OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 007:519.816

ИНТЕГРАЦИЯ АППАРАТА ХРАНИЛИЩ ДАННЫХ И ТЕМПОРАЛЬНЫХ МОДЕЛЕЙ

Еремеев А.П., Еремеев А.А.

НИУ «Московский энергетический институт»,
г. Москва, Российская Федерация
eremeev@appmat.ru
YeremeevAA@mpei.ru

Хранилища данных – основное средство для консолидированного представления данных в пределах организации. Время – одна из размерностей, позволяющая сравнивать различные периоды, а темпоральные модели – средство для представления и оперирования временными (темпоральными) зависимостями в данных и знаниях. В работе рассматриваются возможности реализации темпорального хранилища данных, позволяющего оперировать с темпоральными данными и знаниями и передавать их между различными временны̀ми версиями, что крайне необходимо для современных интеллектуальных систем типа интеллектуальных динамических систем и систем поддержки принятия решений реального времени. Работа выполнена при финансовой поддержке РФФИ (проект № 11-01-00140).

Ключевые слова: интеллектуальная система, хранилище данных, темпоральная база данных, технология OLAP.

Введение

Большинство крупных компаний и организаций для долгосрочного анализа, сбора и хранения данных с целью предоставления результирующей информации пользователям в настоящее время используют *хранилища данных* (ХД) [Kimball et al., 2002]. ХД ориентировано на хранение и анализ данных, охватывающих большой временной промежуток. Интеграция различных информационных систем организации – трудоемкое и достаточно длительное действие, особенно, если используются различные программные продукты, которые были разработаны независимо. Чтобы получить полную информацию об организации и использовать перспективные интеллектуальные системы поддержки принятия управленческих решений, необходимо интегрировать данные, полученные от различных источников, учитывая временной фактор, и сделать эти данные доступными для использования.

При помощи технологии ХД может быть выполнена интеграция, в которой оперативные данные водятся и обрабатываются совместно с информацией из классических информационных системам (баз данных), содержащими данные, необходимыми для функционирования организации. Данные, сохраненные в ХД, предоставляют информацию на протяжении времени (временном интервале), которое необходимо для того, чтобы эффективно управлять (принимать решения) организацией или ее компонентами (объектами). В результате процессов принятия решений данные могут быть возвращены в системы оперативной обработки информации.

Темпоральное ХД можно охарактеризовать как объектно-ориентированный, интегрируемый, поддерживающий временной фактор и долговременный (неразрушающийся) набор данных (совокупность наборов данных), ориентированных на использование в интеллектуальных системах (ИС), в том числе ИС поддержки принятия решений реального времени (ИСППР РВ) [Вагин и др., 2001]. В ИСППР РВ вместе с ХД могут использоваться различные модели и методы для поддержки принятия решений, в том числе и темпоральные модели, ориентированные на представление и оперирование данными и знаниями, содержащими временные (темпоральные) зависимости [Еремеев и др., 2011 а, б]. Для того, чтобы дополнить стандартные средства ХД по подготовке различных отчетов средствами анализа и интерактивного доступа к данным применяются средства (технология) оперативной аналитической обработки данных (On-Line Analitical Processing, OLAP), или

OLAP-технология многомерного анализа данных. Темпоральные ХД часто используются, чтобы интегрировать различные данные, включая темпоральные [Herden, 2002]. Такие ХД часто являются основой для систем управления взаимоотношениями с клиентами (CRM-систем), систем планирования ресурсов предприятия (ERP-систем) и других современных информационных систем. Перспективным использованием технологии ХД является интеграция с OLAP-технологией с применением многомерного представления данных. Средства OLAP-технологии позволяют агрегировать и сравнить данные по различным измерениям (размерностям), относящимся к области определения приложения. К типичным размерностям, содержащимся часто в бизнес-ХД, относятся – время, организационная структура (подразделения, отделы, и т.д.), территориальное нахождение (города, области, страны) и данные управляемого объекта или производимого товара.

