УДК 004

ПРИНЦИПЫ ОРГАНИЗАЦИИ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ В ХРАНИЛИЩЕ*

© 2012 Е.Ю. Виноградова**

Ключевые слова: информационные технологии, интеллектуальный анализ данных, хранилище данных.

Описываются основные принципы обработки данных в хранилище данных интеллектуальной информационной системы для реализации поддержки принятия решений руководством хозяйствующих субъектов.

Одной из важнейших составных частей современных аналитических систем являются средства интеллектуального анализа данных. Выполнение большинства аналитических запросов требует сложной статистической обработки данных, применения методов искусственного интеллекта. Современные СУБД, предназначенные для реализации аналитических систем, включают довольно обширный набор средств для статистической обработки информации. Однако задачи пользователей могут потребовать выполнения специфических операций над данными, поэтому средства анализа могут встраиваться также и в клиентские приложения.

В аналитических системах для обработки данных используется очень широкая номенклатура методов. Это и традиционные статистические методы регрессионного, факторного, дисперсионного анализа, анализа временных рядов, а также новые, получившие распространение в последнее время методы, основанные на искусственном интеллекте. К последним, как правило, относят нейронные сети, нечеткую логику, генетические алгоритмы, методы извлечения знаний. В совокупности они именуются методами интеллектуального анализа данных. Часто используется англоязычный термин “data mining” (дословно - извлечение знаний). Эти методы развивают традиционные статистические подходы, находя применение там, где обычные приемы невозможно использовать в силу отсутствия точных зависимостей, описывающих анализируемые процессы. Технологии интеллектуального анализа данных способны существенно расширить круг практически значимых задач, решаемых с использованием вычислительной техники.

В большинстве случаев средства анализа данных в системах поддержки принятия решений (СППР) на основе хранилищ данных используются для решения следующих задач:

- выделения в группах сходных по некоторым признакам записей (кластерный анализ)[1];
- нахождения и аппроксимации зависимостей, связывающих анализируемые параметры или события, а также поиска параметров, наиболее значимых в терминах конкретной задачи[2];
- поиска данных, существенно отклоняющихся от выявленных закономерностей (анализ аномалий);
- прогнозирования развития объектов различной природы на основе хранящейся ретроспективной информации об их состоянии в прошлом.

Кластерный анализ (также употребляются термины “кластеризация”, “самообучение”, “обучение без учителя”) - это метод выделения из множества элементов групп (кластеров) схожих между собой элементов. Предполагается, что элементы одного и того же кластера похожи, а элементы различных кластеров отличаются друг от друга. Как правило, число кластеров заранее не определяется. Кластерный анализ записей баз данных осуществляется на основе значений их количественных и качественных атрибутов. При этом делается попытка автоматически разнести имеющиеся записи по различным группам. Кластерный анализ применяют при ре-

* Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (грант 09-03-83306а/У).

** Виноградова Екатерина Юрьевна, кандидат экономических наук, доцент Уральского государственного экономического университета, г. Екатеринбург. E-mail: katerina@usue.ru.

шении большого числа задач. В социологии его используют для обработки результатов опросов общественного мнения, в медицине - для выявления типичных клинических случаев, в маркетинге - для поиска родственных групп клиентов. Часто выделение кластеров - отправная точка для других алгоритмов интеллектуального анализа данных.

Применение данной процедуры позволяет во многих случаях перейти от обработки всего массива записей к анализу относительно небольшого числа кластеров[3].

Системы интеллектуального анализа данных эффективно используются для автоматического нахождения взаимосвязей и нелинейных зависимостей в данных. Учет подобных зависимостей позволяет лучше осмыслить предметную область, повысить качество решений, принимаемых на основе анализа ее состояния. В отличие от традиционных корреляционных методов, способных выявлять линейную взаимосвязь между переменными, методы интеллектуального анализа данных обнаруживают и сложные нелинейные зависимости. Пакеты программ на их основе позволяют при обнаружении зависимостей определять их статистические характеристики, производить визуализацию области действия зависимости и выпадающих точек. Некоторые продукты интеллектуального анализа, например, система IDIS (The Information Discovery System) фирмы Intelligence Ware, способны выражать выявленные зависимости в виде правил на естественном языке. Современные средства интеллектуального анализа данных позволяют также определять переменные, оказывающие наибольшее влияние на значение заданных атрибутов. Например, анализируя медицинские данные о больных, получивших травму, можно выполнить автоматический выбор трех атрибутов, наиболее значимых для определения времени восстановления больного после травмы. В качестве таких атрибутов, например, могут быть выделены: “время до момента оказания квалифицированной помощи”, “возраст” и “физическое состояние больного”. Впоследствии именно эти признаки могут быть использованы для многомерного визуального анализа, например, при определении координатных осей.

