

На правах рукописи

Вахитов Александр Робертович

**МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ СИСТЕМЫ
ОПЕРАТИВНОЙ ОБРАБОТКИ И ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА
ДАННЫХ, ИСПОЛЬЗУЮЩЕЙ НЕЧЕТКУЮ ЛОГИКУ**

Специальность 05.13.11 – Математическое и программное
обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата
технических наук

Томск – 2010

Работа выполнена в Национальном исследовательском Томском политехническом университете.

Научный руководитель: доктор технических наук, профессор
Силич Виктор Алексеевич

Официальные оппоненты: доктор технических наук, профессор
Спицын Владимир Григорьевич

кандидат технических наук, доцент
Черкашин Михаил Владимирович

Ведущая организация Иркутский государственный технический
университет, г. Иркутск

Защита состоится «1» декабря 2010 г. в 14⁰⁰ на заседании совета по защите докторских и кандидатских диссертаций Д 212.269.06 при Национальном исследовательском Томском политехническом университете по адресу: 634034, г. Томск, ул. Советская, 84/3.

С диссертацией можно ознакомиться в научно-технической библиотеке Национального исследовательского Томского политехнического университета по адресу: 634034, г. Томск, ул. Белинского, 55.

Автореферат разослан «___» октября 2010 г.

Ученый секретарь совета по защите докторских и кандидатских диссертаций Д 212.269.06, кандидат технических наук, доцент



Сонькин М.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Совершенствование информационных систем (ИС), во многом определяется тем, насколько успешно будут решаться проблемы представления и обработки знаний. Успехи в области создания интеллектуальных информационных систем (ИИС) стимулировали применение технологий и методов искусственного интеллекта для решения двух основных задач: повышения скорости обработки данных и расширения функций системы.

Одним из основоположников классического подхода к обработке баз данных (БД) был Э. Кодд. Данный подход предполагает использование транзакционных БД, набор аналитических функций в которых весьма ограничен. Схемы, используемые в OLTP-приложениях, основанных на транзакционных БД, осложняют создание даже простых отчетов, так как данные чаще всего распределены по множеству таблиц, и для их агрегирования необходимо выполнять сложные операции объединения. Кроме того, что в таких системах хранятся постоянно изменяющиеся данные. По мере осуществления операций записи-чтения суммарные значения меняются очень быстро, поэтому, чаще всего, анализ выполняется по окончании отчетного периода, иначе картина может оказаться искаженной.

В значительной степени проблемы обработки данных на основе классического подхода решает переход к объединению и анализу данных учетной системы с помощью технологии Online Analytical Processing (OLAP). OLAP-моделью в работе называется логическая модель данных, лежащая в основе OLAP-технологии. Использование технологии OLAP для обработки запросов позволяет повысить скорость и удобство обработки данных. Клиент-серверная архитектура OLAP-продуктов обеспечивает одновременный доступ большого числа пользователей, при этом анализ проводится одинаково быстро по всем аспектам информации независимо от размера и сложности структуры БД. Однако зачастую преобразование реляционной модели обработки данных в OLAP-модель не решает проблем скорости обработки данных и их интеллектуального анализа.

Технология data mining, представляющая собой выявление скрытых закономерностей или взаимосвязей между переменными в больших массивах необработанных данных, является расширением OLAP-технологии. Data mining включает методы и модели статистического анализа и машинного обучения, целью которых является автоматическая обработка данных. Инструменты data mining позволяют проводить анализ данных предметными специалистами (аналитиками), не владеющими соответствующими математическими знаниями. Однако на практике часто возникают задачи, для которых использование средств data mining является недостаточным в связи с тем, что постоянно увеличивается объем обрабатываемых данных, а также спектр требуемых функций системы по интеллектуальному анализу данных. Поэтому существует необходимость в создании эффективных алгоритмов по обработке и анализу данных в ИИС.

Другой проблемой ИИС систем является то, что они строятся как самостоятельные программы, имея собственную организацию хранения данных и знаний. Поэтому их применение для решения сложных задач в промышленных масштабах становится проблематичным. Разработка и внедрение ИИС требует не только значительно больших объемов информации, но и значительно более сложных вычислений, учета слабо формализуемых факторов, высокого уровня интерфейса, связи с уже существующими БД.

Таким образом, несмотря на имеющиеся достижения в области создания ИИС и на наличие целого ряда методов и программных средств обработки данных, задача создания информационной системы, обеспечивающей оперативную обработку информации, обладающую возможностями интеллектуального анализа данных и имеющую интерфейс с активно используемыми СУБД, является актуальной и представляет интерес в научном и практическом плане, что определяет актуальность темы диссертационной работы.

Целью работы является создание математического и программного обеспечения системы оперативной обработки и интеллектуального анализа данных, использующей нечеткую логику и обеспечивающей повышение эффективности процесса обработки данных.

Задачи для достижения поставленной цели:

- формирование требований к системе оперативной обработки и интеллектуального анализа данных;
- анализ подходов к построению математических моделей и обоснование выбора класса математической модели для разрабатываемой системы, реализующей функции логического вывода в условиях неопределенности и неполноты знаний исследуемой предметной области;
- анализ подходов к построению программного обеспечения и обоснование выбора программного обеспечения для реализации разрабатываемой системы, включающей функции OLAP, data mining, ключевых индикаторов производительности, деревьев решений и трендов;
- формирование концептуально-информационной модели предметной области (КИМПО) и реляционной модели обработки данных для обработки информации о НИРС в вузе;
- создание БД и базы знаний для обработки информации о НИРС в вузе;
- разработка математической модели системы с использованием аппарата нечеткой логики, в том числе описание исследуемой предметной области в терминах нечеткой логики и проектирование правил нечеткого логического вывода;
- разработка OLAP-модели данных на основе математической модели, КИМПО, реляционной схемы данных;

- расширение состава функций OLAP-модели путем разработки таких средств интеллектуального анализа данных, как data mining, трендов, ключевых индикаторов производительности и деревьев решений;
- программная реализация системы оперативной обработки и интеллектуального анализа данных на основе разработанных моделей и алгоритмов для обработки информации о НИРС в вузе;
- визуализация данных, применение алгоритмов кластеризации для интеллектуальной обработки данных;
- использование моделей, алгоритмов и программ для организации обработки данных в Учебном центре ТПУ & SoftLine Academy;
- использование моделей, алгоритмов и программ для организации обработки данных в ФГУП «Томской электротехнический завод».

