На правах рукописи

Вохминцев Александр Владиславович

МЕТОДИКА ИЗВЛЕЧЕНИЯ СТРУКТУРНЫХ ЗНАНИЙ ИЗ ЕСТЕСТВЕННЫХ ТЕКСТОВ НА ОСНОВЕ НЕЧЕТКИХ СЕМАНТИЧЕСКИХ ГИПЕРСЕТЕЙ

Специальность 05.13.01 – "Системный анализ, управление и обработка информации (промышленность)"

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Челябинск – 2002

Работа выполнена в Южно-Уральском государственном университете.

Научный руководитель – доктор технических наук, профессор Мельников А.В.

Официальные оппоненты:
доктор технических наук, профессор Ширяев В.И.,
кандидат технических наук, доцент Крушный В.В.

Ведущая организация –
Санкт-Петербургский государственный технический университет.

Защита состоится 4 декабря 2002 года, в 14 часов, на заседании диссертационного совета Д 212.298.03 при Южно-Уральском государственном университете по адресу: 454080, г. Челябинск, пр.им. В.И.Ленина, 76 (конференц-зал, ауд.244).

С содержанием диссертации можно ознакомиться в библиотеке Южно-Уральского государственного университета.

Автореферат разослан "29" октября 2002 г.

Ученый секретарь
диссертационного совета А.М. Коровин

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В предлагаемой диссертационной работе представлены результаты, полученные аспирантом в ходе научно-исследовательской и проектно-конструкторской деятельности в период с 1998 года по настоящее время, связанные с разработкой методики извлечения структурных знаний из естественных текстов и созданием на ее основе информационно-аналитической системы (ИАС) интеллектуального анализа информации.

Проблемы создания моделей, методик и систем для интеллектуального анализа информации рассматриваются в научных трудах Аверкина А.Н., Александера Т., Грэя П., Инмона Б., Кендалла М., Кисилева М., Кнута Д., Кодда С., Коровкина С.Д., Коско В., Куинлена Р., Логиновского О.В., Малышева Н.Г., Мельчука И.А., Минского М., Нелюбина Л.Л., Новака Д., Орловского С.А., Осипова Г.С., Осуги С., Попова Э.В., Попова Э.Ю., Поспелова Д.А., Рубашкина В.Ш., Скороходько Э.Ф., Скрэгга Г., Уоссермена Ф., Шенка Р., Эдвардса Д. и др.

В диссертационной работе использованы подходы и методы проектирования информационно-аналитических систем или их составных частей, представленных в научных работах Инмона Б., Кнута Д., Коско В., Куинлена Р., Малышева Н.Г., Осуги С., Попова Э.В., Новака Д., Скороходько Э.Ф.

Актуальность темы. Современный этап развития общества характеризуется взрывным развитием информационных технологий, которые позволяют накапливать большие объемы информации. Однако огромный размер не позволяет использовать эту информацию для непосредственного анализа ситуации и выработки управленческих решений в промышленности. Поэтому особенно актуальными на сегодняшний день являются технологии анализа больших объемов информации, которые можно поделить по уровню анализа знаний на технологии класса OLTP, OLAP и Data Mining. Наибольшие возможности для аналитиков предоставляет технология Data Mining, которая позволяет получать ранее неизвестные закономерности, носящие не универсальный характер, в больших объемах информации.

Важными источниками исходной информации для аналитической обработки являются средства массовой информации, Internet, хранилища данных различных организаций и учреждений, в которых информация в основном представлена в виде естественных текстов. Задача аналитической обработки естественных текстов является достаточно сложной и в общем случае связана с построением систем с искусственным интеллектом. Основной тенденцией, особенно в России, при разработке подобных систем является упор на задачи классификации документов по релевантности, составления автореферата документа, контекстного анализа документов или автоматического перевода текстов. В данном направлении разработано достаточно много коммерческих программных продуктов, среди которых следует отметить Intelligent Miner for Text (IBM), Natural Language Projects at ISI (University of Southern California), Text Analyst (Научно-производственный инновационный центр "Микросистемы"), Web-Analyst (Мегапьютер), Semantic Exproler (НейроОК), Text Miner (SAS), Russian Context Optimizer (Гарант-Парк-Интернет), Cibola Oleada (Computing

Research Laboratory New Mexico State University), Серия программных продуктов фирмы LingSoft, Галактика-Zoom (Корпорация Галактика), Интернет-паук (NooLab), ПРОМТ (ЗАО "ПРОект-МТ"). Данные системы пользуются коммерческим успехом на рынке программного обеспечения, несмотря на это поставленная задача в них не решается на достаточном уровне.