Такое многомерное представление обеспечивает наличие долгосрочных данных, которые могут быть проанализированы по оси времени, тогда как большинство ХД только предоставляют снимки данных во времени. Известные OLAP-системы позволяют иметь дело с изменениями значений, например, параметров или характеристик организации (объекта, товара и т.п.), но они не в состоянии оперировать с модификациями в размерностях, например, при новом ответвлении или разделении, несмотря на то, что время обычно явно представляется как размерность в ХД.

Причина этого явления состоит в том, что технология ХД базируется на предположении (ограничении), что размерности являются ортогональными. Ортогональность относительно размерности времени означает, что другие размерности должны быть независимыми от времени. Это ограничение запрещают надлежащую обработку изменений в данных размерности.

Для корректности результатов OLAP-запросов необходимо, чтобы модификации во времени данных (размерности) были приняты во внимание. Например, когда анализируется функционирование некоторой системы и ее компонентов за достаточно длительный период времени, необходимо учитывать, добавление новых или удаление некоторых компонентов за этот период. По данной проблематике в настоящее время проводятся серьезные исследования ввиду ее актуальности.

1. Темпоральные хранилища данных

Для того, чтобы получить корректные результаты OLAP-запросов важно отследить модификации размерностей данных. В связи с этим необходимо ввести временные расширения в средства ХД. Следовательно, все элементы измерения и все иерархические ссылки между этими элементами измерения должны быть представлены с временными метками из некоторого интервала $[T_s; T_e]$, где T_s и T_e – соответственно начало и конец действительного времени, $T_e > T_s$.

Используя временную проекцию и временную выборку, как это определено в Глоссарии темпоральных баз данных [Dyreson et al., 1994], можно построить базовую структуру темпорального ХД, ориентированную на определенный временной период (интервал). При этом каждая модификация элемента измерения или иерархического отношения приводит к новой версии структуры, если такая версия для данного временного интервала не существует. Такие модификации могут быть сделаны при помощи трех основных операций ХД – INSERT (Добавить), UPDATE (Обновить) и DELETE (Удалить) и некоторых более сложных операций, например, SPLIT (Разделить) и MERGE (Слить).

Рассмотрим возможные подходы к организации запросов к темпоральному ХД. Первый, так называемый *защитный (defensive) подход*, базируется на том, что для каждого сформулированного запроса выполняется проверка - пересекает ли запрос временные границы определенных версий структуры. Если границы между версиями пересекаются, то для такого запроса возможны следующие альтернативы: отклонить запрос или выдать предупреждение.

Однако данный подход не обеспечивает правильность всех возможных запросов. Рассмотрим пример. Пусть отдел А организации с сентября 2011 года разделяется на два подразделения - А1 и А2. Если мы хотим проанализировать все месяцы 2011 года, то для отдела А есть данные за январь-август, а с сентября есть данные только для подразделений А1 и А2. Поэтому специалист по анализу должен иметь достаточное знание предметной области и, в частности, знать, как отдел А и подразделения А1 и А2 связаны между собой.

Следовательно, в темпоральных ХД необходимо иметь механизм для представления зависимостей и отношениий между элементами измерения в различных версиях структуры.

Для приведенного примера возможно представить оборот подразделения А1 в течение периодов до сентября 2011 с помощью функции **turnover(A1; period) = 30 % turnover (A; period)**. Также можно показать, что в течение всех периодов с сентября 2011 года численность персонала M#

подразделения A соответствует функции **M#(A; period) = M#(A1; period) +M#(A2; period)**. Используя такие функции, можно гарантировать, что успешный анализ может быть сделан даже в тех случаях, когда могут быть изменения в структуре.

Такого типа функции называются *функциями преобразования* и для их определения предназначена операция MapF.

Темпоральная модель ХД позволяет иметь дело с изменениями на уровне экземпляра. При этом используются следующие типы данных [Salzberg et al., 1999]:

- *структурированные данные,* которые содержат необходимую информацию относительно всех структур ХД, т.е. информацию о размерности и временных метках иерархической структуры;
- *отображение структурных данных*, содержащее все функции MapF, т.е. все функции преобразования, которые необходимы для преобразования данных из одной структуры в другую;
- *фактические данные*, представляющие всю фактическую информацию, содержащуюся в ХД.