Методы исследования. В процессе исследования использовались следующие теории и методы: теория реляционной модели данных, теория математической логики, аппарат нечеткой логики, теории принятия решений, методы объектно-ориентированного программирования, методы аналитической обработки информации, методы извлечения данных.

Научная новизна. Получены следующие основные результаты, обладающие научной новизной:

1. Предложена математическая модель интеллектуальной обработки данных, основанная на аппарате нечеткой логики и позволяющая повысить эффективность обработки данных и знаний при решении широкого круга задач.
2. Разработан оригинальный алгоритм обработки данных, использующий нечеткую логику, аналитическую обработку данных в реальном времени и механизм извлечения данных.
3. Разработан новый алгоритм совместного использования процедур нечеткого логического вывода и процедур обработки данных (data mining, деревья решений, тренды и т.д.).
4. Впервые предложен алгоритм динамического изменения структуры данных на этапе работы пользователя с программной системой, позволяющий повысить эффективность его работы.
5. Создана и исследована программная система оперативной обработки и интеллектуального анализа данных на основе разработанных моделей и алгоритмов.

Практическая значимость работы. Практическая значимость исследования заключается в разработке системы оперативной обработки и интеллектуального анализа данных, использующей нечеткую логику, которая предназначена для решения широкого круга задач по обработке и анализу данных. Использование системы позволяет автоматизировать процесс обработки данных, осуществлять интеллектуальный анализ данных, извлекать отсутствующие данные, увеличить скорость обработки данных и удобство представления данных конечному пользователю по сравнению с классическими методами.

Реализация результатов работы. Разработанные модели и алгоритмы использованы при создании программной системы WinDIFOT (свидетельство о регистрации в «Объединенном фонде электронных ресурсов «Наука и образование» №15692 от 04.05.2010 г.).

Внедрение результатов работы. Результаты диссертационной работы внедрены в Учебном Центре ТПУ & SoftLine Academy и в ФГУП «Томский электротехнический завод», что подтверждено соответствующими актами.

Основные положения, выносимые на защиту:

1. разработанная математическая модель интеллектуальной обработки данных, основанная на аппарате нечеткой логики, позволяющая повысить скорость и эффективность обработки данных применительно к существующим задачам;
2. разработанная OLAP-модель данных, а также средства интеллектуального анализа данных, в том числе правила data mining, тренды, ключевые индикаторы производительности, которые наделяют систему возможностями оперативной аналитической обработки информации;
3. разработанный алгоритм обработки данных, использующий нечеткую логику, аналитическую обработку данных в реальном времени и механизм извлечения данных;
4. модель, включающая процедуры нечеткого логического вывода и процедуры обработки данных в рамках одной программной системы;
5. созданная система оперативной обработки и интеллектуального анализа данных, использующей нечеткую логику, которая позволяет повысить эффективность процесса работы с данными.

Личный вклад автора. Основные результаты диссертационной работы получены автором лично. Программная система WinDIFOT для оперативной обработки и интеллектуального анализа данных разработана и реализована автором лично.

Апробация работы. Основные результаты работы докладывались и обсуждались на следующих конференциях:

- Молодежь и современные информационные технологии: IV Всероссийская научно-практическая конференции студентов, аспирантов и молодых ученых: Томск, 2006.
- Энергия молодых – экономике России: VII Международная научно-практическая конференция студентов и молодых ученых: Томск, 2006.
- Снежинск и наука – 2006: VII Международная научно-практическая конференция: Снежинск, 2006.
- Математическое моделирование и информационные технологии: VIII Школа-семинар молодых ученых: Улан-Удэ, 2006.
- Молодежь и современные информационные технологии: IV Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых: Томск, 2007.

- Современные техника и технологии СТТ 2007: XIII Международная научно-практическая конференция студентов и молодых ученых: Томск, 2007.
- Информационные системы и модели в научных исследованиях, промышленности и экологии: Всероссийская научно-техническая конференция: Тула, 2007
- Молодежь и современные информационные технологии: IV Всероссийской научно-практической конференция студентов, аспирантов и молодых ученых: Томск, 2008.
- Технологии Microsoft в теории и практики программирования: VI Всероссийская научно-практическая конференции: Томск, 2009.
- Технологии Microsoft в теории и практики программирования: VI Всероссийская научно-практическая конференция: Томск, 2010.

Публикации. По теме диссертации опубликовано 13 работ: 2 – в рецензируемых отечественных изданиях, рекомендованных ВАК, 10 – в материалах российских и международных конференций, 1 – свидетельство о регистрации электронного ресурса. Список публикаций приведен в конце автореферата.

Структура и объем работы. Диссертация включает в себя: введение, три главы, заключение, список литературы (111 наименований) и приложения, иллюстрирующие технические детали работы. Общий объем работы составляет 152 страницы, включая 46 рисунков и 12 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность работы, приводятся цель, основные положения, выносимые на защиту, задачи и методы исследования, формулируется научная новизна, практическая значимость исследования, а также сведения о внедрении результатов, публикациях, апробации работы, личном вкладе автора, объеме и структуре диссертации.

Глава 1

В первой главе проводится анализ подходов к построению математических моделей и обосновывается выбор класса математической модели для разрабатываемой системы, реализующей функции логического вывода в условиях неопределенности и неполноты знаний, характерных для решения задач оперативной обработки и интеллектуального анализа данных. Также в этой главе обосновывается выбор модели обработки данных, включающей функции OLAP, data mining, ключевых индикаторов производительности, деревьев решений и трендов. Кроме того, в первой главе выявляются требования к системе оперативной обработки и интеллектуального анализа данных.

Создание математического обеспечения разрабатываемой системы предполагает обоснование выбора класса математической модели из множества X альтернативных вариантов x_i , а также непосредственное описание предметной области в терминах выбранного класса. К числу основных логических моделей, для которых разработаны формальные методы логического вывода, были отнесены: x_1 – исчисление высказываний, x_2 – исчисление предикатов, x_3 – семантические сети, x_4 – дескриптивная логика, x_5 – нечеткая логика.

С учетом особенностей предметной области, связанных с неполнотой и неопределенностью знаний были определены следующие частные критерии выбора: F_1 – полнота описания системы, F_2 – возможность построения логического вывода, F_3 – возможность работы с неполными и субъективными данными, F_4 – возможность осуществления логических операций, F_5 – удобный интерфейс, F_6 – высокая скорость обработки данных, F_7 – наличие интеллектуальных возможностей обработки данных.