Информационно-аналитические службы не нуждаются в извлечении всех закономерностей из естественных текстов, поэтому нет необходимости в построении модели естественного текста, реализующей глубинный семантический анализ текста. Одной из важных задач, решаемых аналитиками, является определение отношений между объектами, которыми являются физические и юридические лица. Наиболее адекватно отношения между объектами представляются семантическими сетями. Традиционная интерпретация семантической сети позволяет получать только представление о структуре отношений между объектами, которой недостаточно для проведения полноценного аналитического исследования. Поэтому в диссертационной работе предлагается расширение семантической сети до нечеткой семантической гиперсети для представления информации о типах отношений между объектами и о принадлежности объектов к классам предметной области. Модель знаний в виде нечеткой семантической гиперсети позволяет перейти на более высокий уровень представления информации (естественный для мышления человека), который связан с введением качественных категорий отношений между объектами. Однако для организации эффективной работы аналитиков необходимо разработать методику построения базы знаний из естественных текстов на основе автоматических процедур, которые соответствуют устоявшимся технологиям работы информационно-аналитических служб. На основании сказанного выше сформулирована основная цель диссертационной работы и задачи исследования.

Цель диссертационной работы и задачи исследования. Целью диссертационной работы является разработка методики извлечения структурных знаний из естественных текстов на основе нечетких семантических гиперсетей.

Для достижения поставленной цели в диссертационной работе решаются следующие задачи:

1) создание модели знаний для представления отношений между объектами в естественном тексте;

2) создание модели метатекста, формализующей конструкции русского языка на основе неполного синтаксического анализа;

3) разработка метода классификации объектов на основе деревьев решений;

4) разработка процедур построения базы знаний на основе метатекста;

5) разработка методов извлечения структурных знаний из базы знаний;

6) разработка архитектуры программного комплекса ИАС для извлечения ассоциаций из естественных текстов.

Методы исследования. Теоретической и методологической основой диссертационного исследования являются методы теории нечетких множеств, теории графов, теории баз данных и знаний, методы искусственного интеллекта и прикладной лингвистики.

Научная новизна диссертационной работы заключается в следующем:

- Предложена новая формализация представления знаний в виде нечеткой семантической гиперсети, которая отражает дуальный характер отношений между объектами в естественных текстах;
- Разработана модель метатекста русского языка, которая позволяет формализовать процесс обработки естественного текста за счет отказа от глубинного семантического анализа естественного текста;
- Предложена методика на основе нечеткой семантической гиперсети и модели метатекста, которая позволяет извлекать структурные знания из естественных текстов в виде отношений между объектами при помощи многоуровневого подхода к процессу обработки информации.

Практическая ценность и реализация результатов работы заключается в создании методики извлечения структурных знаний из естественных текстов для решения актуальной задачи по автоматизации процессов аналитической обработки больших объемов информации в информационно-аналитических службах. Полученная методология закладывает основу для создания ИАС класса Data Mining, которая позволяет:

- существенно увеличить эффективность работы информационно-аналитических служб за счет автоматизации времяемких процессов аналитической обработки информации, требующих привлечения высококвалифицированных кадров;
- извлекать ассоциации, которые ранее были неизвестны аналитику и носят не универсальный характер, из информации в виде естественных текстов.

Предложенные в диссертационной работе методики, алгоритмы и программные компоненты использованы: в практической работе информационно-аналитических служб Южно-Уральской железной дороги, ЗАО "Интерсвязь", ЗАО "Цветлит", при получении дополнительного профессионального образования в Челябинском институте развития профессионального образования, а также составили основу спецкурса по технологиям Data Mining для направления – 654600 "Информатика и вычислительная техника" и для специальности 220100 – "Вычислительные машины, комплексы, системы и сети" Южно-Уральского государственного университета.

Апробация работы и публикации. Основные положения и результаты, полученные в диссертационной работе, представлены и обсуждены на следующих конференциях и семинарах:

- Международной конференции "Информационные технологии в управлении промышленностью и экономикой субъектов РФ" (Челябинск, 2002)
- Международной конференции "Актуальные проблемы современной науки" (Самара, 2002);
- Межрегиональном научно-практическом семинаре "Компьютерные системы поддержки принятия решения руководителей" (Челябинск, 2001);
- Межрегиональном научно-практическом семинаре "Информационно-аналитические компьютерные системы и технологии в региональном и муниципальном управлении" (Челябинск, 2000).

По теме диссертационной работы опубликовано 9 печатных работ.

Связь с государственными и региональными программами. Диссертационная работа связана с тематикой работ, осуществляемых в соответствии с Федеральной целевой программой "Электронная Россия".

Структура и объем работы. Диссертационная работа включает введение, четыре главы, заключение, список литературы (142 наименования), а также приложение. Диссертация содержит 210 страниц (181 страницу основного текста), 48 иллюстраций, 17 таблиц.

Основные положения, выносимые на защиту:

- адекватным способом представления отношений между объектами в естественных текстах является нечеткая семантическая гиперсеть;
- для эффективного анализа естественных текстов с учетом особенностей русского языка необходим переход к модели метатекста на основе теории нечетких множеств;
- результативность алгоритмов построения базы знаний из естественных текстов связана с ограничением общей задачи до подзадачи определения отношения между объектами;
- целостность моделей, методик и программной системы основана на многоуровневом подходе к представлению процесса извлечения отношений между объектами.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Извлечение ассоциаций из естественных текстов представляет собой многоуровневый процесс обработки информации, состоящий из следующих уровней: уровень исходной информации, хранилище данных, метатекст, база знаний, ассоциации. Информация из различных источников: средства массовой информации, государственные и ведомственные службы, информационно-аналитические агентства, собирается по определенным регламентам в хранилище данных. Из хранилища данных информация отбирается для построения базы знаний, при этом для агрегации и учета релевантности информации используются фильтры. Отобранная информация обрабатывается на основе модели метатекста лингвистическим процессором, который в результате своей работы генерирует метатекст. Метатекст используется методами Data Mining для построения базы знаний. Из базы знаний с помощью методов извлечения ассоциаций аналитик получает ассоциации, на основе которых можно восстанавливать исходные естественные тексты.

