УДК 681.3.07

П. П. Урбанович, профессор; Р. Х. Хтажа, аспирант;
М. Беднарчик (университет KUL, г. Люблин, Польша)

ИНТЕЛЛЕКТУАЛЬНЫЕ БАЗЫ ДАННЫХ В СИСТЕМАХ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

The article is concerned of intelligent databases. The intelligent database characterized as databases, that manage information in a natural way, making the information to be easy to store, access and use. Instrumental recourses for this purpose are analyzed. It is noted, that the most perspective is the direction of development and use of intellectual DB – the information systems, allowing to form administrative and management decisions for various practical areas. Directions and methods of integration of systems of operative analytical processing and the intellectual analysis of data, structure and tool means of such integrated systems are considered.

Введение. В последние годы в мире оформился ряд новых концепций хранения и анализа корпоративных данных:

– хранилища данных, или склады данных (Data Warehouse);

– оперативная аналитическая обработка (OAO – On-Line Analytical Processing, OLAP);

– интеллектуальный анализ данных (ИАД – Data Mining).

Технологии OLAP тесно связаны с технологиями построения систем Data Warehouse и методами интеллектуальной обработки. Поэтому *интеллектуальные базы данных* (ИБД) используются для решения проблемы передачи, хранения и обработки информации с большими объемами, а также для потребностей пользователя с целью получения доступной к пониманию информации [1].

Очень часто информационно-аналитические системы, создаваемые в расчете на непосредственное использование лицами, принимающими решения, оказываются чрезвычайно просты в применении, но жестко ограничены в функциональности. Такие статические системы называются в литературе информационными системами руководителя. Они содержат в себе предопределенные множества запросов и, будучи достаточными для повседневного обзора, не способны ответить на все вопросы к имеющимся данным, которые могут возникнуть при принятии решений.

Динамические системы поддержки принятия решения (СППР), напротив, ориентированы на обработку нерегламентированных (ad hoc) запросов аналитиков к данным. Наиболее глубоко требования к таким системам рассмотрел E. F. Codd в статье [2], положившей начало концепции OLAP. Работа аналитиков с этими системами заключается в интерактивной последовательности формирования запросов и изучения их результатов.

Для того чтобы существующие хранилища данных способствовали принятию управленческих решений, информация должна быть представлена аналитику в нужной форме, т. е. он должен иметь развитые инструменты доступа к данным хранилища и их обработки.

В данной статье рассматриваются и анализируются методы интеграции систем OAO и ИАД, структура и инструментальные средства таких интегрированных систем.

Интеллектуальный анализ данных. ИАД (Data Mining) – это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации). При этом накопленные сведения автоматически обобщаются до информации, которая может быть охарактеризована как знания.

В общем случае процесс ИАД состоит из трех стадий:

1) выявление закономерностей (свободный поиск);

2) использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование);

3) анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях.

Иногда в явном виде выделяют промежуточную стадию проверки достоверности найденных закономерностей между их нахождением и использованием (стадия валидации).

Все методы ИАД подразделяются на две большие группы по принципу работы с исходными обучающими данными [3].

В первом случае исходные данные могут храниться в явном детализированном виде и непосредственно использоваться для прогностического моделирования и/или анализа исключений; это так называемые методы рассуждений на основе анализа прецедентов. Главной проблемой этой группы методов является затрудненность их использования на больших объемах данных, хотя именно при анализе больших хранилищ данных методы ИАД приносят наибольшую пользу.

Во втором случае информация вначале извлекается из первичных данных и преобразуется в некоторые формальные конструкции (их вид зависит от конкретного метода). Согласно предыдущей классификации, этот этап выполняется

на стадии свободного поиска, которая у методов первой группы в принципе отсутствует.

Таким образом, для прогностического моделирования и анализа исключений используются результаты этой стадии, которые гораздо более компактны, чем сами массивы исходных данных. При этом полученные конструкции могут быть либо «прозрачными» (интерпретируемыми), либо «черными ящиками» (нетрактуемыми).

Интеграция систем ОАО и ИАД. Оперативная аналитическая обработка и интеллектуальный анализ данных – две составные части процесса поддержки принятия решений. Большинство систем ОАО заостряет внимание только на обеспечении доступа к многомерным данным, а большинство средств ИАД, работающих в сфере закономерностей, имеют дело с одномерными перспективами данных. Эти два вида анализа должны быть тесно объединены, т. е. системы OLAP должны фокусироваться не только на доступе, но и на поиске закономерностей.

В плане практической реализации и с точки зрения формализации выполняемых операций существующие подходы в построении систем рассматриваемого класса можно разделить на три группы:

– на основе выполнения интеллектуального анализа, который обеспечивается над любым результатом запроса к многомерному концептуальному представлению, т. е. над любым фрагментом любой проекции гиперкуба показателей;

– подобно данным, извлеченным из хранилища, результаты интеллектуального анализа должны представляться в гиперкубической форме для последующего многомерного анализа;

– гибкий способ интеграции позволяет автоматически активизировать однотипные механизмы интеллектуальной обработки над результатом каждого шага многомерного анализа (перехода между уровнями обобщения, извлечения нового фрагмента гиперкуба и т. д.).

