но-хозяйственной деятельности предприятия; используется как комплексное средство реорганизации предприятия или отдельных ее организационных единиц.

ЛИТЕРАТУРА

1. *Дун И.* Реинжиниринг: опережающее решение // Рынок ценных бумаг. 1998. № 6.
2. *Евсеев О.* Динамическое моделирование и реинжиниринг бизнес-процессов // Рынок ценных бумаг. 1998. № 6.
3. Hammer H.M. «Reengineering Work: Don't Automate, Obliterate» // Harvard Business Review. July — August. 1990.
4. *Шеер А.В.* Бизнес-процессы: основные понятия, теории, методы. М., 1999.

ТЕХНОЛОГИИ И ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА АНАЛИТИЧЕСКИХ СИСТЕМ

В.В. Лабоцкий
канд. техн. наук, доцент кафедры ИТУ ВШУБ

К концу XX в. дисбаланс между объемом информации и возможностями аналитиков резко увеличился. Поток и объем информации стал больше, в то время как способы их анализа остались прежними.

Возникшие в самом конце XX в. аналитические системы, отчасти способны снять остроту этой проблемы.

Аналитические системы позволяют решать следующие задачи: ведение отчетности, анализ информации в реальном времени (OLAP) и интеллектуальный анализ данных.

Сервис *отчетности* помогает организации справиться с созданием всевозможных информационных отчетов, справок, документов, сводных ведомостей, особенно когда число выпускаемых отчетов велико и их формы часто меняются. Эти средства, автоматизируя выпуск отчетов, позволяют перевести их хранение в электронный вид и распространять по корпоративной сети между служащими компании.

OLAP-сервис (On-Line Analitycal Processing) представляет собой инструмент для анализа больших объемов данных в режиме реального времени. Взаимодействуя с OLAP-системой, пользователь может осуществлять гибкий просмотр информации, получать произвольные срезы данных и выполнять аналитические операции детализации, свертки, сквозного распределения, сравнения во времени. Вся работа с OLAP-системой происходит в терминах предметной области. OLAP-системы являются частью более общего понятия Business Intelligence, которое включает в себя помимо традиционного OLAP-сервиса средства организации совместного исполь-

зования документов, возникающих в процессе работы пользователей хранилища. Технология Business Intelligence обеспечивает электронный обмен отчетными документами, разграничение прав пользователей, доступ к аналитической информации из сетей Интернет и Интранет.

При помощи *интеллектуального анализа данных* (*Data Mining*) можно проводить глубокие исследования данных. Эти исследования включают в себя поиск зависимостей между данными (например, «Верно ли, что рост продаж продукта А обусловлен ростом продаж продукта В?»); выявление устойчивых бизнес-групп (например, «Какие группы клиентов, близкие по поведенческим и другим характеристикам, можно выделить? Какие характеристики клиентов при этом оказывают наибольшее влияние на классификацию?»), прогнозирование поведения бизнес-показателей (например, «Какой объем перевозок ожидается в следующем месяце?»), оценка влияния решений на бизнес компании (например, «Как изменится спрос на товар А среди группы потребителей Б, если снизить цену на товар С?»); поиск аномалий (например, «С какими сегментами клиентской базы связаны наиболее высокие риски?»).

Что такое Data Mining?

Data Mining переводится с английского как «добыча» или «раскопка данных». Нередко рядом с Data Mining встречаются синонимичные выражения «обнаружение знаний в базах данных» (knowledge discovery in databases — KDD) и «интеллектуальный анализ данных» (ИАД).

До начала 1990-х гг. не было особой нужды переосмысливать ситуацию в этой области. Все шло в рамках направления, называемого прикладной статистикой. Вместе с тем, практики всегда знали, что попытки применить теорию для решения реальных задач обработки небольших локальных баз данных в большинстве случаев оказываются бесплодными.

В связи с совершенствованием технологий записи и хранения данных на людей обрушились колоссальные потоки информации. Деятельность любого предприятия (коммерческого, производственного, медицинского, научного и т.д.) теперь сопровождается регистрацией и записью всех подробностей его деятельности. Что делать с этой информацией? Стало ясно, что без продуктивной переработки в потоках данных сложно разобраться.

Специфика современных требований к такой переработке следующая: данные имеют неограниченный объем, они являются разнородными (количественными, качественными, текстовыми), результаты должны быть конкретны и понятны, инструменты для обработки новых данных должны быть просты в использовании.

Традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, откровенно спасовала перед лицом возникших проблем. Главная причина — концепция ус-

реднения по выборке, приводящая к операциям над фиктивными величинами (например, средняя температура тела пациентов по больнице и т.п.). Методы математической статистики оказались полезными для проверки заранее сформулированных гипотез и для «грубого» разведочного анализа, составляющего основу оперативной аналитической обработки данных OLAP.