Используя приведенные типы данных, можно организовать запросы на основе двух видов так называемого *наступательного (offensive) подхода*, который будут рассмотрен далее.

Темпоральное ХД задается как ряд размерностей, например размерности *Время*, *Факты* и *Продукты*. При этом, по сути не различаются размерность и факты, так как факты могут быть определены при помощи размерности *Факты*. Следовательно, рассматриваются факты, которые находятся в иерархическом порядке. Подобно тому, как можно описать размерность *Рынок*, где хранится, например, принадлежность к областям или государствам, можно описать размерность *Факты* с введенным иерархическим порядком, в частности, между *Прибылью*, *Оборотом* и *Затратами*.

Каждая размерность состоит из уровней, например, уровни *День*, *Месяц* и *Год* принадлежат размерности *Время*. Иерархическая структура для этих уровней размерности определена через ряд иерархических назначений (присваиваний) (Hierarchical Level Assignment), например, уровни измерения в течение размерности *Время* размерности находятся в иерархическом порядке *День → Месяц → Год*. Экземпляр уровня размерности называют элементом размерности, например, июнь - экземпляр уровня размерности *Месяц*. Эти элементы измерения находятся снова в иерархическом порядке (Hierarchical Member Assignment). Каждая размерность – данные некоторой ячейки, использует ссылку на ряд элементов размерности.

2. Организация запросов к темпоральным хранилищам данных

Для выполнения запроса пользователь сначала должен выбрать определенную (основную) версию структуры. Эта основная версия определяет, какая структура должна использоваться для анализа. В большинстве случаев это будет текущая версия структуры. Однако, в некоторых случаях пользователю необходимо использовать "более старую" (одну из предыдущих) версию структуры. Рассмотрим, например, темпоральное ХД, которое содержит данные обо всех странах Европейского союза (ЕС) с 1988 года [Eder et al, 2001]. В пределах этого темпорального ХД произошли два изменения структуры, связанные с объединением Германии в 1990 г. и вступлением Швеции, Австрии и Финляндии в ЕС в 1995 г.. Поэтому можно выделить три версии структуры, которые допустимы для периодов времени, отмеченных в табл. 1, где T_s - начало действительного времени, T_e - конец действительного времени для соответствующей версии структуры.

Таблица 1 – пример версий структур.

Версия структуры	T_s	T_e
SV_1	1988	1989
SV_2	1990	1994
SV_3	1995	∞

Предположим, что пользователь выбирает версию структуры SV_3 в качестве основной версии и запрашивает данные на 1994 г. и 1995 г. для анализа. В этом случае требуются функции преобразования для отображения данных, которые

допустимы и для версии SV2, и для версии SV3. Однако тот же самый анализ мог быть выполнен и с использование версии структуры SV2 в качестве основной версии. Тогда для выполнения запроса потребуются функции для отображения данных из SV3 в SV2.

Организация запросов должна быть такова, чтобы система не только была способна правильно ответить на запросы, охватывающие многократные периоды и возможно различные версии структуры (данных размерности), но и могла сообщить пользователю, какие структурные модификации имели место. Например, если пользователь утверждает, что запрос "дать число жителей Германии в 1998, 1999 и 1990 гг.", то необходимо сообщить пользователю, что для размерности *Государство* в 1990 г. произошла модификация структуры, а именно - объединение Германии. В зависимости от выбранной основной версии структуры результат запроса может иметь вид, как показано на рис. 1 а) или на рис. 1 б).