При разработке теоретико-методологического описания ИАС определены требования к модели знаний, на основе которых произведен выбор в качестве базовой модели знаний семантической сети, которая далее расширяется за счет представления информации о типах отношений между объектами и о принадлежности объектов к классам предметной области. Расширенная семантическая сеть, которая в дальнейшем будет называться *семантической гиперсетью*, позволяет аналитику изменять уровень агрегации информации на основе иерархии классов предметной области. На основе теории множеств введем основные формализации уровня метатекста (объект и типовое отношение).

Назовем кортеж $O = \langle a_{название}, a_1, a_2, \ldots, a_n \rangle$ **объектом**, где $a_{название}$ – название объекта, $a_1, a_2, \ldots, a_n$, $a_i, i = \overline{1,n}$ – свойства объекта из предметной области, такие что $\bigcap_{i=1}^{n} a_i = \varnothing$.

Назовем кортеж $R = \langle r_{название}, r_1, r_2, \ldots, r_m \rangle$ **типовым отношением**, где $r_{название}$ – название типового отношения, $r_1, r_2, \ldots, r_m$, $r_i, i = \overline{1,m}$ – действия между объектами из предметной области такие, что $\bigcap_{i=1}^{m} r_i = \varnothing$ и $r_{название} \bigcap \{r_1, r_2, \ldots, r_m\} \neq \varnothing$.

Свойствами объекта является ограниченный на основе иерархии классов, набор имен нарицательных, действиями между объектами является ограниченный на основе предметной области, набор глаголов, причастий и деепричастий. Типовое отношение, по сути, представляет действие, которое является результатом агрегирования группы глаголов и его форм вокруг глагола на основании семантической идентичности группы данному глаголу. При этом для типовых отношений справедливо утверждение: $R_\alpha \bigcap R_\beta = \varnothing \,|\, \forall(\alpha, \beta)$.

Назовем кортеж $R_j^{\Delta} = \langle O_\alpha, O_\beta, R_j \rangle$ **дифференциальным отношением между объектами**, где O_α, O_β – объекты, такие что $\alpha \neq \beta$, $R_j, j = 1, p$ – типовое отношение, p – количество типовых отношений в модели знаний. Назовем кортеж $R^{\Sigma} = \langle O_\alpha, O_\beta, R_1, R_2, \ldots, R_p \rangle$ **интегральным отношением между** объектами, где O_α, O_β – объекты, такие что $\alpha \neq \beta$, $R_1, R_2, \ldots, R_p$, $R_j, j = 1, p$ – полное множество типовых отношений.

Объект O соответствует вершине семантической гиперсети, а типовое отношение R – ребру семантической гиперсети. Далее для краткости, если характер отношения – дифференциальное или интегральное – не имеет значения, будем называть его просто "отношением" и обозначать $R^{\Sigma-\Delta}$.

Вершиной семантической гиперсети назовем кортеж свойств $x = \langle a_{название}, a_{синонимы}, a_{ключ\ норм}, a_{классы\ объекта}, a_{R^{\Sigma-\Delta}(\tilde{H})}, a_{ассоц} \rangle$:

1) *название* является уникальным для каждой вершины семантической гиперсети и позволяет идентифицировать вершину при извлечении ассоциаций;

2) *синонимы* позволяют устанавливать связи между вершинами семантической гиперсети на основе их принадлежности к одному объекту в реальном мире, список синонимов для вершины определяется аналитиком;

3) *ключ нормализации* находится во взаимооднозначном соответствии со свойством название, при этом название является метаименем для свойства ключ нормализации;

4) *классы объекта* устанавливают принадлежность объекта, соответствующего вершине семантической сети, к определенному классу предметной области в пределах одной статьи;

5) *матрица отношений* содержит информацию об отношениях между объектом из данной вершины с объектами из других вершин семантической гиперсети;

6) *ассоциации* содержат список файлов закономерностей с данной вершиной, полученных в результате извлечения ассоциаций.

Ребром семантической гиперсети назовем кортеж свойств $e=\left\langle \mu_{R(\tilde{H})^{\Sigma-\Delta}}(x_\alpha,x_\beta)/(x_\alpha,x_\beta), a_{тип\ отн}, a_{ключ\ норм}, a_{сила\ отн}, a_{парам\ ассоц}, a_{ссылки}\right\rangle$:

1) *тип отношения* определяется только для дифференциальных отношений и идентифицирует принадлежность к одному из типовых отношений;

2) *ключ нормализации* находится во взаимно однозначном соответствии со свойством тип отношения, при этом тип отношения является метаименем для свойства ключ нормализации;

3) *сила отношения* определяется: классами объектов в отношении, типом отношения, частотой появления отношения, лингвистической характеристикой отношения;

4) *параметры ассоциации* позволяют определить значимость данной ассоциации среди других ассоциаций в семантической гиперсети;

5) *ссылки* устанавливают соответствие между отношениями в семантической гиперсети и анализируемым естественным текстом.

