СПИСОК ЛИТЕРАТУРЫ

1. Berg Insight: Strategic Analysis of the European Mobile LBS Market (Report in LBS Research Series) [Электронный ресурс]. – Режим доступа: http://www.berginsight.com/ShowReport.aspx?m_m=3&id=44. – 20.04.2009.
2. O'Reilly T. What Is Web 2.0 [Электронный ресурс]. – Режим доступа: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. – 20.04.2009.
3. Wind R., Jensen C., Pedersen K., Torp K. A Testbed for the Exploration of Novel Concepts in Mobile Service Delivery // Proc. of Mobile Data Management Int. Conf. – Mannheim, Germany, 2007. – P. 218–220.
4. Kupper A. Location-Based Services: Fundamentals and Operation. – Chichester: John Wiley & Sons Ltd, 2005. – P. 365.
5. Simmons R. et al. Learning to Predict Driver Route and Destination Intent. // Proc. of IEEE Intelligent Transportation Systems Conf. – Toronto, 17-20 Sept., 2006. – P. 127–132.
6. Froehlich J., Krumm J. Route Prediction from Trip Observations. //Society of Automotive Engineers World Congress. Paper 2008-01-0201. – Detroit, 22 April, 2008. – P. 103–117.
7. Froehlich J., Krumm J. The Microsoft Multiperson Location Survey. MSR-TR-2005-103 [Электронный ресурс]. – Режим доступа: ftp://ftp.research.microsoft.com/pub/tr/TR-2005-103.doc. – 20.04.2009.
8. Brilingaite A., Jensen C. Enabling Routes of Road Network Constrained Movements as Mobile Service Context // Geoinformatica. – 2007. – V. 11. – № 1. – P. 55–102.
9. Brilingaite A., Jensen C. Online Route Prediction for Automotive Applications // Proc. of the 13th World Congress and Exhibition on Intelligent Transport Systems and Services. – London, October, 2006. – P. 168–175.
10. Welcome to OpenStreetMap [Электронный ресурс]. – Режим доступа: http://wiki.openstreetmap.org/wiki/Main_Page. – 20.04.2009.

Поступила 21.04.2009 г.

УДК 004.89

ИСПОЛЬЗОВАНИЕ KPI, ТЕХНОЛОГИЙ OLAP И DATA-MINING ПРИ ОБРАБОТКЕ ДАННЫХ

А.Р. Вахитов

Томский политехнический университет
E-mail: var-sasha@tpu.ru

Рассматривается способ обработки данных, основанный на совместном использовании аналитической обработки в реальном времени, а также ключевых индикаторов производительности и технологии извлечения данных. Обсуждаются принципы реализации способа, области применения, базовые термины, а также преимущества по сравнению с классическими способами решения подобных задач. Особое внимание уделяется практическому применению данного подхода в предметной области, связанной с НИРС в вузе.

Ключевые слова:
OLAP, обработка данных, data mining, ключевые индикаторы производительности.

В современном мире особую ценность приобретают эффективные способы обработки информации. Базы данных (БД), а также системы управления этими базами (СУБД) стали необходимыми в любой организации. Учебные заведения, банки, страховые, коммерческие и прочие компании собирают и хранят в своих базах гигабайты информации о сотрудниках, предоставляемых услугах, товарах и т. д. Ценность подобных сведений несомненна: они используются в различных целях (управление материально-техническими запасами, решение вопросов, связанных с перераспределением полномочий, отслеживание тенденций развития организации и другое).

Подобные БД называют операционными или транзакционными, поскольку они характеризуются огромным количеством небольших транзакций (операций записи-чтения). Компьютерные системы, осуществляющие учет операций и, собственно, доступ к транзакционным базам, принято называть системами оперативной обработки транзакций Online Transactional Processing (OLTP) или учетными системами [1].