Затем каждый из альтернативных вариантов x_i был описан относительно возможности применения в качестве класса математической модели для разрабатываемой системы.

Для выбора класса математической модели были использованы метод анализа иерархий, предложенный Томасом Саати, а также интегральные критерии, являющиеся функциями от частных критериев. Указанные методы нашли широкое распространение для обоснования выбора на множестве альтернативных вариантов.

На рис. 1. приведена иерархия целей для выбора класса математической модели системы с использованием метода Саати.



Рис. 1. Иерархия цели для выбора класса математической модели системы

Затем для каждого уровня иерархии были построены матрицы парных сравнений в терминах доминирования одного элемента над другим. На рис. 2 приведены матрицы парных сравнений, построенные для второго и третьего уровней иерархии. Для каждого из экспертов была построена своя матрица

парных сравнений. Для построения обобщенной матрицы чаще всего был использован метод средней оценки влияния, т.е. каждый элемент обобщенной матрицы сравнений представляет из себя среднее арифметическое от соответствующих элементов матриц отдельных экспертов.

	a ₁	a ₂
a ₁	1	1/2
a ₂	2	1

	F ₁	F ₂	F ₃	F ₄
F ₁	1	1	1/3	4
F ₂	1	1	2	2
F ₃	3	1/2	1	6
F ₄	1/4	1/2	1/6	1

	F ₅	F ₆	F ₇
F ₅	1	1	1
F ₆	1	1	5
F ₇	1	1/5	1

матрицы для 1-го эксперта

	a ₁	a ₂
a ₁	1	1/4
a ₂	4	1

	F ₁	F ₂	F ₃	F ₄
F ₁	1	2	1/3	3
F ₂	1/2	1	1	4
F ₃	3	1	1	5
F ₄	1/3	1/4	1/5	1

	F ₅	F ₆	F ₇
F ₅	1	1/4	3
F ₆	4	1	7
F ₇	1/3	1/7	1

матрицы для 2-го эксперта

	a ₁	a ₂
a ₁	1	1/3
a ₂	3	1

	F ₁	F ₂	F ₃	F ₄
F ₁	1	3	1/5	2
F ₂	1/3	1	5	1
F ₃	5	1/5	1	4
F ₄	1/2	1	1/4	1

	F ₅	F ₆	F ₇
F ₅	1	1/2	2
F ₆	2	1	1
F ₇	1/2	1	1

матрицы для 3-го эксперта

	a ₁	a ₂
a ₁	1	1/3
a ₂	3	1

	F ₁	F ₂	F ₃	F ₄
F ₁	1	2	1/4	3
F ₂	1/2	1	1/3	2
F ₃	4	3	1	5
F ₄	1/3	1/2	1/5	1

	F ₅	F ₆	F ₇
F ₅	1	1/2	2
F ₆	2	1	4
F ₇	1/2	1/4	1

обобщенные матрицы

Рис. 2. Матрицы парных сравнений второго и третьего уровня иерархии

На основе каждой из построенных матриц формируются наборы локальных приоритетов, которые отражают относительные приоритеты сравниваемых элементов. Для этого нужно вычислить множество собственных векторов для каждой матрицы, а затем нормализовать результат к единице, получая тем самым вектор приоритетов. Одним из наилучших путей вычисления нормализованных приоритетов является геометрическое среднее. Его можно получить, перемножая элементы в каждой строке и извлекая корни n -ой степени, где n – число элементов. Полученный таким образом столбец чисел нормализуется делением каждого числа на сумму всех чисел. В итоге для уровня акторов были получены следующие значения нормализованных приоритетов: $a_1=0,25$; $a_2=0,75$.

Для уровня целей и требований акторов были получены следующие значения нормализованных приоритетов: $F_1=0,22$; $F_2=0,15$; $F_3=0,55$; $F_4=0,08$; $F_5=0,286$; $F_6=0,57$; $F_7=0,14$.

На последнем шаге анализа локальные приоритеты были пересчитаны с учетом приоритетов направляемых элементов. Глобальный приоритет варианта x_5 оказался самым высоким (0,418). Таким образом, по методу Саати обоснован выбор нечеткой логики в качестве класса математической модели системы.

Затем обоснование выбора класса математической модели системы производилось при помощи интегральных критериев. Общая формула интегральных критериев, являющихся функцией от частных критериев, имеет следующий вид: $F = f(F_i), i = \overline{1, n}$

Наиболее широко распространены такие интегральные критерии как:

1. Максимум суммы взвешенных оценок: $F = \sum_{i=1}^n w_i \bullet F_i(x_j) \rightarrow \max_{x_j \in X}$
2. Минимум суммы квадратов отклонений от «идеальной точки»:

$$\sum_{i=1}^n w_i \bullet (\tilde{F}_i - F_i(x_j))^2 \rightarrow \min_{x_j \in X}$$

Для определения значений весовых коэффициентов w_i каждого из частных критериев F_i использовался метод парных сравнений. На основе матрицы сравнений, были получены следующие компоненты собственного вектора для каждой из строк: $F_1=1,24$; $F_2=0,999$; $F_3=2,83$; $F_4=0,72$; $F_5=0,407$; $F_6=0,98$; $F_7=0,99$.

После деления каждой из полученных компонент на их сумму, равную 8,16, были получены следующие нормализованные приоритеты: $F_1=0,15$; $F_2=0,12$; $F_3=0,35$; $F_4=0,09$; $F_5=0,05$; $F_6=0,12$; $F_7=0,12$. Эти значения являются весовыми коэффициентами w_i частных критериев F_i . В соответствии с усредненными значениями оценок на основе мнений каждого из экспертов были определены интегральные критерии для каждого из альтернативных вариантов. В качестве класса математической модели для исследуемой системы была выбрана нечеткая логика, поскольку она имеет преимущества перед альтернативными вариантами математического обеспечения системы по всем интегральным критериям, а также имеет самый высокий глобальный приоритет, определенный методом анализа иерархий Саати.

Описанный подход был использован также для выбора модели обработки данных. Было рассмотрено 5 основных моделей обработки данных: y_1 – реляционная модель обработки данных, y_2 – активные и дедуктивные БД, y_3 – постреляционные БД, y_4 – объектно–ориентированные БД, y_5 – OLAP–модель, а также выявлено 5 частных критериев выбора: G_1 – полнота представления предметной области, G_2 – возможность извлечения отсутствующих данных, G_3 – удобный интерфейс, G_4 – высокая скорость обработки данных, G_5 – наличие интеллектуальных возможностей обработки данных.