Понятия дифференциального и интегрального отношения между объектами можно перенести на отношения между вершинами семантической гиперсети. Назовем интегральным отношением между вершинами $x_\alpha, x_\beta \mid \alpha \neq \beta$ семантической гиперсети выражение $R(\tilde{H})^{\Sigma}=\left\langle x_\alpha, x_\beta, R_1, R_2, \ldots, R_p\right\rangle \equiv \left\langle x_\alpha, x_\beta, R_\Sigma\right\rangle$, а дифференциальным отношением между вершинами семантической гиперсети $x_\alpha, x_\beta \mid \alpha \neq \beta$ назовем выражение $R(\tilde{H})^{\Delta}_j=\left\langle x_\alpha, x_\beta, R_j\right\rangle$.

Пусть $X=\{x_i\}$, $i=\overline{1,n}$ – конечное множество вершин семантической гиперсети, где n – количество объектов в модели знаний и $\tilde{E}=\{\tilde{e}_k\}$, $k=\overline{1,m}$ – множество ребер семантической гиперсети. Тогда **нечеткой семантической гиперсетью** (НСГС) назовем кортеж $\tilde{H}=\left\langle X, \tilde{E}\right\rangle$.

Отношения между вершинами семантической гиперсети носят нечеткий характер, который определяется лингвистической переменной A – "сила интегрального отношения". Определим множество ее значений (терм-множества):

T={очень значимое, значимое, значимое больше среднего, средняя значимость, значимое ниже среднего, незначимое, абсолютно незначимое}. В качестве базового множества возьмем последовательность целых чисел с шагом 1: $t^{\Sigma} = \{0,1,2,3,\ldots,S^{\Sigma}\}$, где S^{Σ} – максимальная граница базового множества. Для интегральных отношений между вершинами максимальная граница базового множества определяется по формуле

$$S^{\Sigma} = \bigvee_{i=1}^{n} \bigvee_{i'=1}^{n} \mu_{R(\tilde{H})^{\Sigma}}(x_i, x_{i'}), \qquad (1)$$

где $S^{\Sigma} / K_{A_q}, q = \overline{1,7}$ – коэффициенты термов лингвистической переменной A, которые позволяют определить максимальные значения функций принадлежности, при этом справедливы утверждения: $0 < K_{A_q} \le 1$ и $K_{A_7} < K_{A_6} < K_{A_5} < K_{A_4} < K_{A_3} < K_{A_2} < K_{A_1}$.

Аналогичным образом определим множество лингвистических переменных $B = \{B^1, B^2, \ldots, B^p\}, B^j, j = \overline{1,p}$ – "сила дифференциального отношения" для каждого дифференциального отношения $R(\tilde{H})_j^{\Delta}$. Определим множество их значений: *T={очень значимое, значимое, значимое больше среднего, средняя значимость, значимое ниже среднего, незначимое, абсолютно незначимое}*. В качестве базового множества возьмем последовательность целых чисел с шагом 1: $t_j^{\Delta} = \{0,1,2,3,\ldots,S_j^{\Delta}\}$, где S_j^{Δ} – максимальная граница базового множества. Для дифференциальных отношений между вершинами максимальная граница базового множества определяется по формуле

$$S_j^{\Delta} = \bigvee_{i=1}^{n} \bigvee_{i'=1}^{n} \mu_{R(\tilde{H})_j^{\Delta}}(x_i, x_{i'}), \qquad (2)$$

где $S_j^{\Delta} / K_{B_q^j}, q = \overline{1,7}$ – коэффициенты термов лингвистической переменной B^j, которые позволяют определить максимальные значения функций принадлежности, при этом справедливы утверждения: $0 < K_{B_q^j} \le 1$ и $K_{B_7^j} < K_{B_6^j} < K_{B_5^j} < K_{B_4^j} < K_{B_3^j} < K_{B_2^j} < K_{B_1^j}$.

Значения функций принадлежности μ_F лингвистической переменной A (B^j) определяются относительно максимального значения базового множества S^{Σ} (S_j^{Δ}), которое динамически изменяется в зависимости от информации в НСГС. Поэтому значения функций принадлежности μ_F также будут динами-

чески изменяться. Такой подход в измерении силы отношения позволяет адекватно определять значимость информации в базе знаний.

Назовем **матрицей отношений** для вершины НСГС x_α такую матрицу $R(\tilde{H}) = \| r_{ik} \|_{n \times m}$,

где n – вектор-строк представляет вершины семантической гиперсети;

m – вектор-столбцов представляет ребра семантической гиперсети;

$k=q'$ – для интегральных отношений и $k = q' + j$ – для дифференциальных отношений;

q' – база вершины x_α (номер столбца в матрице отношений, с которого начинают определяться отношения между вершиной НСГС x_α и другими вершинами НСГС x_i).