Учетные системы настраиваются и оптимизируются для выполнения максимального количества транзакций за максимально короткое время. Показателем эффективности является количество транзакций, выполняемых за секунду. Обычно операции над отдельными записями очень просты и не связаны друг с другом. Однако совокупности записей можно использовать для получения качественно новой информации, а именно для создания отчетов и анализа деятельности организации.

Набор аналитических функций в учетных системах обычно весьма ограничен. Схемы, используемые в OLTP-приложениях, осложняют создание даже простых отчетов, так как данные чаще всего распределены по множеству таблиц, и для их агрегирования необходимо выполнять сложные операции объединения. Как правило, попытки создания комплексных отчетов требуют больших вычислительных мощностей и приводят к потере производительности [1].

Уместно также отметить, что в учетных системах хранятся постоянно изменяющиеся данные.

По мере осуществления операций записи-чтения суммарные значения меняются очень быстро, и два комплексных анализа, проведенных с интервалом в несколько минут, могут дать разные результаты, поэтому, чаще всего, анализ выполнятся по окончании отчетного периода, иначе картина может оказаться искаженной.

Приведенными выше соображениями объясняется переход к объединению и анализу данных учетной системы с помощью технологии Online Analytical Processing (OLAP). Этот метод позволяет аналитикам, менеджерам и руководителям проанализировать накопленные данных за счет быстрого и согласованного доступа к широкому спектру представлений информации.

Методология OLAP – это аналитическая обработка в реальном времени (технология обработки информации, включающая составление и динамическую публикацию отчётов и документов), предназначенная для быстрой обработки сложных многотабличных запросов к БД.

Причины использования OLAP для обработки запросов – это скорость и удобство. Реляционные БД хранят сущности в отдельных таблицах, которые обычно хорошо нормализованы. Эта структура удобна для операционных БД (систем OLTP), но сложные многотабличные запросы, обрабатывающие множество строк, в ней выполняются относительно медленно. Кроме того, в этой структуре данные сложно анализировать. OLAP-технология значительно упрощает анализ за счет использования многомерных кубов представления данных. Просматривая сводные таблицы, пользователь видит сначала итоговые значения показателей, и, при необходимости, может их легко детализировать. Клиент-серверная архитектура OLAP-продуктов обеспечивает одновременный доступ большого числа пользователей. При этом анализ производится одинаково быстро по всем аспектам информации независимо от размера и сложности структуры БД.

В качестве объекта исследования была использована БД, содержащая информацию о НИРС, имеющая следующую схему данных в СУБД Microsoft SQL Server 2008, рис. 1.

На основе этой БД была построена OLAP-структура, содержащая рабочие данные и представляющая из себя OLAP-куб, рис. 2.

Куб создаётся из соединения таблиц с применением схемы звезды. В центре звезды находится таблица фактов, которая содержит ключевые факты, по которым делаются запросы. Множественные таблицы с измерениями присоединены к таблице фактов. Эти таблицы показывают, как могут анализироваться агрегированные реляционные данные.

Заявленное время обработки запросов в OLAP составляет около 0,1 % от аналогичных запросов в реляционную БД [2]. В качестве примера был создано 2 одинаковых отчета, содержащих информацию о НИР определенного студента и преподавателя: в первом случае источником данных являлась реляционная БД, во втором – OLAP-модель. В