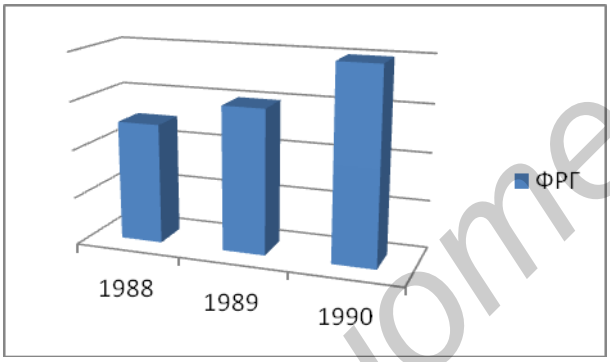

Рисунок 1а) - результат запроса с SV_2 или SV_3

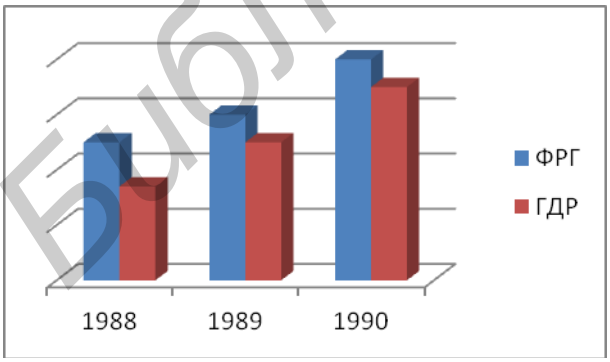

Рисунок 1б) - результат запроса с SV_1

Для решения поставленной задачи по организации запросов возможны два разных подхода. Оба подхода базируются на следующих общих компонентах:

- *административный инструмент*, позволяющий администратору импортировать новые данные в ячейки темпорального ХД и, конечно, задавать изменения многомерной структуры (в предлагаемом прототипе реализуется с помощью языка программирования Java);
- *темпоральное ХД,* содержащее необходимую информацию о версии структуры, данные и функции преобразования, как было описано ранее (в качестве основы для организации темпоральног хранилища данных можно использовать, например, СУБД Oracle);
- *преобразователь (трансформатор),* который преобразует все требуемые значения ячеек данных из всех используемых версий структуры в выбранную пользователем версию структуры с помощью определенных функций преобразования, то есть, используя необходимые функции MapF (реализуется с помощью языка программирования Java).

Для организации запросов в качестве первого шага используется, так называемый косвенный подход, который проще реализовать, нежели базовый прямой подход, так как при косвенном подходе не требуется реализация интерфейса для исследований.

Основная идея косвенного подхода состоит в том, что преобразователь генерирует одно ХД для каждой версии структуры, необходимой пользователю (рис. 2 а, где DM(Dimension member) - один из элементов списка, который составляет измерение). В большинстве случаев это будет фактической версией структуры. Каждое ХД при этом состоит из всех фактических данных, которые допустимы для того же временного интервала, что и соответствующая версия структуры, а также из всех фактических данных, которые могли быть получены (преобразованы) функциями MapF из других версий структуры. Поэтому пользователь определяет основную версию структуры, выбирая определенное ХД.