Элемент матрицы равен степени смежности вершины семантической гиперсети x_α с другими вершинами НСГС x_i. В качестве значения элемента матрицы отношений выбирается терм лингвистической переменной с наибольшим значением функции принадлежности.

Перед тем как приступить к анализу информации в базе знаний, необходимо предварительно исключить большую часть несущественной информации из базы знаний. Для этого используются следующие ***методы извлечения ассоциаций***:

Метод "глубина транзитивных отношений" определяет множество вершин НСГС, достижимых из вершины x_i (объект анализа) при помощи нечетких цепей $\tilde{C}(x_i, x_{q+1})$ с максимально допустимой длиной q.

Метод "принадлежность объекта к классу" определяет принадлежность классов объектов в вершинах НСГС $a^{x_i}_{\text{классы объекта}}$ к классам модели знаний $c = \{c_1, c_2, \ldots, c_M\}, c_{i^2}, i^2 = \overline{1,M}$ для каждого уровня глубины транзитивных отношений, где M – количество классов в модели знаний. Если класс не определен для некоторого уровня $0 < l \le q$, то для последующего анализа отбираются все вершины НСГС x_i^l l- уровня, иначе необходимо проверить условие: $a^{x_i}_{\text{классы объекта}} \cap c^l \neq \varnothing$, где $c^l = \{c_1^l, c_2^l, \ldots, c_s^l\}, c_i^l, i = \overline{1,s}, c^l \subseteq c = \{c_1, c_2, \ldots, c_M\}$ – множество классов l - уровня, определяемое аналитиком. Если объект в вершине НСГС x_i^l принадлежит хотя бы одному из классов во множестве, то вершина НСГС отбирается для дальнейшего анализа, иначе вершина НСГС исключается из базы знаний.

Метод "сила отношения" устанавливает принадлежность силы отношения ребра НСГС к терму лингвистической переменной соответствующей данному отношению. Если значение лингвистической переменной не определено для отношения, то все ребра НСГС, для которых выполняется условие $\tilde{e} = \{< \mu_{R(\tilde{H})^{\Sigma-\Delta}}(x_i, x_j) > | i \neq j\}$, исключаются из дальнейшего анализа. Если значение лингвистической переменной определено для отношения, то при исключении ребер НСГС существует три стратегии: отбирать ребра НСГС по равенству силе отношения, отбирать ребра НСГС по отношению больше (больше или равно) силе отношения, отбирать ребра НСГС по отношению меньше (меньше или равно) силе отношения. Дополнительно для каждого значения лингвистической переменной может определяться степень принадлежности, которая не должна превышать некоторого значения $\Omega(0<\Omega<1)$, задаваемого аналитиком: $a^{\tilde{e}k}_{сила\ отн}\{b_{значение}\} \geq \Omega$.

Метод "тип отношения" определяет принадлежность типа отношения ребра НСГС к одному из типовых отношений в модели знаний. Если $a^{\tilde{e}k}_{тип\ отн} = \varnothing$ (интегральное отношение), то эти ребра отбираются для дальнейшего анализа, иначе $a^{\tilde{e}k}_{тип\ отн} \neq \varnothing$ (дифференциальное отношение) необходимо проверить условие $a^{\tilde{e}k}_{тип\ отн} \bigcap \{R'_1, R'_2, \ldots, R'_z\} \neq \varnothing$, где $R' = \{R'_1, R'_2, \ldots, R'_z\}$ R'_j, $j = \overline{1,z}$ – подмножество типовых отношений, выбранных аналитиком из множества типовых отношений в модели знаний $R = \{R_1, R_2, \ldots, R_p\}$, $R' \subseteq R$. Если множество $R' = \varnothing$, то для последующего анализа отбираются все ребра НСГС, иначе отбираются только те ребра НСГС, для которых справедливо условие.

Методы извлечения асоциаций управляются соответствующими *параметрами извлечения ассоциаций*, тогда запрос к нечеткой семантической гиперсети можно представить в виде кортежа:

$$I = \left\langle x^*_\alpha, q^*, \Gamma_1(x_\alpha)^*, \Gamma_2^2(x_\alpha)^*, \ldots, \Gamma_l^q(x_\alpha)^*, A^*, B^{1^*}, B^{2^*}, \ldots, B^{p^*}, R'^*, X^* \right\rangle, \qquad (3)$$

где x^*_α – параметр объект анализа; q^* – параметр длина нечеткой цепи;

$\Gamma_l^q(x_\alpha)^* = < c_1^{l^*}, c_2^{l^*}, \ldots, c_s^{l^*}, x^*_\rho >$ – кортеж параметров, $c_i^{l^*}$ – параметр класс l-уровня, x^*_ρ – параметр конкретный объект;

$A^* = \left\langle A^*_q, \Omega^*, \Psi^* \right\rangle$ – кортеж параметров, A^*_q – параметр значение лингвистической переменной A, Ω^* – параметр степень принадлежности, $\Psi^* = \{=, >, <, \geq, \leq\}$ – параметр стратегии отбора по методу "сила отношения";

