Вальдемар Вуйцик, Сауле Смаилова, Индира Увалиева

ПРИМЕНЕНИЕ ТЕХНОЛОГИИ "DATA MINING" ДЛЯ АНАЛИЗА ДАННЫХ УЧЕБНОГО ПРОЦЕССА

В статье подчеркнуто, что основной задачей высших учебных заведений является обеспечение качественного образования. Один из способов достижения высокого уровня качества в системе высшего образования – обнаружение знаний, скрытых в базах учебных данных. Показаны возможности оперативного и интеллектуального анализа данных в контексте высшего образования. Реализован OLAP и интеллектуальный анализ данных учебного процесса методами факторного и кластерного анализа, выполнен также анализ влияния факторов; спроектирована архитектура информационно-аналитической системы.

Ключевые слова: *OLAP-анализ, интеллектуальный анализ данных, факторный анализ, кластерный анализ.*

Форм. 1. Табл. 1. Рис. 6. Лит. 10.

Вальдемар Вуйцик, Сауле Смаїлова, Індіра Увалієва

ЗАСТОСУВАННЯ ТЕХНОЛОГІЇ "DATA MINING" ДЛЯ АНАЛІЗУ ДАНИХ НАВЧАЛЬНОГО ПРОЦЕСУ*

У статті підкреслено, що основним завданням вищих навчальних закладів є забезпечення якісної освіти. Один із способів досягнення високого рівня якості в системі вищої освіти – виявлення знань, прихованих в базах навчальних даних. Показано можливості оперативного та інтелектуального аналізу даних в контексті вищої освіти. Реалізовано OLAP та інтелектуальний аналіз даних навчального процесу методами факторного і кластерного аналізу, здійснено також аналіз впливу чинників; спроектовано архітектуру інформаційно-аналітичної системи.

Ключові слова: *OLAP-аналіз, інтелектуальний аналіз даних, факторний аналіз, кластерний аналіз.*

Waldemar Wojcik[1], Saule Smailova[2], Indira Uvalieva[3]

APPLICATION OF DATA MINING TECHNOLOGY TO ANALYZE THE EDUCATIONAL PROCESS DATA

The paper argues that the key objective of higher educational institutions is to provide the quality of education. A way to achieve the high level of quality in the system of higher education is the discovery of knowledge, hidden in the training data bases. The possibility of operational and data mining in the context of higher education is identified. OLAP and data mining are implemented based on the methods of factor analysis, cluster analysis and impact analysis; the architecture of information and analytical system is designed.

Keywords: *OLAP-analysis, data mining, factor analysis, cluster analysis.*

Постановка проблемы. Проведение эффективной политики и реформ в сфере образования требует применения новых методов анализа для подготовки организационных и управленческих решений, адекватных современным задачам. В данной ситуации информационно-аналитическое обеспечение становится одним из главных «сервисов» в решении проблемы модернизации управления качеством образования [1].

* статтю підготовлено на основі доповіді на XII-му міжнародному науковому семінарі «Сучасні проблеми інформатики в управлінні, економіці, освіті та екології» (1–5 липня 2013 р., оз. Світязь – Київ).

[1] Institute of Electronics and Information Technology, Lublin University of Technology, Poland.

[2] D. Serikbaev East Kazakhstan state technical university, Ust-Kamenogorsk, Kazakhstan.

[3] D. Serikbaev East Kazakhstan state technical university, Ust-Kamenogorsk, Kazakhstan.

Анализ последних исследований и публикаций. В настоящее время методы "Data Mining" (интеллектуальный анализ) получили широкое распространение в различных сферах деятельности. Исследованиями в этой области занимаются такие ученые, как А.А. Барсегян [4], М.С. Куприянов [4], Г. Пятецкий-Шапиро [7], Х. Ромесбург [10], Дж. Хан [9]. Проблемы анализа данных образовательного процесса рассматривались в работах таких ученых, как Р. Бакер [8], Л.И. Григорьев [1] и другие.