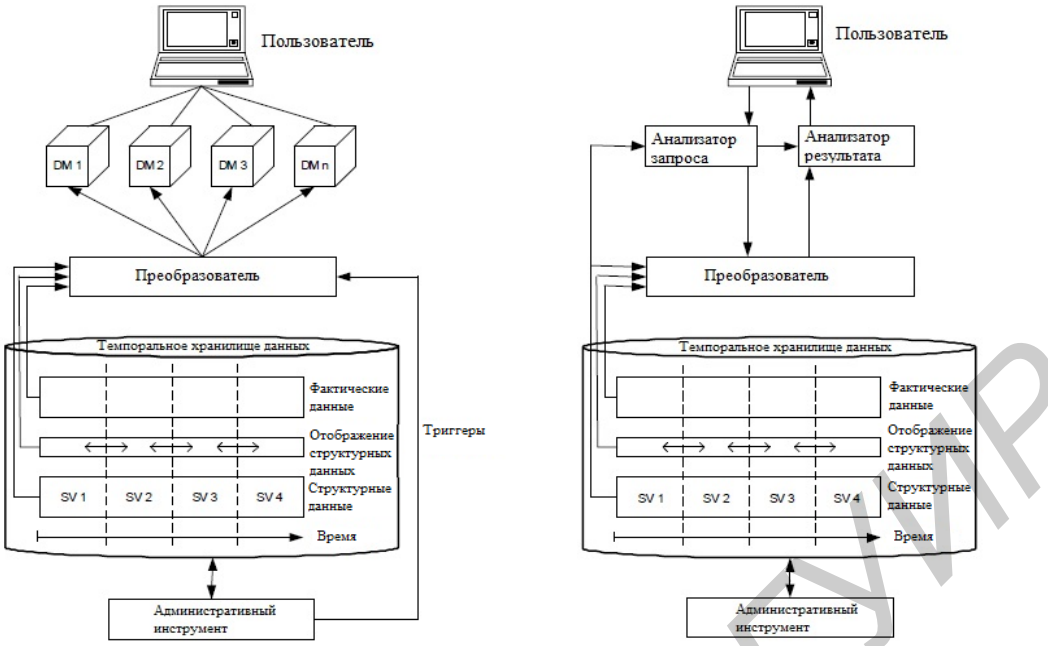

Рисунок 2 - Структура а) косвенного подхода и б) прямого подхода.

Данный подход называется косвенным, потому что преобразователь инициирован административным инструментом, а не пользователем непосредственно. Другими словами - преобразователь начинает генерировать ХД только после того, как была сгенерирована новая версия структуры или были импортированы новые измерения. В обоих случаях преобразователь должен повторно анализировать все существующие ХД, чтобы возвратить непротиворечивую информацию.

Косвенный подход реализован при помощи СУБД Oracle в качестве основы для организации темпорального ХД и Hyperion Essbase в качестве интерфейса, содержащего ХД.

У косвенного подхода имеются и недостатки (ограничения). Основной недостаток заключается в том, что пользователю предоставлен инструмент для организации запросов даже после структурных изменений, однако в этом случае необходимо предварительно сообщить пользователю о том, какие структурные изменения оказали влияние на поставленный запрос. Следовательно, необходимо пополнить результат запроса некоторой дополнительной информацией, как показано на рис. 2 б).

Организация (структура) прямого подхода включает следующие составляющие (см. рис. 2 б):

- анализатор запроса: берет запрос, поставленный пользователем, в качестве входных данных и определяет, какие данные из какой версии структуры необходимы, чтобы ответить на запрос. Результат далее передается в преобразователь и в анализатор результата;
- преобразователь: в отличие от косвенного подхода, преобразователь инициируется пользователем или, другими словами, для каждого установленного запроса преобразователь преобразовывает все необходимые значения ячеек с целью получения ответа на запрос;
- анализатор результата: использует данные, переданные из анализатора запроса и преобразователя, что обогащает результат, полученный преобразователем дополнительной информацией о пользователе, т.е. информацией о том, какие структурные модификации оказали влияние на установленный запрос. В проводимой работе анализатор результата является предметом дальнейшего исследования;
- темпоральное ХД и административный инструмент: эти компоненты описаны ранее.

Можно констатировать, что прямой подход улучшает не только правильность OLAP-запросов после модификации многомерной структуры, но и качество интерпретации ответов на эти запросы.

Заключение

В работе рассмотрены два подхода (структуры) организации темпорального ХД на основе интеграции аппарата ХД и темпоральной модели. Для обоих подходов строится базовое темпоральное ХД, которые отличаются по способу организации запросов и соответственно ответов на запросы. Прямой подход использует анализатор запроса и анализатор результата, чтобы обогатить результат каждого установленного запроса информацией о воздействии структурных модификаций на запрос. Косвенный подход генерирует ХД для определенной версии структуры и преобразовывает данные от всех других версий структуры в это ХД, используя заданную спецификацию функций преобразования. Таким образом, прямой подход предлагает большую гибкость, а косвенный подход предпочтителен по времени отклика. Таким образом, каждый подход имеет свою область предпочтения.

В контексте использования темпоральных ХД (БД) совместно с ИСППР РВ планируется дальнейшее исследование обоих подходов относительно производительности, требований пространства памяти и т.д. Другая важная область исследования связана с построением функций преобразования не только между элементами измерения, но и между размерностями и уровнями измерения. Следует подчеркнуть, что представление функций преобразования на уровне схемы намного сложнее, чем представление на уровне экземпляра темпорального ХД.