На рис. 1 представлена обобщенная структурная схема системы многомерного интеллектуального анализа данных.

Сфера детализированных данных. Это область действия большинства систем, нацеленных на поиск информации. В большинстве случаев реляционные СУБД отлично справляются с возникающими здесь задачами.

Общепризнанным стандартом языка манипулирования реляционными данными является SQL.

Сфера агрегированных показателей. Комплексный взгляд на собранную в хранилище данных информацию, ее обобщение и агрегация, гиперкубическое представление и многомерный анализ являются задачами систем оперативной аналитической обработки данных (OLAP). Здесь можно или ориентироваться на специальные многомерные СУБД [4], или оставаться в рамках реляционных технологий. Во втором случае заранее агрегированные данные могут собираться в БД звездообразного вида, либо агрегация информации может производиться на лету в процессе сканирования детализированных таблиц реляционной БД.

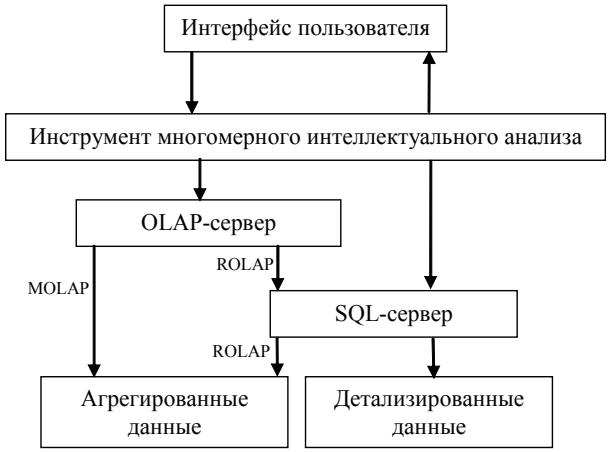

Рис. 1. Архитектура системы многомерного интеллектуального анализа данных

Аббревиатуры ROLAP и MOLAP означают соответственно реляционные (relational) и многомерные (multi-demensional) OLAP.

Многомерное концептуальное представление (multi-dimensional conceptual view) есть множественная перспектива, состоящая из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных. Одновременный анализ по нескольким измерениям определяется как многомерный анализ. Каждое измерение включает направления консолидации данных, состоящие из серии последовательных уровней обобщения, где каждый вышестоящий уровень соответствует большей степени агрегации данных по соответствующему измерению (рис. 2). Так, измерение **Исполнитель** может определяться направлением консолидации, состоящим из уровней обобщения «предприятие – подразделение – отдел – служащий». Измерение **Время** может даже включать два направления консолидации: «год – квартал – месяц – день» и «неделя – день», поскольку счет времени по месяцам и по неделям несовместим. В этом случае становится возможным произвольный выбор желаемого уровня детализации информации по каждому из измерений. Операция спуска (drilling down) соответствует движению от высших ступеней консолидации к низшим; напротив, операция подъема (rolling up) означает движение от низших уровней к высшим.

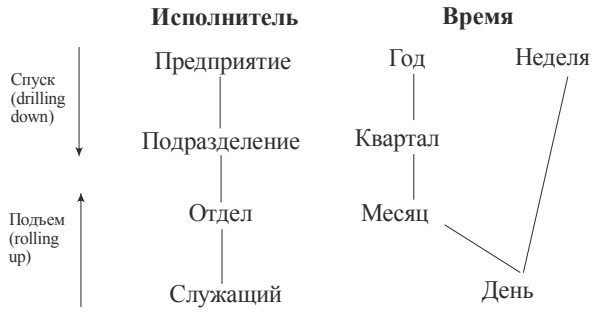

Рис. 2. Пример измерений и направлений консолидации данных

В специализированных СУБД, основанных на многомерном представлении данных, последние организованы не в форме реляционных таблиц, а в виде упорядоченных многомерных массивов типа:

– гиперкубов – все хранимые в БД ячейки должны иметь одинаковую мерность, т. е. находиться в максимально полном базисе измерений;

– поликубов – каждая переменная хранится с собственным набором измерений, и все связанные с этим сложности обработки перекладываются на внутренние механизмы системы.

Использование многомерных БД в системах оперативной аналитической обработки имеет следующие достоинства:

1) в случае использования многомерных СУБД поиск и выборка данных осуществляются значительно быстрее, чем при многомерном концептуальном взгляде на реляционную базу данных, так как многомерная база данных денормализована, содержит заранее агрегированные показатели и обеспечивает оптимизированный доступ к запрашиваемым ячейкам;

2) многомерные СУБД легко справляются с задачами включения в информационную модель разнообразных встроенных функций, тогда как объективно существующие ограничения языка SQL делают выполнение этих задач на основе реляционных СУБД достаточно сложным, а иногда и невозможным.

С другой стороны, имеются существенные ограничения:

– многомерные СУБД не позволяют работать с большими базами данных. К тому же за счет денормализации и предварительно выполненной агрегации объем данных в многомерной базе, как правило, соответствует (по оценке [4]) в 2,5–100 раз меньшему объему исходных детализированных данных;

– многомерные СУБД по сравнению с реляционными очень неэффективно используют внешнюю память.