Под прогнозированием, как правило, понимают процесс формирования вероятностного суждения о состоянии какого-либо объекта, процесса или явления в определенный момент времени в будущем. Методы прогнозирования основаны на принципе инерционности развития, т.е. предполагается, что развитие объекта подчинено определенным закономерностям, которые сохранятся на некоторый период в будущем. При прогнозировании используется способность методов интеллектуального анализа данных выявлять закономерности в исторических данных, описывающих развитие объекта, и использовать в дальнейшем эти тенденции для выработки гипотез о его состоянии в будущем. Особенно широко для предсказаний методы интеллектуального анализа данных применяют в финансовой сфере при прогнозировании доходности акций, курсов валют, экономических индикаторов.

В настоящее время компьютерные аналитические технологии данных переживают этап бурного развития - появляются новые программные продукты и задачи, которые успешно решаются с их помощью. Однако даже самые лучшие программные средства не заменят специалиста, способного провести интегральный анализ наблюдаемых явлений. Тем не менее современные интеллектуальные компьютерные технологии могут быть хорошим помощником аналитика, в значительной мере упрощая ему работу.

Архитектура OLAP-систем. OLAP-система включает в себя два основных компонента:

♦ OLAP-сервер - обеспечивает хранение данных, выполнение над ними необходимых операций и формирование многомерной модели на концептуальном уровне. В настоящее время OLAP-серверы объединяют с хранилищем данных;

♦ OLAP-клиент - представляет пользователю интерфейс к многомерной модели данных, обеспечивая его возможностью удобно манипулировать данными для выполнения задач анализа.

♦ OLAP-серверы скрывают от конечного пользователя способ реализации многомерной модели. Они формируют гиперкуб, с которым пользователи посредством OLAP-клиента выполняют все необходимые манипуляции, анализируя данные. Между тем способ реализации очень важен, так как от него зависят такие характеристики, как производительность и занимаемые ресурсы. Выделяют три основных способа реализации:

1) MOLAP - для реализации многомерной модели используют многомерные БД;

2) ROLAP - для реализации многомерной модели используют реляционные БД;

3) HOLAP - для реализации многомерной модели используют и многомерные, и реляционные БД.

Часто в литературе по OLAP-системам можно встретить аббревиатуры DOLAP и JOLAP.

DOLAP - это настольный (desktop) OLAP. Является недорогой и простой в использовании OLAP-системой, предназначенной для локального анализа и представления данных, которые загружаются из реляционной или многомерной БД на машину клиента.

JOLAP - новая, основанная на Java, коллективная OLAP-API-инициатива, предназначенная для создания и управления данными и метаданными на серверах OLAP. Основной разработчик - Hyperion Solutions. Другими членами группы, определяющей предложенный API, являются компании IBM, Oracle и др.

MOLAP-серверы используют для хранения и управления данными многомерные БД. При этом данные хранятся в виде упорядоченных многомерных массивов. Такие массивы подразделяются на гиперкубы и поликубы.

В гиперкубе все хранимые в БД ячейки имеют одинаковую мерность, т.е. находятся в максимально полном базисе измерений.

В поликубе каждая ячейка хранится с собственным набором измерений, и все связанные с этим сложности обработки перекладываются на внутренние механизмы системы.

Очевидно, что физически данные, представленные в многомерном виде, хранятся в "плоских" файлах. При этом куб представляется в виде одной плоской таблицы, в которую построчно вписываются все комбинации членов всех измерений с соответствующими им значениями мер.

Основными составляющими таких схем являются денормализованная таблица фактов (Fact Table) и множество таблиц измерений (Dimension Tables).

Таблица фактов, как правило, содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться. Обычно говорят о четырех наиболее часто встречающихся типах фактов. К ним относятся:

♦ факты, связанные с транзакциями (Transaction facts). Они основаны на отдельных событиях (типичными примерами которых являются телефонный звонок или снятие денег со счета с помощью банкомата);

♦ факты, связанные с "моментальными снимками" (Snapshot facts). Основаны на состоянии объекта (например, банковского счета) в определенные моменты времени, например на конец дня или месяца. Типичными примерами таких фактов являются объем продаж за день или дневная выручка;

♦ факты, связанные с элементами документа (Line-item facts). Основаны на том или ином документе (например, счете за товар или услуги) и содержат подробную информацию об элементах этого документа (например, количестве, цене, проценте скидки);

♦ факты, связанные с событиями или состоянием объекта (Event or state facts). Представляют возникновение события без подробностей о нем (например, просто факт продажи или факт отсутствия таковой без иных подробностей).

Таблица фактов, как правило, содержит уникальный составной ключ, объединяющий первичные ключи таблиц измерений. При этом как ключевые, так и некоторые не ключевые поля должны соответствовать измерениям гиперкуба. Помимо этого, таблица фактов содержит одно или несколько числовых полей, на основании которых в дальнейшем будут получены агрегатные данные.