Нерешенные части общей проблемы. В настоящее время в вузах Казахстана широко применяется кредитная технология обучения. Учебные достижения студентов фиксируются в течение семестра с помощью контрольных точек, а итоговая оценка рассчитывается по формуле:

$$И = 0{,}6 \times \frac{РК_1 + РК_2}{2} + 0{,}4 \times Э, \quad (1)$$

где $РК_1$, $РК_2$ – цифровые эквиваленты оценок первого и второго рубежного контроля соответственно; $Э$ – цифровой эквивалент оценки на экзамене.

В информационных системах вузов Казахстана накоплены большие объемы информации об учебной деятельности студентов, которые в основном используются для статистической отчетности и при ранжировании студентов по суммарному среднему баллу, полученному ими в течение всего периода обучения.

Необходимо использовать интеллектуальные алгоритмы обработки информации, которые могли бы дать наглядные и понятные результаты для принятия решений в целях совершенствования учебного процесса. Однако остаются нерешенными проблемы применения методов "Data Mining" для анализа данных и принятия решений в сфере образования.

Целью исследования является анализ результатов обучения и деятельности студентов для принятия управленческих и организационных решений с использованием методологии оперативного и интеллектуального анализа данных.

Основные результаты исследования. Для управления качеством образовательного процесса авторами предложен подход, включающий методы технологий OLAP и "Data Mining", который позволяет:

- выявить закономерности и тренды, существующие в системах данных образования;
- формировать кластеры, содержащие объекты образования со сходными характеристиками;
- находить зависимости в больших наборах данных;
- выявлять показатели образования, которые наилучшим образом позволяют прогнозировать результаты образовательного процесса;
- строить модели для прогнозирования результатов образовательной деятельности;
- выявлять слабые и сильные стороны образовательной политики;
- генерировать рекомендации для принятия управленческих решений [3].

"Data Mining" – это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных для интерпрета-

ции знаний, необходимых для принятия решений в различных сферах человеческой деятельности [7]. Процесс создания модели интеллектуального анализа включает следующие этапы:

- усвоение данных (обследование первичных данных);
- проверка данных на полноту;
- анализ ключевых факторов влияния;
- выделение исключений;
- обнаружение категорий;
- анализ сценариев;
- прогнозирование.

Для решения задач анализа данных образовательного процесса используются следующие методы "Data Mining":

- факторный анализ (Factor Analysis);
- кластерный анализ и классификация (Clustering);
- поиск ассоциативных правил (Association);
- деревья принятия решений (Decision Trees).
- последовательная кластеризация (Sequence Clustering);
- временные ряды (Time Series);
- нейронные сети (Neural Network) .

Для реализации предложенного подхода обрабатывались и анализировались данные образовательной статистики высшего учебного заведения города Усть-Каменогорска – Восточно-Казахстанского государственного технического университета (ВКГТУ) им. Д. Серикбаева за 2009–2013 годы. Назовем эти переменные первичными, поскольку они отражают первичные данные. Всего рассматривается более 121 тыс. записей по следующим 17 показателям:

- группа;
- ФИО;
- специальность;
- курс;
- основа обучения;
- дисциплина;
- язык обучения;
- балл единого национального тестирования (ЕНТ);
- рубежный контроль 1;
- рубежный контроль 2;
- экзаменационная оценка;
- итоговая оценка;
- название школы;
- тип аттестата;
- балл аттестата;
- количество часов пропусков;
- ФИО преподавателя.

Алгоритм предлагаемого подхода графически представлен на рис. 1.

Основным требованием к информационной системе, ориентированной на анализ данных, является своевременное обеспечение аналитика всей информацией, необходимой для принятия решения [6]. Для реализации опера-

тивного и интеллектуального анализа данных в качестве первичных данных выступают сведения учебного процесса вуза, источником которых является база данных университета (например, база данных SPortal ВКГТУ им. Д. Серикбаева). Собранные данные, как правило, нуждаются в дополнительной обработке, называемой очисткой. В процессе очистки при необходимости может производиться удаление «выбросов» (нехарактерных и ошибочных значений), обработка отсутствующих значений параметров, численное преобразование (например, нормализация) и т.д. Таким образом, первым шагом анализа являются сбор и очистка данных учебного процесса.