В основу современной технологии Data Mining положена концепция шаблонов (pattern), отражающая фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной форме. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей. Примеры задач с использованием методов OLAP и Data Mining приведены в табл. 1.

Таблица 1

Примеры задач с использованием методов OLAP и Data Mining

OLAP	Data Mining
Каковы средние показатели травматизма для курящих и некурящих?	Встречаются ли точные шаблоны в описаниях людей, подверженных повышенному травматизму?
Каковы средние размеры телефонных счетов существующих клиентов в сравнении со счетами бывших клиентов (отказавшихся от услуг телефонной компании)?	Имеются ли характерные портреты клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании?
Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке?	Существуют ли стереотипные схемы покупок для случаев мошенничества с кредитными карточками?

Важное положение Data Mining — нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать неочевидные, неожиданные (unexpected) регулярности в данных, составляющие так называемые скрытые знания (hidden knowledge). К обществу пришло понимание, что сырые данные (raw data) содержат глубинный пласт знаний, при грамотной раскопке которого могут быть обнаружены настоящие самородки.

Data Mining (по определению Г. Лиатецкого-Шапиро) — исследование и обнаружение алгоритмами, средствами искусственного интеллекта в сырых данных скрытых структур, шаблонов или зависимостей, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком и необходимы для принятия решений в различных сферах деятельности.

При исследовании данных средствами Data Mining используется большое число различных методов и их различные комбинации. К наиболее важным и часто используемым методам относятся:
кластеризация; ассоциация; деревья решений; анализ с избирательным действием; сети уверенности; метод ближайших соседей; нейронные сети; нечеткая логика; генетические алгоритмы; регрессионные методы; эволюционное программирование.

Верификация и оценка полезности моделей

Автоматическая проверка достоверности добытого знания, оценка статистической значимости построенных моделей является необходимым и очень важным моментом любого исследования. Проверка значимости определенных зависимостей, моделей может проводиться стандартными статистическими методами (например, Стентон, 1999).

При исследовании статистической надежности результата наиболее часто основываются на значениях стандартного отклонения и стандартной ошибки. Если значение зависимой переменной для i-й записи p_i ($1 \le i \le N$), а значение этой же переменной, предсказанное найденной моделью P_i, то стандартное отклонение определяется по следующей формуле:

$$S_{dev} = \sqrt{\frac{\sum_{i=1}^{N}(p_i - P_i)}{N-1}},$$

где N — число записей, для которых посчитана регрессионная модель.

Стандартная ошибка, предсказанная данной моделью, переменной определяется по формуле

$$S_{err} = \sqrt{\frac{\sum_{i=1}^{N}(p_i - P_i)}{(N-1)\sigma}},$$

где σ — квадрат дисперсии значений p_i.

Квадрат дисперсии рассчитывается по следующей формуле:

$$\sigma = \frac{\sum_{i=1}^{N}(p_i - \bar{p})}{N-1},$$

где $\bar{p}$ — среднее значение этой переменной.

Другими словами, стандартная ошибка — это стандартное отклонение деленное на дисперсию.

Наиболее значимая, или наиболее точная модель, найденная каким-либо из методов Data Mining — модель, обладающая наименьшим значением стандартного отклонения среди всех найденных моделей.

Значимость — мера вероятности того, что найденная зависимость «истинна» и действительно характеризует исследуемые данные. В качестве такой меры часто используется отрицательный логарифм вероятности того, что зависимость выведена случайно и является результатом статистической флуктуации в данных. Чем ближе значение последней вероятности к нулю, тем выше значимость. Для вычисления *индекса значимости* сравнивается стандартное отклонение результата, основанное на реальных данных, со стандартным отклонением результата, полученным для искусственно созданных данных, в которых значения целевых переменных для разных записей случайным образом перемешаны. Если стандартное отклонение, полученное на реальных данных s_{real}, приблизительно равно стандартному отклонению случайных данных s_{rand}, то индекс значимости близок к единице. В этом случае результат исследования не может рассматриваться как значимый. Если s_{real} много меньше, чем s_{rand} для всех случайно генерированных таблиц, то индекс значимости намного больше единицы. Поэтому результат исследования может быть назван значимым. На практике имеет смысл называть значимыми только те модели, у которых индекс значимости больше двух.