Для многомерного анализа пригодны таблицы фактов, содержащие как можно более подробные данные, т.е. соответствующие членам нижних уровней иерархии соответствующих измерений. В таблице фактов нет никаких сведений о том, как группировать записи при вычислении агрегатных данных. Например, в ней есть идентификаторы продуктов или клиентов, но отсутствует информация о том, к какой категории относится данный продукт или в каком городе находится данный клиент. Эти сведения, в дальнейшем используемые для построения иерархий в измерениях куба, содержатся в таблицах измерений.

Таблицы измерений содержат неизменяемые либо редко изменяемые данные. В подавляющем большинстве случаев эти данные представляют собой по одной записи для каждого члена нижнего уровня иерархии в измерении. Таблицы измерений также содержат как минимум одно описательное поле (обычно с именем члена измерения) и, как правило, целочисленное ключевое поле (обычно это суррогатный ключ) для одно-

значной идентификации члена измерения. Если измерение, соответствующее таблице, содержит иерархию, то такая таблица также может содержать поля, указывающие на "родителя" данного члена в этой иерархии. Каждая таблица измерений должна находиться в отношении "один-ко-многим" с таблицей фактов.

Скорость роста таблиц измерений должна быть незначительной по сравнению со скоростью роста таблицы фактов; например, новая запись в таблицу измерений, характеризующую товары, добавляется только при появлении нового товара, не продававшегося ранее.

Классификация задач Data Mining. Методы Data Mining помогают решить многие задачи, с которыми сталкивается аналитик. Из них основными являются: классификация, регрессия, поиск ассоциативных правил и кластеризация. Ниже приведено краткое описание основных задач анализа данных.

Задача классификации сводится к определению класса объекта по его характеристикам. Необходимо заметить, что в этой задаче множество классов, к которым может быть отнесен объект, заранее известно.

Задача регрессии, подобно задаче классификации, позволяет определить по известным характеристикам объекта значение некоторого его параметра. В отличие от задачи классификации значением параметра является не конечное множество классов, а множество действительных чисел.

При поиске ассоциативных правил целью является нахождение частых зависимостей (или ассоциаций) между объектами или событиями. Найденные зависимости представляются в виде правил и могут быть использованы как для лучшего понимания природы анализируемых данных, так и для предсказания появления событий.

Задача кластеризации заключается в поиске независимых групп (кластеров) и их характеристик во всем множестве анализируемых данных. Решение этой задачи помогает лучше понять данные. Кроме того, группировка однородных объектов позволяет сократить их число, а следовательно, и облегчить анализ.

Необходимо понимать, что концепция хранилища данных (ХД):

♦ это не концепция анализа данных, скорее, это концепция подготовки данных для анализа;

♦ не предопределяет архитектуру целевой аналитической системы. Она говорит о том, какие процессы должны выполняться в системе, но не о том, где конкретно и как они будут выполняться.

Таким образом, концепция ХД определяет лишь самые общие принципы построения аналитической системы и в первую очередь сконцентрирована на свойствах и требованиях к данным, но не на способах их организации и представления в целевой БД и режимах их использования. ХД - это концепция построения аналитической системы, но не концепция ее использования. Она не решает ни одну из следующих проблем:

♦ выбор наиболее эффективного для анализа способа организации данных;

♦ организация доступа к данным;

♦ использование технологии анализа.

Проблемы использования собранных данных решают подсистемы анализа, которые применяют такие технологии, как:

♦ регламентированные запросы;

♦ оперативный анализ данных;

♦ интеллектуальный анализ данных.

Если регламентированные запросы успешно применялись еще задолго до появления концепции хранилищ данных, то оперативный и интеллектуальный анализ в последнее время все больше связывают с хранилищами данных.

В заключение можно отметить, что концепция хранилищ данных не является законченным архитектурным решением СППР и тем более не является готовым программным продуктом. Цель концепции ХД - определить требования к данным, помещаемым в ХД, общие принципы и этапы построения ХД, основные источники данных, дать рекомендации по решению потенциальных проблем, возникающих при их выгрузке, очистке, согласовании, транспортировке и загрузке.

---

[1] См.: *Акофф Р.* Планирование будущего корпораций. М., 1985; *Лотов А.В.* Введение в экономико-математическое моделирование. М., 1984; *Банди Б.* Методы оптимизации. Вводный курс : пер. с англ. М., 1988.

[2] См.: *Гаврилов А.В.* Системы искусственного интеллекта : учеб. пособие. Ч. 1. Новосибирск, 2000, 2001; *Баронов В.В.* Автоматизация управления предприятием. М., 2000; *Варфоломеев В.И., Назаров С.В.* Алгоритмическое моделирование элементов экономических систем. М., 2004.

[3] *Русанова Е.С.* Реинжиниринг на основе информационных технологий // Вестн. Самар. гос. экон. ун-та. Самара, 2011. № 4 (78). С. 71-77.

Поступила в редакцию 14.03.2012 г.