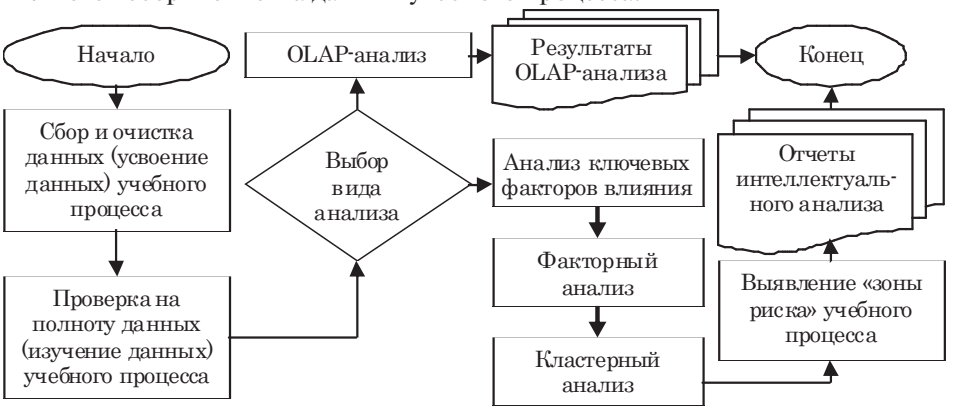

Рис. 1. **Алгоритм анализа данных учебного процесса,** *авторская разработка*

Вторым шагом предлагаемого подхода является изучение данных, которое позволит понять, насколько адекватно подготовленный набор представляет учебный процесс вуза. Здесь может проводиться поиск минимальных и максимальных значений параметров, анализ распределений значений и других статистических характеристик, сравнение полученных результатов с представлениями о деятельности учебного заведения.

Далее выбирается вид анализа: OLAP-анализ или интеллектуальный анализ данных. Решение на базе OLAP позволяет реализовать быстрые операции агрегирования/детализации данных по произвольному набору показателей, предоставляя таким образом аналитику детализированную либо обобщенную оперативную информацию по интересующим его показателям образовательного процесса [4]. Для нашего анализа в качестве измерений, в разрезах которых будет анализироваться данные, могут выступать:

- показатели учебного процесса (оценки за экзамен, итоговый балл, рубежный контроль 1 и 2 и т.д.);
- период (в зависимости от степени детализации год, семестр, рейтинговый период, неделя);
- факультет;
- уровень агрегации (кафедра, специальность, группа).

Многомерная модель визуально представляется с помощью куба (или в случае более трех измерений – гиперкуба). Далее строится куб, в котором значения показателей учебного процесса есть функция от переменных «Показатели», «Год» и «Факультеты». Тогда в качестве измерений будут выступать

«Показатели», «Год» и «Кафедры». На рис. 2 приведен многомерный куб данных для представления данной функции.