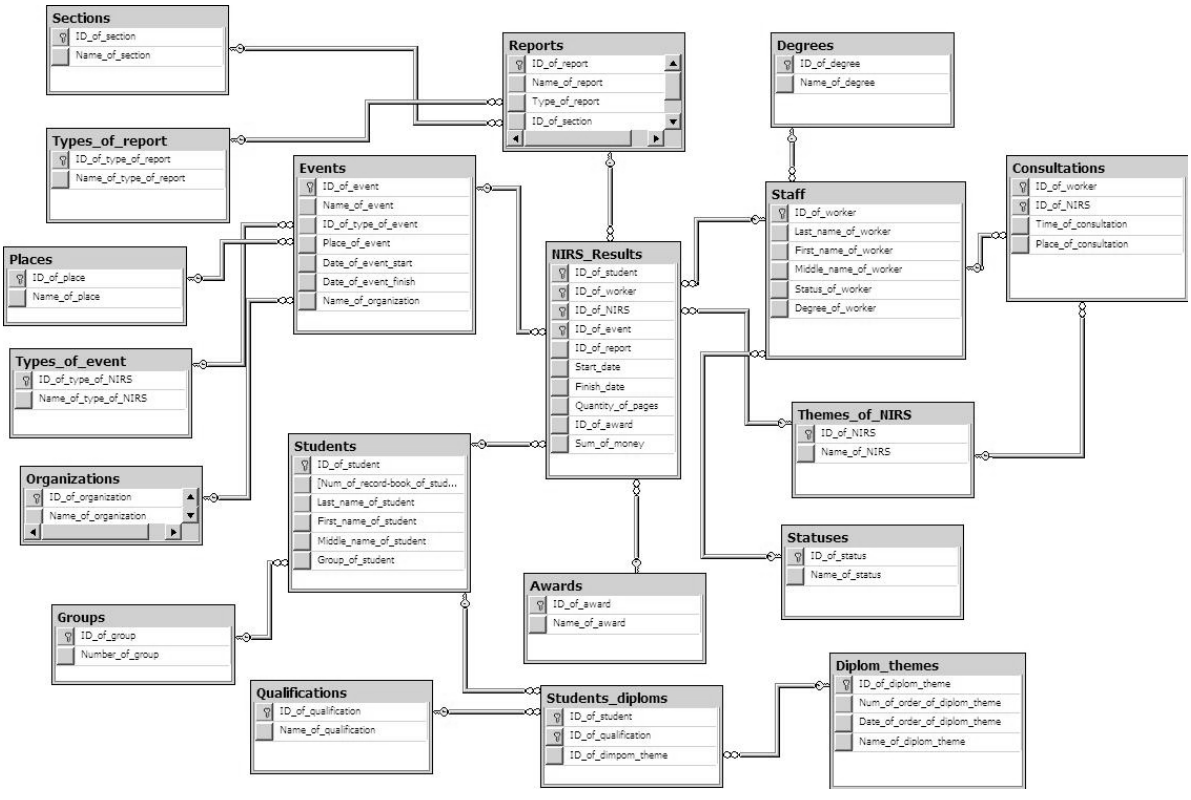

Рис. 1. *Схема данных о НИРС в СУБД Microsoft SQL Server 2008*

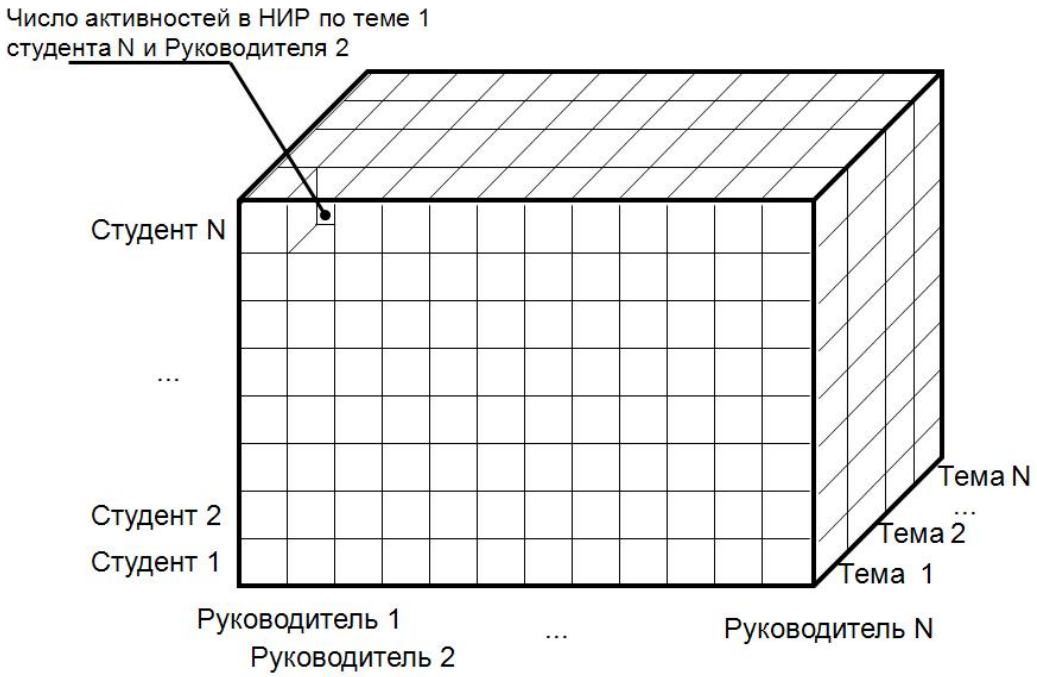

Рис. 2. *OLAP-куб, содержащий информацию о НИРС*

обоих случаях число записей в основной таблице равнялось 5000, аппаратная платформа была идентична. Время создания отчета на основе реляционной БД оказалось равным 15,1 с, а в случае, когда источником данных была OLAP-модель, – 1,33 с.

Действительно, время обработки запроса (создания отчета) на основе OLAP-модели оказалось в 11 раз меньше. Несомненным преимуществом OLAP-технологии является то, что конечный пользователь имеет возможность динамически изменять структуру запроса к базе данных. Например, на рис. 3 показан интерфейс системы, в правой части которого пользователь сам определяет строки и столбцы, которые ему необходимы, а соответствующие этой структуре данные динамически отображаются в левой части интерфейса.