Коэффициент корреляции r ($1 \geq r \geq -1$) между значениями непрерывной целевой переменной и предсказываемыми значениями также характеризует значимость обнаруженной регрессионной модели. Чем ближе значение r^2 к единице, тем больше значимость модели; если значения r^2 небольшие, вблизи нуля, модель может быть отвергнута. Часто бывает полезно составить визуальное представление о значимости модели, построив график зависимости предсказываемых значений целевой переменной от их реальных значений. Если бы предсказываемые значения совпадали с реально наблюдаемыми значениями, мы бы получили прямую, идущую под 45 ° к оси абсцисс. Чем ближе к этой прямой ложатся точки на графике, тем точнее модель описывает данные. Если мы получим примерно сферическое облако, модель смело можно выбрасывать в корзину.

Для оценки значимости коэффициентов регрессионных моделей часто используется t-критерий (критерий Стьюдента), определяемый как частное модуля соответствующего регрессионного коэффициента на величину его стандартного отклонения. Выполняя процесс линейной регрессии, система может тестировать линейно-регрессионные модели, включающие в себя различные комбинации независимых переменных. Для каждой модели определяются значения ее регрессионных коэффициентов, их стандартных отклонений и t-критерии. Если коэффициент регрессионной модели характеризуется значением t-критерия меньшим, чем заданное критическое значение, то он отвергается. Использовать значение t-критерия меньшее корня квадратного из двух, как правило, нецелесообразно.

При оценке статистической значимости модели вычисляется некоторая мера ее точности, но определенное значение значимости применимо только к тем данным, на которых построена модель, данные, к которым в дальнейшем будет применяться модель, могут отличаться от исходных непредсказуемым образом.

Инструментальные средства аналитических систем

В настоящее время можно приобрести достаточно большое число отдельных аналитических систем, совмещенных с хранилищем данных (Data Warehouse — DW). Практически все системы поддерживают архитектуру клиент-сервер, предоставляя пользователям возможность выполнять наиболее трудоемкие процедуры обработки данных на высокопроизводительном сервере. Следует отметить, что в основном это разработки западных компаний из США. Российская компания «Мегапьютер Интеллидженс» прорвалась на этот сегмент рынка во многом благодаря поддержке нетрадиционного для коммерческих систем метода Data Mining — эволюционного программирования, предоставляющего аналитику уникальные возможности. PolyAnalyst является одной из самых мощных систем Data Mining в мире, разработанных для Intel-платформ и операционных систем Microsoft Windows, с относительно низкой стоимостью. Аналогичные системы Data Mining таких ведущих производителей, как IBM (Intelligent Miner, Data Miner), Silicon Graphics (SGI Miner), Integral Solutions (Clementine), SAS Institute (SAS) работают на средних и больших машинах и стоят от десятков до сотен тысяч долларов. Примерами нейросетевых систем являются BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic) их стоимость 1500—8000 дол.

В табл. 2 приводятся адреса наиболее распространенных и интересных аналитических систем. В табл. 3, 4, 5 приводятся краткие сравнительные характеристики этих систем. В табл. 6 представлен мировой рейтинг по продуктам Data Mining за 2001—2002 гг.

Таблица 2

WEB-сайты аналитических систем

Аналитическая система	Адрес в сети Интернет
IBM DB2 Intelligent Miner	www.ibm.com/
SAS Interprise Miner	www.sas.com/
MineSet	www.sgi.com/software/mineset/
Knowledge STUDIO	www.angoss.com/
Kepler Dialogis	www.dialogis.de
Clementine	www.isl.co.uk/
PolyAnalyst	www.megaputer.com www.megaputer.ru

С другими системами можно ознакомиться в Интернет по адресу http://www.kdnuggets.com/ http://www.data-miners.com/.

Таблица 3

Сравнительная характеристика аналитических систем

Продукт, производитель, страна	Платформа, сервер	Платформа, клиент	Доступ к данным	Объем данных
IBM DB2 Intelligent Miner V6R1 IBM International	AIX Sun Solaris Win NT	AIX Win95/98/NT/ 2000/ OS/2	DB2, IBM Visual Warehous SAP Business Warehouse Другие источники через DB2 DataJoiner	> Gb
SAS Interprise Miner SAS Institute USA	AIX Sun Solaris Win NT HP-UX, ALX, Digital Compaq UNIX	AIX Sun Solaris Win NT HP-UX, ALX, Digital Compaq UNIX OS/2	SAS Warehouse, Компоненты SAS, Доступ к внешним источникам через модуль SASbase	>Gb
MineSet Silicon Graphics USA	IRIX Windows NT	IRIX Windows95/98/ 2000/NT	Flat files: Excel, CSV, dBF, SAS, SPSS ODBC: Oracle MSSQL, Informix, DB2	>Gb
Knowledge STUDIO Angoss Software Canada	Windows NT	Windows95/98/ 2000/NT	Excel, CSV, dBF, ODBC	>Gb
Kepler Dialogis Germany	UNIX WindowsNT	Windows95/98/NT/ 2000	Flat files: Excel, CSV, dBF ODBC: Oracle MSSQL, Informix	>Gb
Clementine Integral Solutions UK	Unix WinNT Sun Solaris	Windows95/98/ 2000/NT	Excel, SPSS, ODBC	>100Mb
PolyAnalyst Мегапьютер Интеллидженс Россия	WinNT OS/2	Windows95/98/NT/ 2000 OS/2	Flat files: Excel, CSV, dBF ODBC: Oracle Sybase MSSQL, Informix, DB2, Direct: Oracle Express, IBM VW	>Gb