$B^{P^*} = \left\langle B_q^{P^*}, \Omega^*, \Psi^* \right\rangle$ – кортеж параметров, $B_q^{P^*}$ – параметр значение лингвистической переменной B_q^P, Ω^* – параметр степень принадлежности, $\Psi^* = \{=, >, <, \geq, \leq\}$ – параметр стратегии отбора по методу "сила отношения";

$R^{/^*} = \{R_1^{/^*}, R_2^{/^*}, \ldots, R_z^{/^*}\}$ $R_j^{/^*}$ – параметр типовое отношение;

$X^* = \{x_1^*, x_2^*, \ldots, x_y^*\}$, $x_j^*, j=\overline{1,y}$ – параметр множество объектов, подлежащих исключению из анализа.

В диссертации разработана ***методика построения базы знаний на основе нечеткой семантической гиперсети из естественных текстов на русском языке***. Естественный текст представляет собой трудно формализуемую информацию с преобладанием качественных отношений между именами собственными, которые носят нечеткий нелинейный характер, особенно это свойственно для русского языка. На основе целевой модели знаний можно сделать вывод, при анализе естественного текста интерес представляют только отношения между именами собственными. Поэтому модель метатекста русского языка не требует проведения полноценного лингвистического анализа естественного текста.

Назовем **глагольным отношением между именами собственными** кортеж $R^V = \langle N_\alpha, N_\beta, V \rangle$, где N_α, N_β – имена собственные; V – глагол или его форма, при этом $V = r_i^{R_j}$; r_i – действие типового отношения R_j.

Назовем **ассоциативным отношением между именами собственными** кортеж $R^A = \langle N_\alpha, N_\beta \rangle$, где N_α, N_β – имена собственные.

Смысловая связь через глаголы и его формы между именами собственными (местоимениями) имеет место только при отношениях посредством простых предложений, причастных, деепричастных оборотов, сложноподчиненных предложениях. В остальных случаях можно говорить об ассоциативных отношениях между именами собственными (местоимениями) с некоторой степенью силы отношения.

Отношения между именами собственными носят нечеткий характер, который определяется лингвистической переменной C – "синтаксическая конструкция" теории нечетких множеств. Определим множество ее значений: *T={словосочетание, простое предложение, причастный оборот, деепричастный оборот, сложноподчиненное предложение, сложносочиненное предложение, бессоюзное предложение, предложение без главных членов, транзитивное отношение}*. В качестве базового множества возьмем последовательность целых чисел от 0 до 100 с шагом 1: $t^{V-A} = \{0,1,2,3,\ldots,100\}$. Определение принадлежности синтаксической конструкции между именами собственными к терм-

множеству лингвистической переменной C происходит на основе морфологического анализа, в результате которого каждой морфе – слову или комбинации слова с некоторым знаком пунктуации – в нормализованном естественном тексте присваивается весовой эквивалент. После этого для каждого глагольного или ассоциативного отношения определяется *функция лингвистики*:

$$f(R^{V-A}) = 100 - \sum_{i=1}^{E} w_i , \qquad (4)$$

где E – количество морф в отношении R^{V-A}; w_i – весовой эквивалент морфы.

По значению функции лингвистики отношения определяется принадлежность отношения между именами собственными R^{V-A} к терм-множеству $C_q, q = \overline{1,9}$ лингвистической переменной.

На основании модели метатекста определена структура лингвистического процессора. Естественный текст обрабатывается фильтром, который производит его унификацию. Унифицированный естественный текст обрабатывается морфологическим анализатором, результатом работы которого является нормализованный естественный текст (рис.1).

Слово	Часть речи	Ключ нормальной формы	Признаки слова

Рис.1. Формат нормализованного естественного текста

Нормализованный естественный текст обрабатывается синтаксическим анализатором, результаты работы которого представляются в виде таблиц отношений между именами собственными (рис.2-а) и таблиц связей между именами собственными и именами нарицательными (рис.2-б).

Имя собственное	Глагол или его форма	Функция лингвистики	Имя собственное

а)

Имя собственное	Имя нарицательное

б)

Рис.2. Результаты синтаксического анализа: а) отношения между именами собственными; б) связи между именами собственными и именами нарицательными

Генератор метатекста преобразует отношения между именами собственными R^{V-A} в отношения между объектами $R^{\Sigma-\Delta}$ (рис.3-а). При генерации метатекста также происходит агрегация таблиц (рис.2-б) в таблицу свойства объекта (рис.3-б). Кроме таблиц формата (рис.3) в метатекст включается нормализованный естественный текст, при этом на основе индекса осуществляется взаимно однозначное соответствие между таблицей (рис.3-а) и нормализованным

естественным текстом. Индекс содержит номер предложения в нормализованном естественном тексте, в котором встречается соответствующее отношение между объектами.