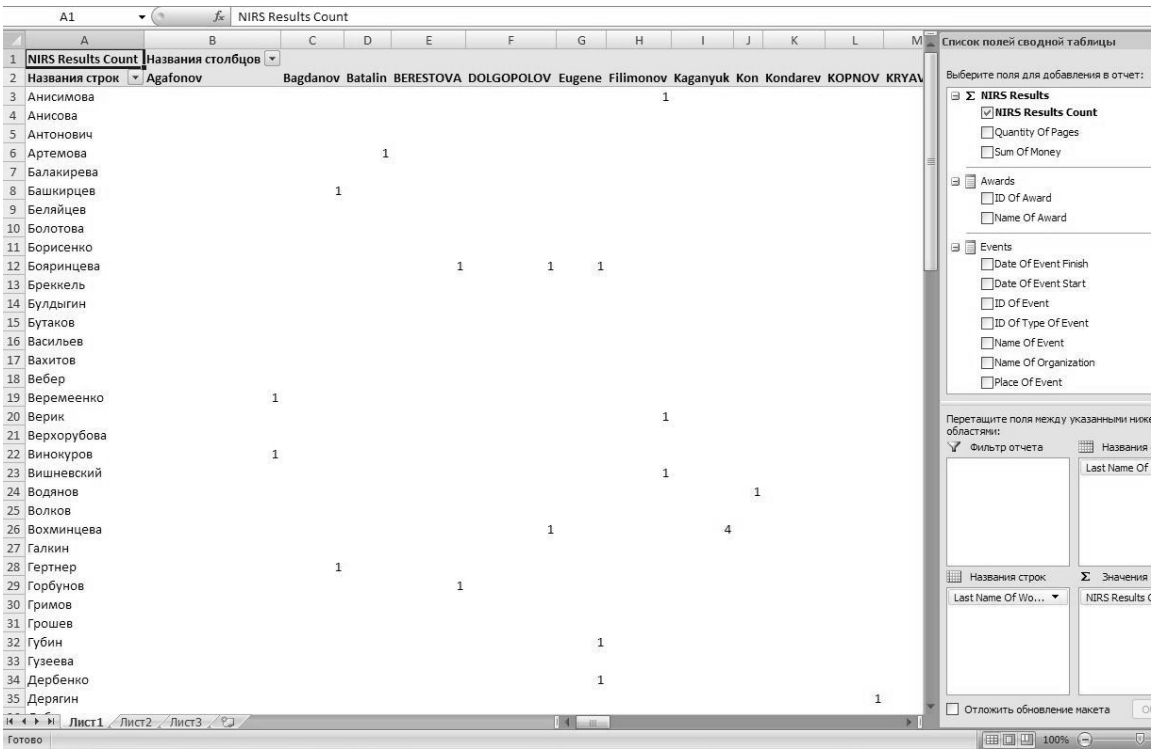

Рис. 3. *Интерфейс для отображения OLAP-куба*

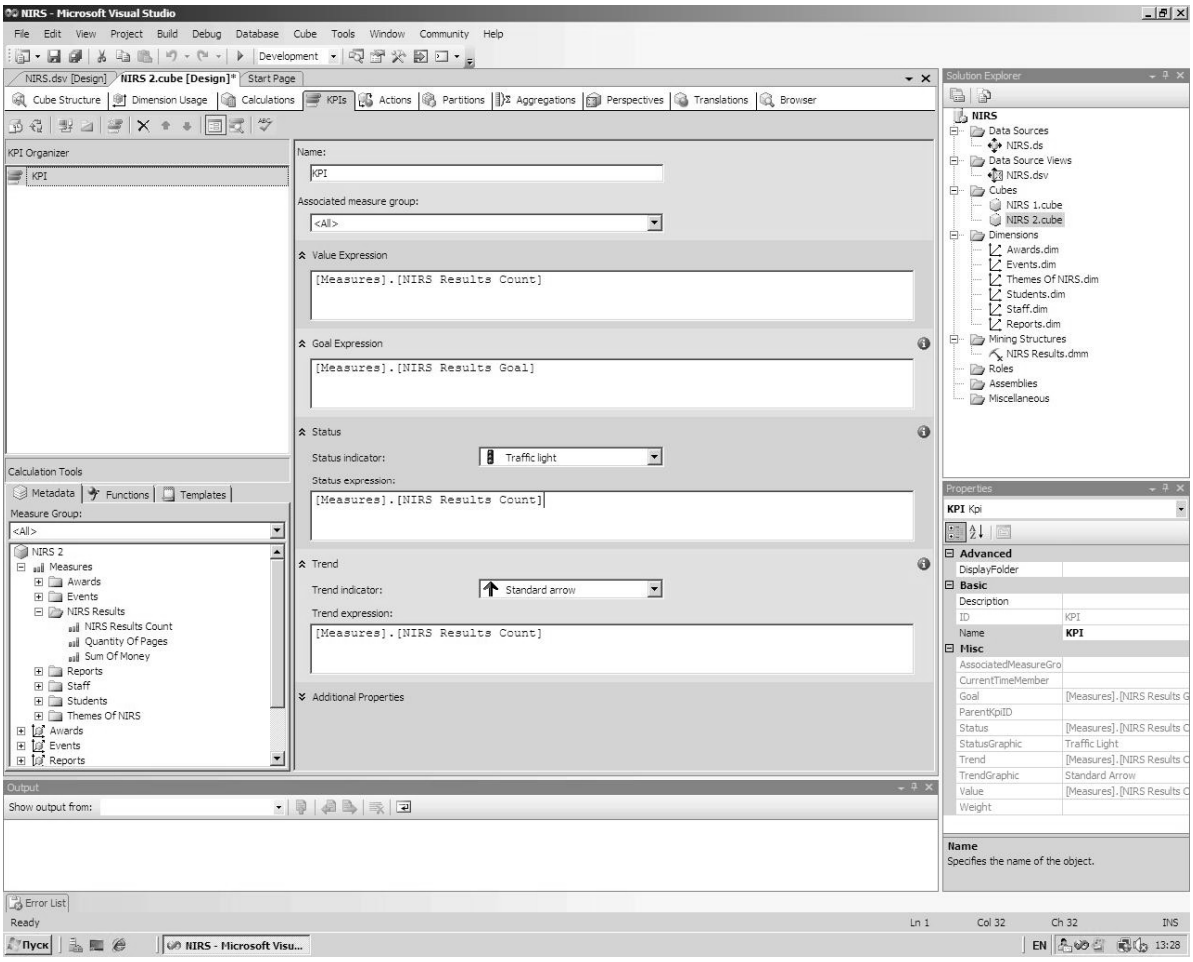

Рис. 4. *Создание KPI, отражающего общее число НИРС активностей*

А если учесть тот факт, что процесс преобразования реляционной схемы данных в OLAP достаточно прост и не занимает много времени, то преимущества OLAP становятся очевидными. Таким образом, проведенные исследования свидетельствуют о целесообразности преобразования реляционных моделей и использования OLAP-моделей при обработке данных.

OLAP-технология использовалась при работе с трендами и ключевыми индикаторами производительности – *key performance indicator* (KPI). KPI представляет собой ключевой индикатор производительности – систему оценки, которая помогает организации определить достижение стратегических целей. Их использование дает организации возможность оценить свое состояние и помочь в формировании стратегии развития. KPI позволяет производить контроль деловой активности в реальном времени. В исследуемой системе были выделены следующие KPI: общее число активностей по НИР определенного студента, число активностей по отдельным видам НИРС, число преподавателей, являющихся руководителями НИРС, число студентов, имеющих результаты НИР и др. Далее, на рис. 4 показан пример создания KPI, в котором задается мера для оценки общего числа НИРС активностей, выражение для расчета целевого значения, а также вид индикатора, который будет сигнализировать конечному пользователю о достижении стратегических целей, либо о том, что те или иные показатели деловой активности организации нуждаются в улучшении.

Тренд представляет собой выраженную направленность изменения показателей любого временного ряда. Графики могут быть описаны различными уравнениями – линейными, логарифмическими, степенными и т. д. Фактический тип графика устанавливается на основе графического изображения данных временного ряда, путем осреднения показателей динамики ряда, на основе статистической проверки гипотезы о постоянстве параметров графика. В дальнейшем эти данные используются для осуществления предсказательного анализа данных или *data mining*.

Data mining – выявление скрытых закономерностей или взаимосвязей между переменными в больших массивах необработанных данных. Английский термин «*data mining*» не имеет однозначного перевода на русский язык (добыча данных, вскрытие данных, информационная проходка, извлечение данных/информации), поэтому в большинстве случаев используется в оригинале. В рамках данного исследования технология *data mining*