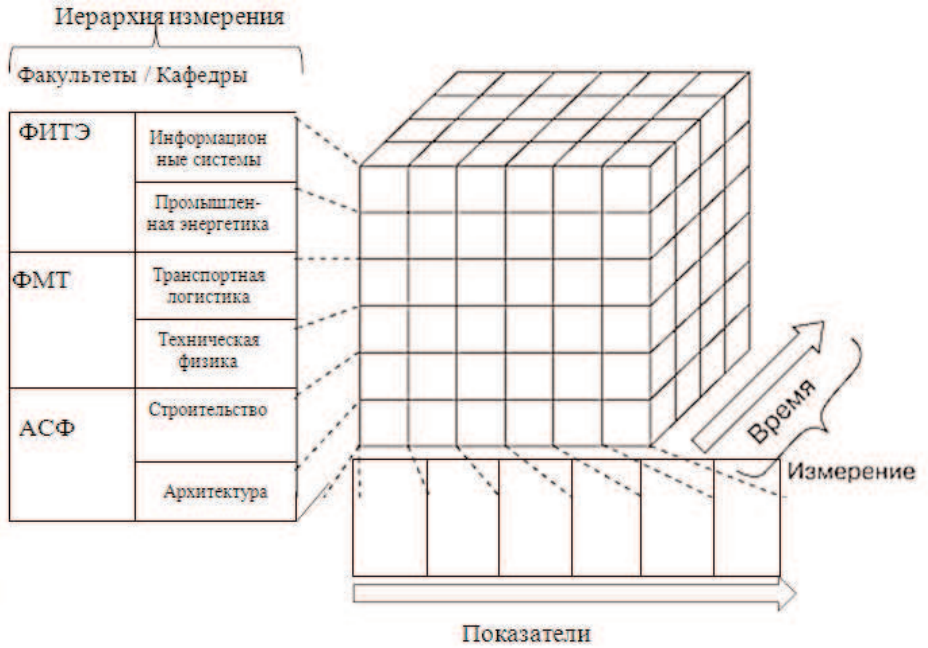

Рис. 2. **Куб данных учебного процесса,** *авторская разработка*

На этих измерениях могут быть заданы иерархии, представленные на рис. 3.

Рис. 3. **Иерархии куба данных учебного процесса,** *авторская разработка*

Результаты OLAP-анализа позволяют принимать своевременные и обоснованные решения для управления качеством образовательного процесса, что является гарантией успеха деятельности вуза.

Кроме OLAP-анализа данная система позволяет реализовать интеллектуальный анализ, который включает следующие этапы: анализ влияния факторов, факторный анализ, кластерный анализ.

Анализ влияния факторов позволяет определить, как зависит результат образовательного процесса от других параметров (факторов обучения). В данном исследовании был проведен анализ влияния факторов на экзаменационную оценку по дисциплинам. Таким образом, данный этап позволяет оценить степень влияния разных параметров образовательного процесса друг на друга, при этом следует убрать из рассмотрения полностью независимые и, наоборот, полностью зависимые факторы.

Результаты этапа интеллектуального анализа данных «Анализ влияния факторов» представлены на рис. 4.

Отчет по ключевым факторам влияния для "Экзамен"

Ключевые факторы влияния и их воздействие результаты экзамена

Столбец	Значение	Подходит	Относительное влияние
Курс	2	не сдал	
Курс	1	не сдал	
Основа обучения	Договор	не сдал	
Специальность	5В090100	не сдал	
Специальность	5В042000	не сдал	
Язык обучения	каз.	не сдал	
Специальность	5В070900	не сдал	
Специальность	5В071300	не сдал	
Специальность	5В071800	не сдал	
Специальность	5В070700	не сдал	

Рис. 4. **Отчет по ключевым факторам влияния на экзаменационную оценку,** *авторская разработка*

Из представленного отчета видно, что на экзаменационную оценку наибольшее влияние оказывает курс, основа и язык обучения. Таким образом, данный этап анализа данных выявляет показатели образовательного процесса, в наибольшей степени влияющие на результаты обучения.

На втором шаге интеллектуального анализа данных выполняется генерация вариантов, обобщающих фактор, а на третьем – их сравнение между собой и разделение по группам.

В результате проведения операции факторизации (уменьшения данных) выделено 3 ведущих фактора, объясняющих 62,9% совокупной дисперсии. Попытаемся интерпретировать данные, используя матрицу повернутых компонент (табл. 1).

Фактор 1 наиболее тесно связан с такими показателями, как PK_1, PK_2, экзаменационная оценка, которые можно охарактеризовать как «Успеваемость». ***Фактор 2*** (индивидуальный фактор) связывает такие показатели, как «Курс» и «Школа». ***Фактор 3*** включает в себя показатель «Специальность».