Объект	Типовое отношение	Синтаксическая конструкция C_q	Объект	Индекс

а)

Объект (название)	Свойство объекта	. . .	Свойство объекта

б)

Рис.3. Метатекст: а) отношения между объектами; б) свойства объектов

Классификация объектов, соответствующих вершинам НСГС, осуществляется на основе иерархии классов предметной области. Иерархия классов представляет дерево с произвольным количеством ветвлений, при этом множества свойств классов потомков являются вложенными во множества свойств классов родителей.

Назовем кортеж $\delta = \langle a_{название}, a_{соб}, a_{влож} \rangle$ **классом** объекта, если выполняются условия: $a_{влож} \setminus a_{соб} = \varnothing$, $\bigcap_{i=1}^{p} a_i = \varnothing$, $\bigcap_{j=1}^{s} b_j = \varnothing$, $a_i \cap \delta_1\{b_j\} = \varnothing \mid \forall (i,j)$, где $a_{название}$ – название класса; $a_{соб} = \{a_1, a_2, \ldots, a_p\}$, a_i, $i = \overline{1, p}$ – собственные свойства предметной области класса; $a_{влож} = \{b_1, b_2, \ldots, b_s\}$, b_j, $j = \overline{1, s}$ – вложенные множества свойств классов, где s – число вершин потомков для данной родительской вершины.

Иерархия классов составляется аналитиками-экспертами на основе анализа предметной области их деятельности. Правила построения иерархии классов гарантируют, что каждый класс будет иметь уникальное множество собственных свойств $a_{соб}$ и уникальные множества вложенных свойств $a_{влож}$. При этом корневой класс δ_1 будет содержать все вложенные множества в иерархии классов, а терминальные классы, которые не имеют потомков, не будут содержать ни одного вложенного множества.

На основе иерархии классов осуществляется классификация объектов по методу С4.5. Критерий разбиения в стандартной версии метода С4.5 определяется следующим образом:

$$I_{\nabla}(тест) = I(O^{тест}) - I_{тест}(O^{тест}), \qquad (5)$$

где $I(O^{тест})$ – энтропия множества $O^{тест}$, объекты $O_1^{тест}, O_2^{тест}, \ldots, O_k^{тест}$ получены при разбиении исходного множества объ-

ектов O^{mecm} по проверке (свойству); $I_{mecm}(O^{mecm})$ – энтропия множества O^{mecm} после его разбиения по проверке. Критерий определяется для всех свойств дерева решений, а затем выбирается свойство с максимальным значением $I_{\nabla}(mecm)$. После чего по этому свойству производится дальнейшее построение дерева. Стандартная версия алгоритма C4.5 может строить смещенные деревья решений с большим количеством ветвлений, поэтому необходимо модифицировать критерий разбиения (5) с помощью семантического коэффициента поправки:

$$I_{\nabla}^{несмещ}(mecm) = k_{\nabla}^{сем} I_{\nabla}\ (mecm) = \left(\sum_{j=1}^{k} \frac{\left|O_j^{mecm}\right|}{\left|O^{mecm}\right|} \log_2 \frac{\left|O_j^{mecm}\right|}{\left|O^{mecm}\right|} \right)^{-1} I_{\nabla}\ (mecm). \quad (6)$$

Для перехода с уровня метатекста на уровень базы знаний необходимо определить *функцию семантики*, которая должна учитывать классы объектов в отношении, тип отношения, частоту появления отношения, лингвистическую характеристику отношения.

Лингвистическая характеристика отношения определяется лингвистической переменной C. В соответствие значениям C поставлены весовые эквиваленты $g_k, k = \overline{1,9}$, которые не позволяют дать высокую оценку отношению между объектами, полученному в результате опосредованных ассоциативных отношений R^A между именами собственными в естественном тексте. Для весовых эквивалентов справедливы отношения: $g_1 > g_2 > g_3 > g_4 > g_5 >> g_6 > g_7 > g_8 > g_9$. Соответствие между значением лингвистической переменной C_q в отношении $R^{\Sigma-\Delta}$ и весовым эквивалентом g_k устанавливает функция $F^g(O_x, O_y, R_j^{\Delta}, C)$. Так как отношение R_j^{Δ} между объектами в пределах одного метатекста может повторяться более одного раза, необходимо при определении функции семантики складывать лингвистические характеристики этих отношений.

Каждое типовое отношение оказывает различное влияние на силу отношения между вершинами НСГС. Это связано как с предметной областью деятельности аналитиков, так и с неравномерным распределением частоты появления разных типовых отношений в метатекстах. Поэтому необходимо введение *функции коррекции типового отношения* $\mu^{\Delta}(R_j)$, которая позволяет ранжировать типовые отношения. Функция $\mu^{\Delta}(R_j)$ является функцией принадлежности нечеткого множества $\tilde{R}$ с базовым множеством

$R = \{R_1, R_2, \ldots, R_P\}$, $R_j, j = \overline{1,p}$. Таблица значений функции $\mu^{\Delta}(R_j)$ определяется аналитиками-экспертами.