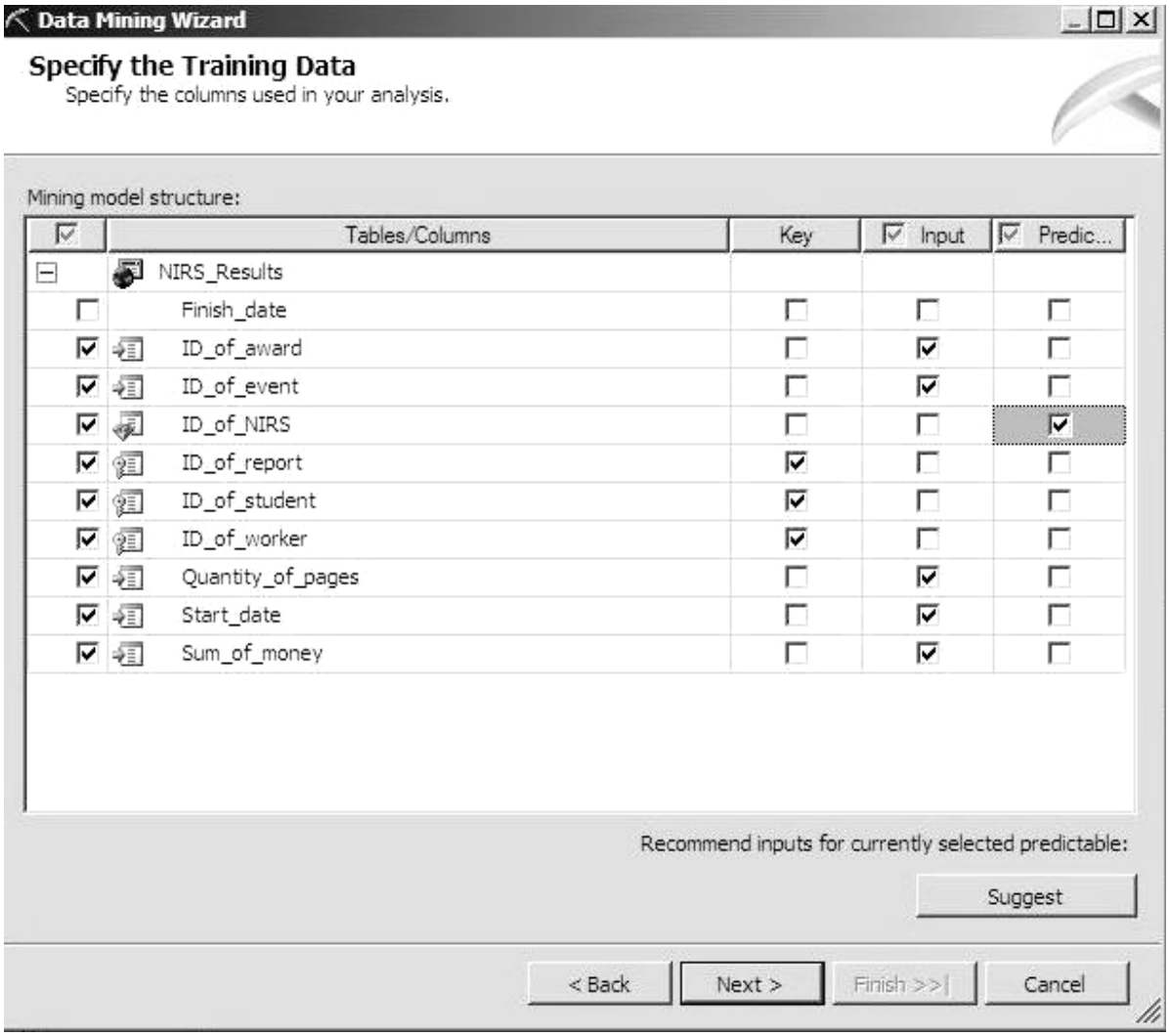

Рис. 5. *Использование data mining для определения вероятностных значений атрибутов*

позволяет доопределить недостающие данные в базе. Далее, на рис. 5 показан пример применения подобного анализа.

В таблице фактов OLAP-куба задается отсутствующий для некоторых записей атрибут, который необходимо определить: в данном случае это область науки, склонность заниматься которой есть у студента. Кроме того, задаются входные данные для расчета отсутствующего значения: сведения об участии других студентов в конференциях, научных конкурсах; успеваемость по определенным дисциплинам и т. д. В итоге получим сводную таблицу, в которой записям, у которых отсутствует атрибут «область НИРС», с определенной долей вероятности присваиваются значения соответствующего атрибута, имеющего сходные с искомым объектом входные параметры. Таким образом, с помощью однократной тренировки системы имеется возможность получить множество недостающих в БД сведений.

СПИСОК ЛИТЕРАТУРЫ

1. Brachman R., Sefridge P. Knowledge representation support for data archeology // Intelligent and Cooperative Information Systems. – 1993. – № 2. – P. 159–186.

2. Совместное использование учетных систем и технологии OLAP [Электронный ресурс]. – 2006. – режим доступа: http://www.cit-forum.ru/database/articles/olap_oltp.shtml. – 17.04.2009.

Поступила 17.04.2009 г.