Таблица 4

Сравнительные характеристики систем в использовании технологий

Технология	IBM Intelligent Miner IBM International	SAS Enterprise Miner 3.0 SAS Institute USA	MineSet Silicon Graphics USA	Knowledge STUDIO Angoss Software Canada	Kepler Dialogis Germany	Clementine 5.0 Integral Solutions GB	PolyAnalyst 4.0 Мегапьютер Интеллидженс Россия
Деревья решений	Да	Да	Да	Да	Да	Да	Да
Нейронные сети	Да	Да	Нет	Да	Да	Да	Да
Генетические алгоритмы	Да	Да	Нет	Да	Да	Да	Да
Эволюционное программирование	Нет	Нет	Нет	Нет	Нет	Нет	Да
K-Nearest Neibours	Да	Да	Нет	Нет	Нет	Да	Да
Кластеризация	Да	Да	Да	Нет	Нет	Да	Да
Нечеткая Классификация	Нет	Да	Нет	Нет	Нет	Да	Да
Регрессия	Да	Да	Нет	Да	Нет	Да	Да
Ассоциация	Да	Да	Да	Нет	Да	Нет	Да
Индукция правил	Нет	Нет	Нет	Нет	Да	Нет	Нет
Многомерный Анализ	Нет	Да	Да	Нет	Нет	Нет	Да
Визуализация	Да	Да	Да	Да	Да	Да	Да

Таблица 5

Мировой рейтинг продуктов Data Mining (2001 г.) по данным http://www.kdnuggets.com

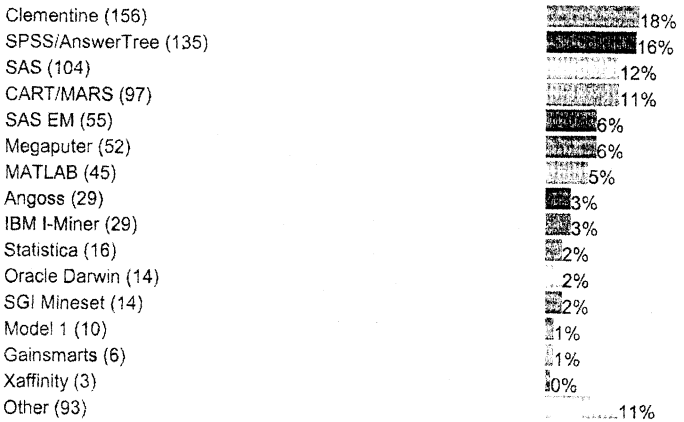
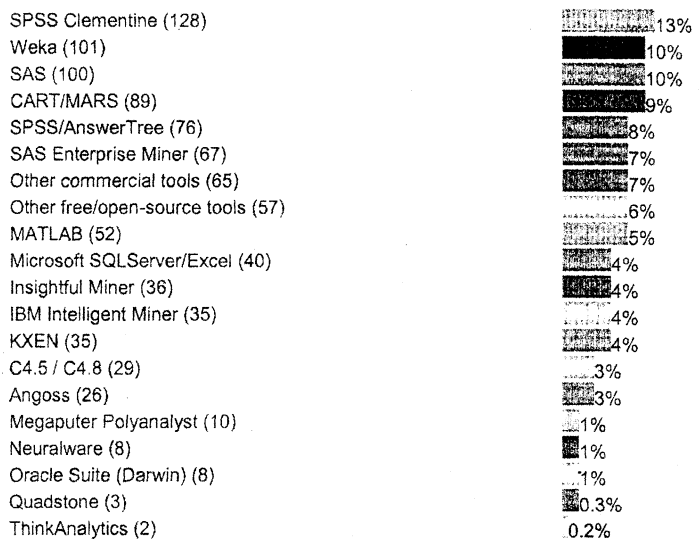


Таблица 6

Мировой рейтинг продуктов Data Mining (2002 г.) по данным http://www.kdnuggets.com



ЛИТЕРАТУРА

1. *Дюк В., Самойленко А.* Data Mining: Учеб. курс. СПб., 2001.