Таблица 1. **Матрица повернутых компонент** [2]

Показатели	Факторы		
	Фактор 1	Фактор 2	Фактор 3
Группа	-0.071328	0.603215	0.33152
Специальность	-0.088037	-0.007445	0.718477*
Курс	0.287471	-0.842134*	0.05706
Основа обучения	0.614993	0.309936	-0.065313
Дисциплина	0.032099	-0.462158	0.231556
Язык обучения	0.087639	0.513117	0.231897
Балл ЕНТ	0.636538	0.230571	-0.093444
РК1	0.811277*	0.083258	0.108555
РК2	0.838042*	0.096895	0.112544
Экзамен	0.808068*	-0.00942	0.031197
Школа	-0.302153	0.789023*	0.024618
Тип аттестата	0.316242	0.176647	-0.563136
Балл аттестата	-0.024008	0.109804	-0.426023
Пропуски	-0.669468	0.058263	-0.164445

* значения факторных нагрузок, имеющих наибольшее значение связи фактора и соответствующего показателя.

Важной характеристикой факторного анализа является то, что получаемые факторы, в отличие от исходных признаков, являются независимыми. Более рациональным в условиях мультиколлинеарности можно считать построение уравнения регрессии на главных компонентах, которые являются линейными функциями всех исходных показателей и не коррелированы между собой [10].

Третий этап – классификация по нескольким обобщающим показателям (главным компонентам), полученным с помощью методов факторного анализа.

При классификации использовались методы кластерного анализа. Преимущество кластерного анализа состоит в том, что распределение объектов можно проводить не по одной переменной, а по набору признаков. Это позволяет не только выявлять группы схожих объектов, но и создает предпосылки для установления того, что означает такое распределение на кластеры, чем оно вызвано [5].

Средние значения показателей для четырех кластеров представлены на рис. 5.

Первый кластер характеризуется низким значением таких показателей, как балл ЕНТ, PK_1, PK_2, экзамен. Второй кластер характеризуется средним значением вышеуказанных показателей. Для третьего и четвертого кластеров характерны высокие значения показателей.

Предложенный в работе алгоритм был реализован в информационно-аналитической системе анализа данных образовательного процесса, архитектура которой представлена на рис. 6.

Программные средства проектирования информационно-аналитической системы включают следующие компоненты:

- многомерная база данных спроектирована в Microsoft SQL Server 2008 R2;
- для разработки клиентского приложения выбрана среда Visual Studio 2008 C# NET;

- для создания модели оперативного анализа данных применялся компонент Microsoft SQL Server Analysis Services;
- для создания модели интеллектуального анализа данных применялась среда SQL Server Business Intelligence Development Studio.

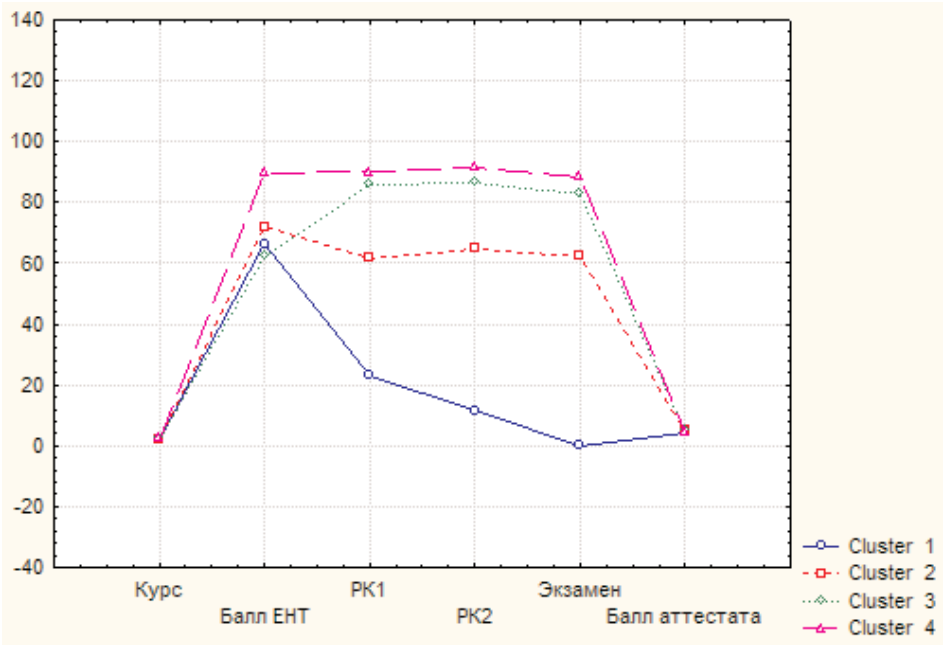

Рис. 5. **Средние значения показателей для каждого кластера,** *авторская разработка*