В подавляющем большинстве случаев информационный гиперкуб является сильно разреженным, а поскольку данные хранятся в упорядоченном виде, неопределенные значения удается удалить только за счет выбора оптимального порядка сортировки, позволяющего организовать данные в максимально большие непрерывные группы. Но даже в этом случае проблема решается только частично. Кроме того, оптимальный с точки зрения хранения разреженных данных порядок сортировки, скорее всего, не будет совпадать с порядком, который чаще всего используется в запросах. Поэтому в реальных системах приходится искать компромисс между быстродействием и избыточностью дискового пространства, занятого базой данных.

Уровни интеллекта в интеллектуальных базах данных. Различают три инструментальных уровня интеллекта, которые делают базу данных интеллектуальной (рис. 3):

– инструменты высокого уровня;

– интерфейс с пользователем высокого уровня;

– машина базы данных.

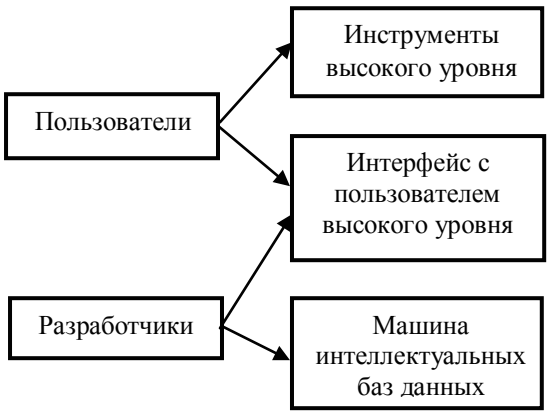

Рис. 3. Взаимосвязь компонентов и уровней интеллекта в интеллектуальных базах данных

Инструменты высокого уровня (high-level tools). Одной из отличительных особенностей интеллектуальных БД является то, что они сочетают в себе ряд методов, которые ранее были изолированы друг от друга. Кроме того, эти методы лежат в основе структуры интеллектуальных БД. Существуют инструменты высокого уровня типа приложений, которые являются дополнениями к функции интеллектуальных баз данных. Эти инструменты могут быть применены пользователями БД для реализации ряда услуг (таких, например, как интеллектуальный поиск).

Будем различать следующие виды инструментов высокого уровня:

– *инструменты обнаружения знаний* (этот тип содержит инструменты для статистического анализа данных);

– *инструменты для контроля и совершенствования данных* (инструменты, которые обнаруживают и автоматически напоминают об ошибках в БД);

– *инструменты управления* (этот тип отражает тот факт, что информация может быть выражена в различных формах и на различных уровнях);

– *инструменты предоставления данных* (вполне естественно, что большинство пользователей нуждается в предоставлении некоторых из этих данных);

– *инструменты поддержки резолюции* (имеется особый вид средства информации, ее поиска и использования с целью принятия управленческих решений);

– *инструменты управления форм данных* (данные не всегда хранятся в той форме, в которой они нам нужны; инструменты этого типа позволяют пользователю преобразовать данные из одной формы в другую);

– *инструменты разработки интеллектуальной системы* (предоставляют услуги по разработке интеллектуальных БД).

Перечисленные инструменты позволяют разработчикам и системным администраторам улучшить разработку и эксплуатацию интеллектуальных баз данных.

Интеллект на уровне интерфейса с пользователем. Пользовательский интерфейс является частью системного программного обеспечения, которым занимается пользователь.

Интерфейс может быть интеллектуальным и зависит от простоты и формы использования. В этом случае интерфейс использует предположения для вывода информации, которые не могут быть получены непосредственно.

Интеллект на уровне машины БД. Пользовательский интерфейс поддерживается набором возможностей базы данных. Эти возможности предоставляют набор методов, позволяющих СУБД выполнять требуемые функции. Примером может служить обработка запроса.

Заключение. В статье рассмотрены и проанализированы тенденции развития информационных систем хранения, обработки и использования больших массивов слабо структурированных данных.

Отмечено, что наиболее перспективным является направление разработки и использования интеллектуальных БД, т. е. информационных систем, позволяющих формировать управленческие и менеджерские решения для различных практических приложений.

Рассмотрены направления и методы интеграции систем ОАО и ИАД, структура и инструментальные средства таких интегрированных систем.

Литература

1. Intelligent Databases / K. Saye [et al.]. – New York: John Wiley & Sons. – 1989. – 326 p.

2. Codd, E. F. Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate / E. F. Codd, S. B. Codd, C. T. Salley. – E. F. Codd & Associates. – 1993. – 462 p.

3. Parsaye, K. A. Characterization of Data Mining Technologies and Processes / K. A. Parsaye // The Journal of Data Warehousing. – 1998. – № 1. – 226 p.

4. Harinarayan, V. Implementing Data Cubes Efficiently / V. Harinarayan, A. Rajaraman, J. Ullman // SIGMOD Conference. – Montreal, CA. – 1996. – 412 p.