Отметим, что рассмотренные подходы улучшают правильность интерпретации ответов на OLAP-запросы и избавляют пользователя от необходимости иметь подробные знания об истории изменения размерности данных.

В настоящее время на кафедре прикладной математики Национального исследовательского университета «МЭИ» ведутся научно-исследовательские работы по созданию базового математического и программного обеспечения для конструирования ИСППР РВ для управления сложными техническими и организационными системами типа объектов энергетики и транспортных систем [Вагин и др., 2001] с использованием нетрадиционных логик, включая темпоральные логики и соответствующие темпоральные БД [Еремеев и др., 2011 а, б].

Библиографический список

[Kimball et al., 2002] Kimball R., Ross M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition). John Wiley & Sons. 2002. P. 188 - 198.

[**Вагин и др., 2001]** Вагин В.Н., Еремеев А.П. Некоторые базовые принципы построения интеллектуальных систем поддержки принятия решений реального времени // Известия РАН. Теория и системы управления. 2001. № 6. С. 114-123.

[**Еремеев и др., 2010 а]** Еремеев А.А., Еремеев А.П., Пантелеев А.А. Темпоральная модель данных и возможности ее реализации на основе технологии OLAP // Двенадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2010 (20-24 сентября 2010гю, г.Тверь, Россия): Труды конференции. В 4-х томах. Т. 3. - М.: Физматлит, 2010. С. 345-353.

[**Еремеев и др., 2010 б]** Еремеев А.П., Еремеев А.А., Пантелеев А.А. Темпоральные базы данных и их применение в интеллектуальных системах // Интеллектуальные системы. Коллективная монография. Выпуск 4/ Под. ред. В.М. Курейчика. – М.: Физматлит, 2010. С. 253-276.

[Herden, 2002] Herden O. TOLAP: Temporal Online Analytical Processing // International Baltic Conference on Databases and Information Systems 2, 2002. P. 55-66.

[Dyreson et al., 1994] Dyreson C. E., Grandi F., Snodgrass R. T. et al. A Consensus Glossary of Temporal Database Concepts // *ACM SIGMOD Record*, 23(1), 1994. P. 52-64.

[Salzberg et al., 1999] Betty Salzberg and Vassilis J.Tsortas. A Comparison of Access Methods for Temporal Data. ACM Computing Surveys, 31(2), 1999. P.158-221.

[Eder et al, 2001] Johann Eder, Christian Koncilia, Tadeusz Morzy, A model for a temporal warehouse // Proceedings of the Open enterprise solutions: systems, experiences and organizations Workshop (OES-SEO 2001), Rome, Italy. P. 48-54

[**Еремеев и др., 2011 а]** Еремеев А.П., Еремеев А.А., Пантелеев А.А. Возможности реализации темпоральной базы данных для интеллектуальных систем // Программные продукты и системы. №2(94), 2011. С. 3-7.

[**Еремеев и др., 2011 б]** Еремеев А.П., Еремеев А.А., Пантелеев А.А. Применение технологий хранилищ данных и темпоральных баз данных в интеллектуальных системах // Вестник РГУПС. №3, 2011. С. 66-72.

INTEGRATION OF THE DATA WAREHOUSES TOOLS AND THE TEMPORAL MODELS

Eremeev A.P., Eremeev A.A.

National Research University ‘Moscow Power Engineering Institute’, Moscow, Russia

Data warehouses are the main means for consolidated data representation in frames of an organization. The time is one of measurements allowing to compare various periods, and temporal models are the means to present and operate by temporal dependencies in data and knowledge. The possibilities of implementing temporal date warehouses allowing to operate with temporal data and knowledge and transfer them between different temporal versions are viewed. It is extremely important for modern intelligent systems of the type of intelligent dynamic systems and real time decision support ones. The works is performed by financial support of RFBR (the project No 11-01-00140).

Keywords: intelligent system, data warehouse, temporal data base, OLAP-technology