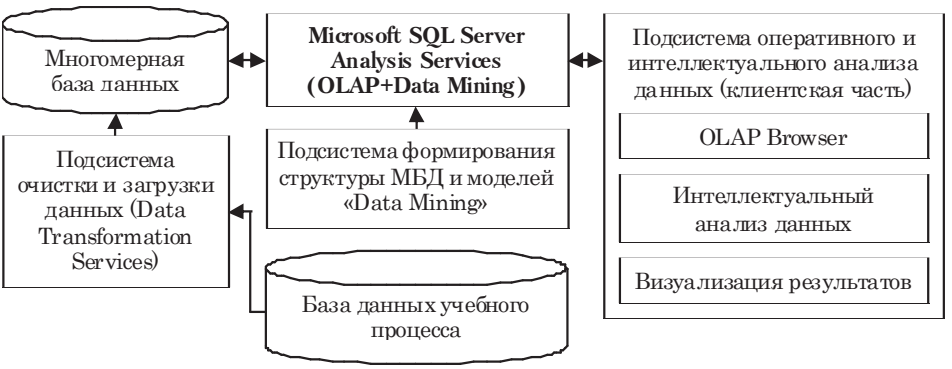

Рис. 6. **Архитектура информационно-аналитической системы анализа данных образовательного процесса,** *авторская разработка*

Выводы. Таким образом, существующий подход к управлению качеством образования в вузах имеет ряд таких недостатков, как низкий уровень аналитической обработки данных образовательного процесса, поэтому был предложен новый подход управления качеством образовательной системы с использованием методов оперативного и интеллектуального анализа данных, а также

спроектирована архитектура информационно-аналитической системы для управления качеством образовательного процесса. Результаты оперативного и интеллектуального анализа позволяют улучшить управленческую деятельность в сфере образования и могут использоваться в системах поддержки принятия решений.

1. *Григорьев Л.И.* Научно-методические и технологические основы информационной системы управления качеством учебного процесса. – М.: Нефть и газ, РГУ нефти и газа им. И.М. Губкина, 2008. – 132 с.

2. *Жуковская В.М., Мучник И.Б.* Факторный анализ в социально-экономических исследованиях. – М.: Статистика, 1976. – 151 с.

3. *Константиновский Д.Л., Агранович М.Л., Дымарская О.Я.* От сбора статистических данных – к информационному обеспечению принятия решений. – 2-е изд., доп. и перераб. – М.: Логос, 2006. – 160 с.

4. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. Куприянов и др. – СПб.: БХВ-Петербург, 2004. – 336 с.

5. *Нейский И.М.* Классификация и сравнение методов кластеризации // Интеллектуальные технологии и системы: Сборник учебно-методических работ и статей аспирантов и студентов. – Вып. 8. – М., 2006. – С. 130–142.

6. *Паклин Н.Б., Орешков В.И.* Бизнес-аналитика: от данных к знаниям (+CD): Учеб. пособие. – 2-е изд., перераб. и доп.– СПб.: Питер, 2010. – 704 с.

7. *Пятецкий-Шапиро Г.* Data Mining и перегрузка информацией: Вступительная статья // Анализ данных и процессов / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров и др. – 3-е изд. перераб. и доп. – СПб.: БХВ-Петербург, 2009. – С. 13–14.

8. *Baker, R.S.J.d., Yacef, K.* (2009). The State of Educational Data Mining in 2009: A Review and Future Visions.

9. *Han, J., Kamber, M.* (2001). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

10. *Romesburg, H.C.* (2004). Cluster Analysis for Researchers. Krieger Pub. International Conference on Computer Science and Software Engineering (pp. 452–455). Washington, DC.

Стаття надійшла до редакції 15.07.2013.