В итоге в качестве модели обработки данных для исследуемой системы была выбрана OLAP-модель, поскольку она имеет преимущества перед

альтернативными вариантами программного обеспечения системы по всем интегральным критериям, а также имеет самый высокий глобальный приоритет, определенный методом анализа иерархий Саати.

В первой главе описан ряд экспериментов с целью определения преимуществ OLAP-модели по скорости обработки данных в сравнении с реляционными базами данных. Эксперименты заключались в последовательном создании следующих отчетов:

1. Выборка информации об общем числе активностей по НИРС и их характеристиках.
2. Выборка информации о числе активностей по НИРС и их характеристиках определенного студента.
3. Выборка информации о числе активностей по НИРС и их характеристиках определенного студента и преподавателя.

Во всех экспериментах в первом случае источником данных являлась реляционная БД, во втором – OLAP-модель. Число записей в основной таблице сначала равнялось 500, затем было увеличено до 1000, 5000, 10000, 20000 и до 100000, аппаратная платформа для эксперимента была идентична¹.

В результате экспериментов среднее время обработки запроса (создания отчета) на основе OLAP-модели оказалось в 11,6 раз меньше. Результаты эксперимента стали дополнительным основанием для выбора OLAP-модели в качестве модели обработки данных разрабатываемой системы.

Глава 2

Во второй главе формируются КИМПО, реляционная модель обработки данных, БД и базы знаний для исследуемой предметной области. Также в этой главе описана математическая модель системы с использованием аппарата нечеткой логики, исследуемая предметная область в терминах нечеткой логики и правила нечеткого логического вывода. Кроме того, в третьей главе производится разработка OLAP-модели данных на основе математической модели, КИМПО, реляционной схемы данных и расширение OLAP-модели путем разработки таких средств интеллектуального анализа данных, как data mining, трендов, ключевых индикаторов производительности и деревьев решений.

Основной экспериментальной предметной областью, исследуемой в рамках работы, является НИРС в вузе. В дальнейшем модели, алгоритмы, а также результаты экспериментов используются при внедрении системы в других областях. В главе производится описание основных характеристик и особенностей исследуемой предметной области: обозначены основные цели, задачи, виды НИРС и способы поощрения за активное участие в работе. В соответствии с этим описанием были выявлены 2 основные функции разрабатываемой системы: упорядочивание и централизованное хранение

¹ В данном эксперименте значения записей не носят информативных характер, их высокое число используется для сравнения скорости обработки данных.

информации о НИРС в вузе и интеллектуальная обработка информации о НИРС целью автоматизации процесса принятия решений в области НИРС.

Проектирование КИМПО началось с выявления сущностей. В результате анализа информационных потребностей пользователей, а также на основе общего описание предметной области были выделены следующие сущности: студенты, преподаватели, НИРС, достижения, поощрения, наказания, принимаемые решения. Далее для этих сущностей были определены атрибуты, кроме того были выявлены дополнительные сущности в соответствии с выявленными атрибутами, проведена нормализация отношение по трем нормальным формам. В результате связывания сущностей и нормализации отношений была определена схема КИМПО для исследуемой системы. В соответствии с описанием предметной области и схемой КИМПО в системы были выделены 2 OLAP-куба: один содержит сведения о результатах НИРС, другой – о поощрения руководства вуза, связанных с НИРС.

В исследуемой системе для оценки результатов НИРС, а также для генерирования решений в области НИРС используется аппарат нечеткой логики. В данной главе произведено описание процесса нечеткого логического вывода в системе.

Система описывается следующими входными параметрами A_i и выходными параметрами B_i , таблица 1:

Таблица 1. Входные и выходные параметры системы.

Обозначение	Описание
A_1	Количество результатов НИРС
A_2	Сумма денег, полученных за НИРС
A_3	Число страниц, опубликованных по итогам НИРС
A_4	Число наград за НИРС
A_5	Число поощрений по итогам НИРС
A_6	Сумма денег, затраченных на поощрения
A_7	Число полученных грантов
A_8	Число участия в научных конкурсах
A_9	Число призовых мест по итогам научных конкурсов
B_1	Премирование
B_2	Назначение преподавателю руководства НИРС
B_3	Рекомендация для поступления в аспирантуру и магистратуру
B_4	Рекомендация на стажировку
B_5	Командировка на конференцию

Все универсумы находятся в пределах измеримого диапазона с 5 степенями градации (термами): очень низкий $[0..x_1]$, средний $[x_1..x_2]$, высокий $[x_2..x_3]$. Конкретные значения x_i зависят от масштабов анализа данных (на уровне конкретного студента или преподавателя, кафедры, факультета либо вуза в целом) и особенности измерения данного параметра. Далее приведены некоторые правилами вывода в системе:

$L_1: (A_5 \in [0..x_1] \wedge A_6 \in [0..x_1]) \wedge (A_1 \in [x_2..x_3] \vee A_2 \in [x_2..x_3] \vee A_3 \in [x_2..x_3] \vee A_4 \in [x_2..x_3]) \rightarrow B_1 \in [x_1..x_2];$

$L_9: (A_1 \in [x_2..x_3] \vee A_2 \in [x_2..x_3] \vee A_3 \in [x_2..x_3]) \rightarrow B_4 \in [x_1..x_2];$

$L_{15}: (A_1 \in [0..x_1] \vee A_6 \in [0..x_1]) \rightarrow A_5 \in [0..x_1].$

Алгоритм нечеткого вывода был рассмотрен на конкретном примере. У одного из студентов необходимо доопределить значение A_5 , зная значения A_1 и A_6 , используя затем полученные параметры для генерирования решения о том, заслуживает ли студент каких-либо дополнительных поощрений.

Универсум значения числа поощрений A_5 для этого студента находится в отрезке $[0..6]$. Начальное множество термов – низкое, среднее, высокое. Функции принадлежности $\mu(A_5)$ имеют следующий вид, рис. 3:

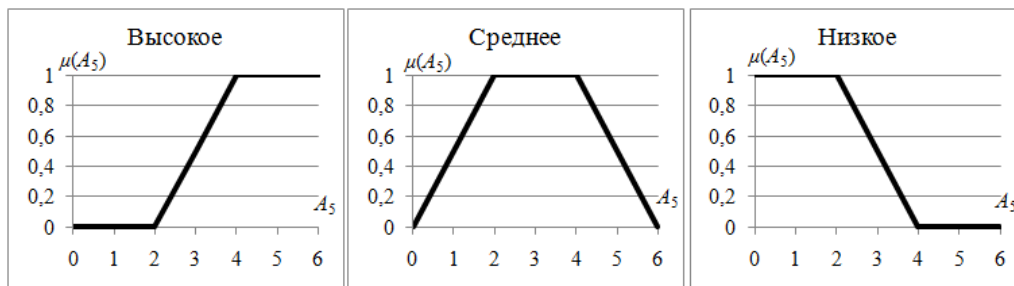


Рис. 3. Функции принадлежности $\mu(A_5)$

Нечеткий логический вывод был осуществлен в 4 этапа:

1. Этап фаззификации.

