопрос и из них формировать гиперкуб;

3) гиперкуб состоит из сгруппированных по определенным измерениям параметров (числовых или другого вида данных), которые присущи, то есть имеют определенные значения параметров, не всем объектам предметной области; те ячейки гиперкуба, в которых нет параметров, содержат пустоты, то есть гиперкуб с большим количеством агрегированных измерений более, чем на 50% состоит из пустот, что плохо сказывается как на быстродействии и рациональном использовании памяти, так и на объеме хранимого объекта [5].

III. ЗАКЛЮЧЕНИЕ

Итак, использование OLAP-средств имеет ряд недостатков, однако, главным его достоинством является простота и наглядность работы конечного пользователя с большими объемами информации, а также возможность визуализации и проведения немедленного анализа полученных результатов, что способствует повышению быстродействия аналитической работы. Конечно, это происходит только в том случае, когда OLAP-средства используются верно и построение гиперкубов происходит правильно.

ЛИТЕРАТУРА

1. Ядро OLAP системы. Часть 3 - построение срезов куба / Алексей Стариков - www.basegroup.ru. - 21.05.2003.
2. Ядро OLAP системы. Часть 2 - внутри гиперкуба / Алексей Стариков - www.basegroup.ru. -21.05.2003.
3. Ядро OLAP системы. Часть 1 - принципы построения. / Алексей Стариков - www.basegroup.ru. - 21.05.2003.
4. Концепции построения и реализации информационных систем, ориентированных на анализ данных. / Сахаров А.А. - www.olap.ru. - 11.03.2003.
5. Разработка модели данных для целей оперативной аналитической обработки финансовой информации университета. / Б.А.Горелов, Б.Б. Горелов. - www.ecsocman.edu.ru. - 21.03.2003.
6. Структура базы данных. Базы данных. Многомерные базы данных. - www.glossary.ru. - 11.03.2003.
7. Операции над векторами. - www.ispu.ru. - 8.9.2003.

Поступила в редакцию 17.10.2005 г., принята к печати 20.12.2005 г.