На основе значений $A_1=15$ и $A_6=7000$ была осуществлена фаззификация, в результате которой были получены следующие степени уверенности в значениях входных переменных:

- Число достижений A_1 большое – 0,65;
- Число достижений A_1 среднее – 0,7;
- Число достижений A_1 малое – 0,35;
- Сумма денег A_6 большая – 1;
- Сумма денег A_6 средняя – 0,5;
- Сумма денег A_6 малая – 0.

2. Этап нечеткого вывода.

На данном этапе были вычислены степени уверенности посылок правил $L_{13} - L_{15}$, представляющих из себя нечеткие импликации:

- $L_{13}: \min(A_1 \in [x_2..x_3] \wedge A_6 \in [x_2..x_3]) = \min(0,65; 1) = 0,65;$
- $L_{14}: \min(A_1 \in [x_2..x_3] \wedge A_6 \in [0..x_1]) = \min(0,65; 0) = 0;$
- $L_{15}: \max(A_1 \in [0..x_1] \vee A_6 \in [0..x_1]) = \max(0,35; 0) = 0,35.$

3. Этап композиции.

Степень уверенности заключения задается функцией принадлежности соответствующего терма. Поэтому с использованием определения нечеткой импликации как минимума левой и правой частей были получены новые нечеткие переменные, соответствующие степеням уверенности в значении выходных данных при применении к заданным входам соответствующего

правила, а затем была проведена аккумуляция – объединение результаты применения всех правил, рис. 4:

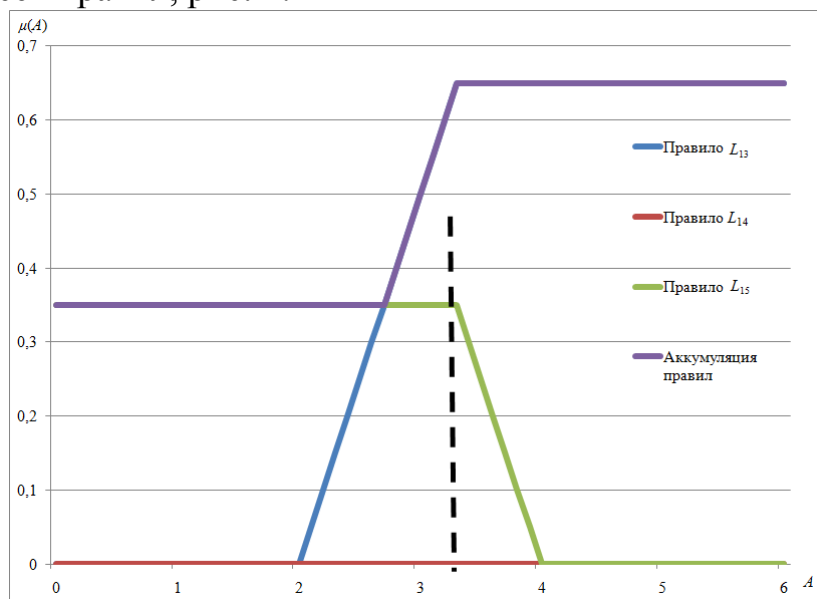


Рис. 4. Аккумуляция правил.

В результате была получена функция принадлежности для числа поощрений A_5 , которая говорит о степени уверенности в значении искомого параметра на основе входных параметров и правил нечеткого логического вывода.

4. Этап дефаззификации.

Для преобразования нечеткого набора значений к точным был использован метод первого максимума, в результате чего было определено, что число поощрений находится в диапазоне «среднее» и равно примерно 3.

Затем полученные данные были использованы для определения выходных параметров B_i . Зная, что $A_1 = 15$, $A_5 = 3$, $A_6 = 7000$ согласно правилу нечеткого логического вывода L_3 : $(A_5 \in [x_1..x_2] \wedge A_6 \in [x_1..x_2]) \wedge (A_1 \in [x_2..x_3] \vee A_2 \in [x_2..x_3] \vee A_3 \in [x_2..x_3] \vee A_4 \in [x_2..x_3]) \rightarrow B_1 \in [0..x_1]$ было определено, что с данными показателями НИР этот студент заслуживает премирования в размере $[0..2000]$.

В разрабатываемой системе в качестве инструментов интеллектуального анализа данных были использованы технологий data-mining, ключевые индикаторы производительности КРІ и тренды.

В качестве наиболее характерного типа закономерностей для осуществления процесса data-mining были определены классификация и ассоциация, используемым методом исследования данных был метод деревьев решений. Применение модели деревьев решений для разрабатываемой системы рассмотрено на конкретном примере. Булевскими атрибутами, которые следует доопределить являются: склонность студентам в НИР, склонность преподавателя к руководству НИРС. Для измерения значений целевой функции используется интегральная оценка НИРС. Дерево решений для определения склонности студента к НИР и преподавателей к руководству НИРС (p –

вероятность исхода в определенном узле, p – число публикаций, s – сумма заработанных денег) показано на рис. 5:

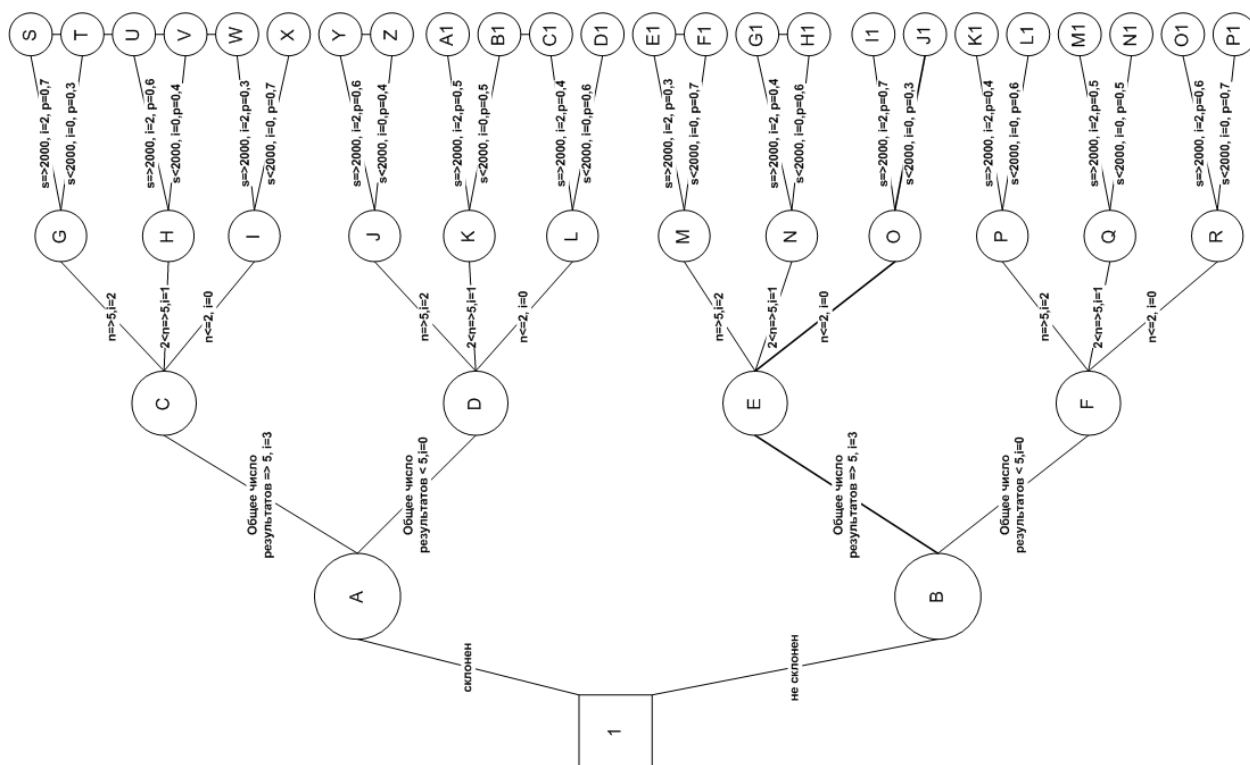


Рис. 5. Дерево решений для определения склонности к НИРС

Зная, что общее число результатов=5, $s=3000$, $p=2$, после отсеечения лишних ветвей получилось, что $EMV(A)=1,5$, а $EMV(B)=3,5$, то есть при данных показателях студент не склонен к НИР (преподаватель не склонен к руководству НИРС) с вероятностью 0,7.

OLAP-технология использовалась также при работе и ключевыми индикаторами производительности – key performance indicator (KPI). В исследуемой системе были выделены следующие KPI: общее число достижений в НИРС на факультете, общее число достижений в НИР определенного студента, число активностей по отдельным видам НИРС, число преподавателей, являющихся руководителями НИРС, а также число студентов, имеющих результаты НИР. Эти показатели сигнализируют конечному пользователю о достижении стратегических целей, либо о том, что те или иные показатели деловой активности организации нуждаются в улучшении.

Кроме того, при анализе данных были использованы тренды. 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. С помощью трендов можно сделать прогноз об активности студентов в НИР в будущем и в соответствии с этим принять решение о

финансировании НИРС на данном факультете и сделать вывод об общих перспективах НИРС на факультете.

Глава 3

В третьей главе проводится программная реализация системы оперативной обработки и интеллектуального анализа данных на основе разработанных моделей и алгоритмов для обработки информации о НИРС в вузе. Здесь рассматриваются такие аспекты реализации системы как выбор инструментов реализации системы, варианты использования системы, физические модели данных и OLAP-кубов, создание иерархий измерений, KPI, трендов, дерева принятий решений, использования data mining. Кроме того, приведены экранные снимки интерфейса программной системы, а также описание процесса визуализации данных для повышения эффективности их интеллектуальной обработки. Также в этой главе описано применение моделей и алгоритмов для организации обработки данных в Учебном центре ТПУ & SoftLine Academy и в ФГУП «Томской электротехнический завод».

В начале главы производится обоснование выбора инструментов для реализации системы. Инструмент для реализации системы должен поддерживать все технологии и методики, выделенные на этапе проектирования. В соответствии с этим было выделено несколько сред разработки программных систем, а затем проведен их сравнительный анализ при помощи интегральных критериев, весовых коэффициентов и матрицы парных сравнений.

Было рассмотрено 4 альтернативных варианта: z_1 – Microsoft SQL Server Analysis Services 2008, z_2 – Oracle Database 11g, z_3 – Deductor Studio 4, z_4 – Crystal Analysis Professional 9 и выделено 10 частных критериев: H_1 – поддержка технологии OLAP, в том числе возможность преобразования реляционной схемы данных в OLAP-модель; H_2 – возможность реализации математической модели нечеткой логики; H_3 – программная поддержка нечеткого логического вывода; H_4 – поддержка data mining; H_5 – использование KPI; H_6 – использование трендов; H_7 – динамическое изменение базы знаний системы; H_8 – удобство представления информации конечному пользователю; H_9 – быстрота обработки информации; H_{10} – стоимость среды разработки.

В результате был выбран вариант z_1 – Microsoft SQL Server Analysis Services 2008. Указанная среда разработки имеет преимущества перед альтернативными вариантами инструментария для реализации системы по всем интегральным критериям.

Согласно КИМПО была создана схема данных в Microsoft SQL Server 2008, в соответствии с которой было построено 2 OLAP-куба, определенных на этапе проектирования системы. При создании OLAP-кубов в среде разработки Business Intelligence Development Studio (BIDS) в качестве основного источника данных выступала база данных SQL Server 2008. Схема OLAP-кубов результатов НИРС и обобщений за НИРС в BIDS представлена на рис. 6:

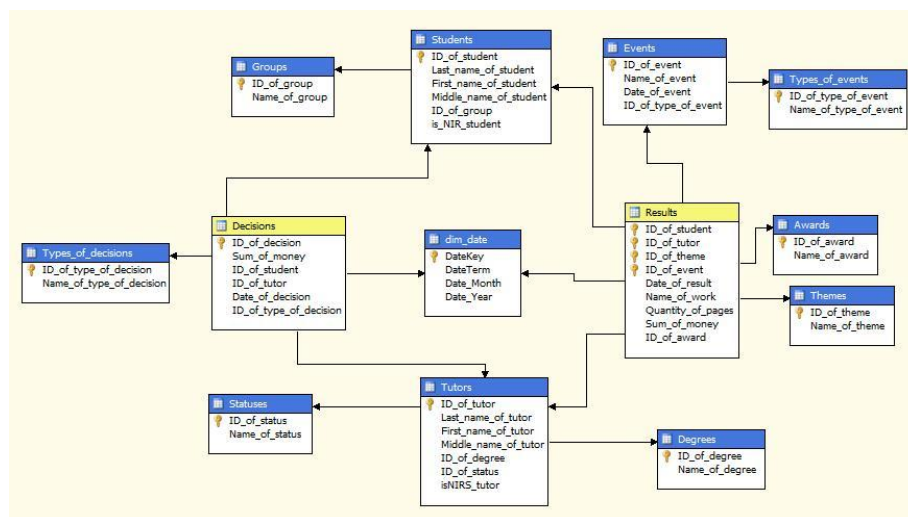


Рис. 6. Схема OLAP-кубов

Бизнес-логика для указанных на этапе проектирования измерений была внедрена путем создания иерархий. Каждая мера была определена с помощью MDX (Multi Dimension Expression) выражения – языка запросов для многомерных баз данных. Для мер, являющихся KPI, были указаны MDX-выражения для расчета целевого значения, выражение показателя эффективности, выражение для расчета состояния, а также вид используемого индикатора. В основе использования трендов находится выражения для расчета тренда при описании мер и вида индикатора для отображения тренда.

Для внедрения data mining использовалась методология добычи данных Microsoft Decision Trees, инструменты реализации которой входят в программную среду разработки BIDS. В таблице фактов OLAP-куба был задан отсутствующий атрибут, который необходимо определить: в данном случае это склонность студента к НИР. Кроме того, задаются входные данные для расчета отсутствующего значения: этими данными являются год поступления, средняя успеваемость, число результатов НИР у студентов, для которых известен атрибут «Склонность к НИРС».

В итоге была получена таблица, в которой записям с определенной долей вероятности присваиваются значения атрибута «склонность НИРС», имеющего сходные с искомым объектом входные параметры. Таким образом, с помощью однократной тренировки системы была реализована возможность получить множество недостающих в БД сведений. Далее приведен MDX-код, предназначенный для создания модели данных:

```
SELECT
    (t.[l_name]) as [Фамилия],(t.[f_name]) as [Имя],(t.[m_name]) as [Отчество],
    ([Students 8].[Is NIR Student]) as [Значение атрибута],
    (PredictProbability([Students 8].[Is NIR Student])) as [Вероятность]
From[Students 8]PREDICTION JOIN
    OPENQUERY([NIRS],
        'SELECT [l_name],[f_name],[m_name],[marks] FROM [dbo].[students_data] ') AS t
ON [Students 8].[Marks] = t.[marks]
```

Основным клиентским инструментом для отображения данных является Microsoft Office Excel 2007, который предполагает использование в качестве источника данных службы аналитики Microsoft SQL Server. Следует отметить, что конечный пользователь имеет возможность динамически изменять структуру запроса к базе данных. Пользователю также доступны формы для просмотра исходной информации. Формы, реализованные с помощью технологии ASP.Net, представляют из себя web-интерфейс для просмотра и(или) редактирования данных. Программирование правил логического вывода в системе осуществлялось при помощи языка программирования C#.Net. Далее приведена часть программного кода, в котором описываются правила нечеткого логического вывода в системе:

```
namespace NIRS_app
{
    public partial class Rules : System.Web.UI.Page
    {
        protected void Page_Load(object sender, EventArgs e)
        {
            int a1,b1,a2,b2,a3,b3,a4,b4,a5,b5,x1,x2,x3,a6;
            if (((a5 >= 0) & (a5 <= x1)) & ((a6 >= 0) & (a6 <= x1)) & ((a1 >= x2) & (a1 <= x3))
            | ((a2 >= x2) & (a2 <= x3)) | ((a3 >= x2) & (a3 <= x3)) | ((a4 >= x2) & (a4 <= x3))) {b1 =1;}
            //правило L1
            if (((a5 >= 0) & (a5 <= x1)) & ((a6 >= 0) & (a6 <= x1)) & ((a1 >= x2) & (a1 <= x3))
            & ((a2 >= x2) & (a2 <= x3)) & ((a3 >= x2) & (a3 <= x3)) & ((a4 >= x2) & (a4 <= x3))) { b1 = 1; }
            // правило L3 }}}
        }
    }
}
```

Отдельная страница системы используется для просмотра и принятия управленческих решений, рис. 7:

Предлагаемое решение		Отметка о принятии
Принять решение	Премирование студента (Петров И.С., сумма 1000 р.)	☒
Принять решение	Назначение преподавателю руководства НИРС (Сергеев И.И., 3 ст.)	☐
Принять решение	Рекомендация для поступления в аспирантуру (Гребенщиков С.М.)	☐
Принять решение	Премирование преподавателя (Сысоев П.И.)	☐

Рис. 7. Страница просмотра и принятия управленческих решений

Система предполагает разграничение полномочий после ввода пользователями логина и пароля. Механизм авторизации реализован при помощи технологии Windows Forms.

В данной работе визуализация является дополнительным средством для эффективного отображения и интеллектуального анализа данных о результатах НИРС. В качестве инструментального средства визуализации данных использовался Novospark Visualizer 2.1. В среду Novospark Visualizer были импортированы представления данных из OLAP-куба результатов НИРС, в которых отражены общие количественные показатели результатов НИРС. На основе этих данных была построена трехмерная поверхность, оси которой соответствуют порядковым номерам студентов, числу результатов и времени этих результатов. Поверхность сформирована из совокупности кривых, отражающих результаты НИРС каждого из студентов. Трехмерное визуальное

представление данных в виде поверхности позволяет сделать определенные выводы о тенденциях, связанных с НИРС. Кроме того, была проведена кластеризация данных с целью выделения студенческих проектных групп для выполнения курсовых, дипломных проектов, участия в научных мероприятиях и т.д. В системе были выявлены 6 кластерных групп.

Следующим набором данных, загруженным в среду Novospark Visualizer стала выборка студентов, содержащая сведения числовых характеристики результатов НИРС по отдельным видам работ. На основе этих данных была построена трехмерная поверхность, оси которой соответствуют порядковым номерам студентов, числу результатов и виду НИРС. Зависимость количественных показателей результатов НИРС от вида НИРС также была представлена в виде кластеров, содержащих группы студентов для реализации командных проектов. В результате были определены 6 кластерных групп. Средняя степень сходства соответствующего кластера из данной выборки с кластером выборки количественных показателей результатов НИРС от времени является достаточно высокой и составляет 71%, что может служить основанием для определения студенческих групп на основе этих кластеров. Таким образом, визуализация данных служит не только для повышения наглядности отображаемой информации, но и содержит в себе возможности интеллектуальной обработки данных, позволяя находить закономерности в наборах данных.

Кроме того, в третьей главе описывается использование алгоритмов и моделей системы в ФГУП «Томской электротехнический завод» Учебном центре ТПУ & Softline Academy.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Исследования, проведенные в диссертационной работе, направлены на решение актуальной важной народнохозяйственной задачи создания математического и программного обеспечения системы оперативной обработки и интеллектуального анализа данных, использующей нечеткую логику и обеспечивающей повышение эффективности процесса обработки данных.

К основным результатам, в которых отражена научная новизна и практическая значимость данной работы, относятся:

1. Проведен анализ моделей обработки данных и классов математических моделей описания данных, на основе которого обоснован выбор классов моделей разрабатываемой системы.
2. Разработана математическая модель системы интеллектуальной обработки данных, основанная на аппарате нечеткой логики, позволяющая повысить эффективность обработки данных применительно к существующим задачам.
3. Разработана OLAP-модель данных системы, а также средства интеллектуального анализа данных, которые наделяют систему возможностями оперативной аналитической обработки информации.

4. Разработан алгоритм обработки данных в системе, использующий нечеткую логику, аналитическую обработку данных в реальном времени и механизм извлечения данных.
5. Разработана модель, включающая процедуры нечеткого логического вывода и обработки данных в рамках одной программной системы.
6. Создана система оперативной обработки и интеллектуального анализа данных, использующей нечеткую логику, которая позволяет повысить эффективность процесса работы с данными.
7. Проведена визуализация данных на основе программного пакета Novospark Vizualizer, позволяющая представить данные в наглядной форме и обнаружить скрытые закономерности в наборах данных.
8. Разработанные модели, алгоритмы и технологии созданной системы применены для организации обработки данных в Учебном центре ТПУ & SoftLine Academy и в ФГУП «Томской электротехнический завод», что подтверждено соответствующими актами о внедрении.

В процессе исследования использовались следующие теории и методы: теория реляционной модели данных, аппарат математической логики, методы объектно-ориентированного программирования, методы аналитической обработки информации. В процессе разработки использовались технологии ASP.Net, OLAP, data mining, KPI, тренды. Для исследования были применены следующие программные средства: Design/IDEF 3.7, MS Office Word 2007, MS Office Excel 2007, MS Office Visio 2007, Rational Rose 2000, MS SQL Server 2008, MS Visual Studio 2008, Novospark Visualizer 2.1.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Вахитов, А.Р. Использование KPI, технологий OLAP И data-mining при обработке данных / А.Р. Вахитов // Известия Томского политехнического университета. – 2009. – Т. 314. – № 5. – С. 175–179.
2. Вахитов, А.Р. Преимущества использования OLAP-технологии при обработке данных / А.Р. Вахитов // Технологии Microsoft в теории и практики программирования: Сборник трудов VI Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых: Томск, 2009. – С. 138–140.
3. Вахитов, А.Р., Новосельцев, В.Б. OLAP-технология как способ обработки данных / А.Р. Вахитов, В.Б. Новосельцев // Современные техника и технологии СТТ 2007: Материалы XIII Международной научно-практической конференции студентов и молодых ученых: Томск, 2007. – С. 342–344.
4. Вахитов, А.Р., Новосельцев, В.Б. Дескриптивная логика как способ обработки знаний / А.Р. Вахитов, В.Б. Новосельцев // Молодежь и современные информационные технологии: Материалы IV Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых: Томск, 2008. – С. 57–59.

5. Вахитов, А.Р., Новосельцев, В.Б. Обработка знаний при помощи дескриптивной логики / А.Р. Вахитов, В.Б. Новосельцев // Информационные системы и модели в научных исследованиях, промышленности и экологии: Материалы всероссийской научно-технической конференции: Тула, 2007. – С. 11–13.
6. Вахитов, А.Р., Новосельцев, В.Б. Оценка эффективности преобразования реляционной модели в OLAP-модель / А.Р. Вахитов, В.Б. Новосельцев // Молодежь и современные информационные технологии: Материалы IV Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых: Томск, 2007. – С. 146–148.
7. Вахитов, А.Р., Новосельцев, В.Б. Преимущества дескриптивной логики при обработке знаний / А.Р. Вахитов, В.Б. Новосельцев // Известия Томского политехнического университета. – 2008. – Т. 313. – № 5. – С. 73–76.
8. Вахитов, А.Р., Соколова, В.В. Decision support system for quality management / А.Р. Вахитов, В.В. Соколова // Снежинск и наука – 2006: Материалы VII Международной научно-практической конференции: Снежинск, 2006. – С. 172–174.
9. Вахитов, А.Р., Соколова, В.В. Информационные технологии в системах поддержки принятия решений / А.Р. Вахитов, В.В. Соколова // Математическое моделирование и информационные технологии: Материалы VIII Школы-семинара молодых ученых: Улан-Удэ, 2006. – С. 231–236.
10. Вахитов, А.Р., Соколова, В.В. Система поддержки принятия решений руководителя кафедры / А.Р. Вахитов, В.В. Соколова // Молодежь и современные информационные технологии: Материалы IV Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых: Томск, 2006. – С. 35–37.
11. Вахитов, А.Р., Соколова, В.В. Система поддержки принятия решений для менеджмента качества / А.Р. Вахитов, В.В. Соколова // Энергия молодых – экономике России: Материалы VII Международной научно-практической конференции студентов и молодых ученых: Томск, 2006. – С. 89–91.
12. Вахитов, А.Р. Использование интеллектуального анализа данных на основе методологии деревьев решений и его реализация в среде BIDS / А.Р. Вахитов // Технологии Microsoft в теории и практики программирования: Сборник трудов VII Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых: Томск, 2010. – С. 81–83.
13. Вахитов А.Р. Программная система интеллектуальной обработки данных на основе нечеткой логики «WinDIFOT» // Свидетельство о регистрации в объединенном фонде электронных ресурсов «Наука и образование» №15692 от 12.05.2010г.