Аналогично объекты в отношениях $R^{\Sigma-\Delta}$ оказывают различное влияние на силу отношения между вершинами НСГС, поэтому необходимо введение *функции коррекции объекта* $\mu^{\delta}(c_{i^2}{}^{x_i})$, которая позволяет определять степень значимости появления объекта в отношении $R^{\Sigma-\Delta}$ на основе принадлежности объекта к одному из классов предметной области. Функция коррекции объекта является функцией принадлежности нечеткого множества $\tilde{\delta}$ с базовым множеством $c = \{c_1, c_2, \ldots, c_M\}, c_{i_2}, i^2 = \overline{1,M}$, где M – количество классов в модели знаний.

Таблица значений функции коррекции объекта определяется аналитиками-экспертами. Так как объект в вершине НСГС может иметь несколько классов, то после определения значения функции $\mu^{\delta}(c_{i^2}{}^{x_i})$ для каждого класса объекта из последовательности $\mu^{\delta}(c_1{}^{x_i}), \mu^{\delta}(c_2{}^{x_i}), \ldots, \mu^{\delta}(c_m{}^{x_i})$ выбирается минимальное значения функции: $f^{\delta}(O_x) = \underset{i^2=1}{\overset{m}{\&}} \mu^{\delta}(c_{i^2}{}^{x_i})$.

Так как отношение $R^{\Sigma-\Delta}$ устанавливается между двумя объектами, то функция $f^{\delta}(O_x)$ определяется для каждого объекта, после чего из двух значений функций выбирается минимальное: $f^{\delta}(O_x, O_y) = f^{\delta}(O_x) \,\&\, f^{\delta}(O_y)$.

Таким образом, функции семантики для интегрального $R(\tilde{H})^{\Sigma}$ и множества дифференциальных отношений $R(\tilde{H})^{\Delta}_j$ будут представимы в виде системы уравнений (7).

$$\begin{cases} A = \mu_{R(\tilde{H})^{\Sigma}}\left(f^{\delta}(O_x, O_y) \sum_{i=1}^{N} \sum_{j=1}^{p} \mu^{\Delta}(R_j) F_i^{g}(O_x, O_y, R_j, C) \right) \\ B^1 = \mu_{R(\tilde{H})^{\Delta}_1}\left(f^{\delta}(O_x, O_y) \sum_{i=1}^{N} F_i^{g}(O_x, O_y, R_1, C) \right) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ B^P = \mu_{R(\tilde{H})^{\Delta}_P}\left(f^{\delta}(O_x, O_y) \sum_{i=1}^{N} F_i^{g}(O_x, O_y, R_P, C) \right) \end{cases}, \qquad (7)$$

где N – количество отношений между объектами O_x, O_y в метатексте.

Практическим результатом применения методики извлечения структурных знаний из естественных текстов на основе нечетких семантиче-

ских гиперсетей является разработка архитектуры ИАС класса Data Mining на основе НСГС и программного комплекса "Analyst Wizard". Для хранилища данных ИАС выбрана технология хранилища данных Oracle. Чтобы попасть в хранилище данных информация должна представлять собой естественный текст на русском языке. Унификация естественных текстов в ИАС осуществляется программной системой Intelligent Miner for Text (IBM). Из хранилища данных информация отбирается для построения базы знаний. При этом для агрегации информации относительно признаков и учета релевантности информации используется программная система Glimpse. Отфильтрованная информация поступает на вход лингвистического процессора, который генерирует метатекст. Для проведения лингвистического анализа используется программная система Lingvo_Rus, которая является авторской разработкой. Лингвистический процессор при работе обращается к базе данных словаря русского языка. Для генерации баз данных словаря используется ht://Dig версии 3.1.1 и русский словарь Лебедева версии 0.99b3. В качестве СУБД базы данных словаря русского языка выбран Oracle 8. Метатекст выступает в качестве исходной информации для процесса построения базы знаний. Для классификации объектов по модифицированному алгоритму C4.5, выбрана программная система Darwin 3.1.1. Классификация объектов осуществляется на основе иерархии классов, для создания которой используется авторская разработка – программная система Builder Class. Для определения отношений между вершинами НСГС используется авторская разработка – программная система Links, которая из метатекста и классификации объектов строит базу знаний в виде нечеткой семантической гиперсети. Информационные схемы Построения базы знаний и Извлечения ассоциаций из базы знаний представлены на рис.4 и рис.5 соответственно. Для извлечения ассоциаций из базы знаний используется авторская разработка – программная система Miner.

Постановка аналитической задачи:

1) определить связи Владимира Лисина с металлургическими компаниями, банками и политиками;

2) определить связи физических и юридических лиц в окружении Лисина с и Искандером Махмудовым.

Определим параметры запроса к нечеткой семантической гиперсети:

$I = \left\langle x_a^* = Лисин, q^* = 2, \Gamma_1(x_a)^* = Металлургия, Банк, Политик, \Gamma_2^2(x_a)^* = Искандер\ Махмудов, A^* = A_2, 0.1, =\geq \right\rangle$.

Полученная таким образом НСГС (рис.6) непосредственно используется для решения поставленной аналитической задачи. После изучения структуры НСГС аналитик обращается к естественным текстам, в которых встречается интересующая его ассоциация.

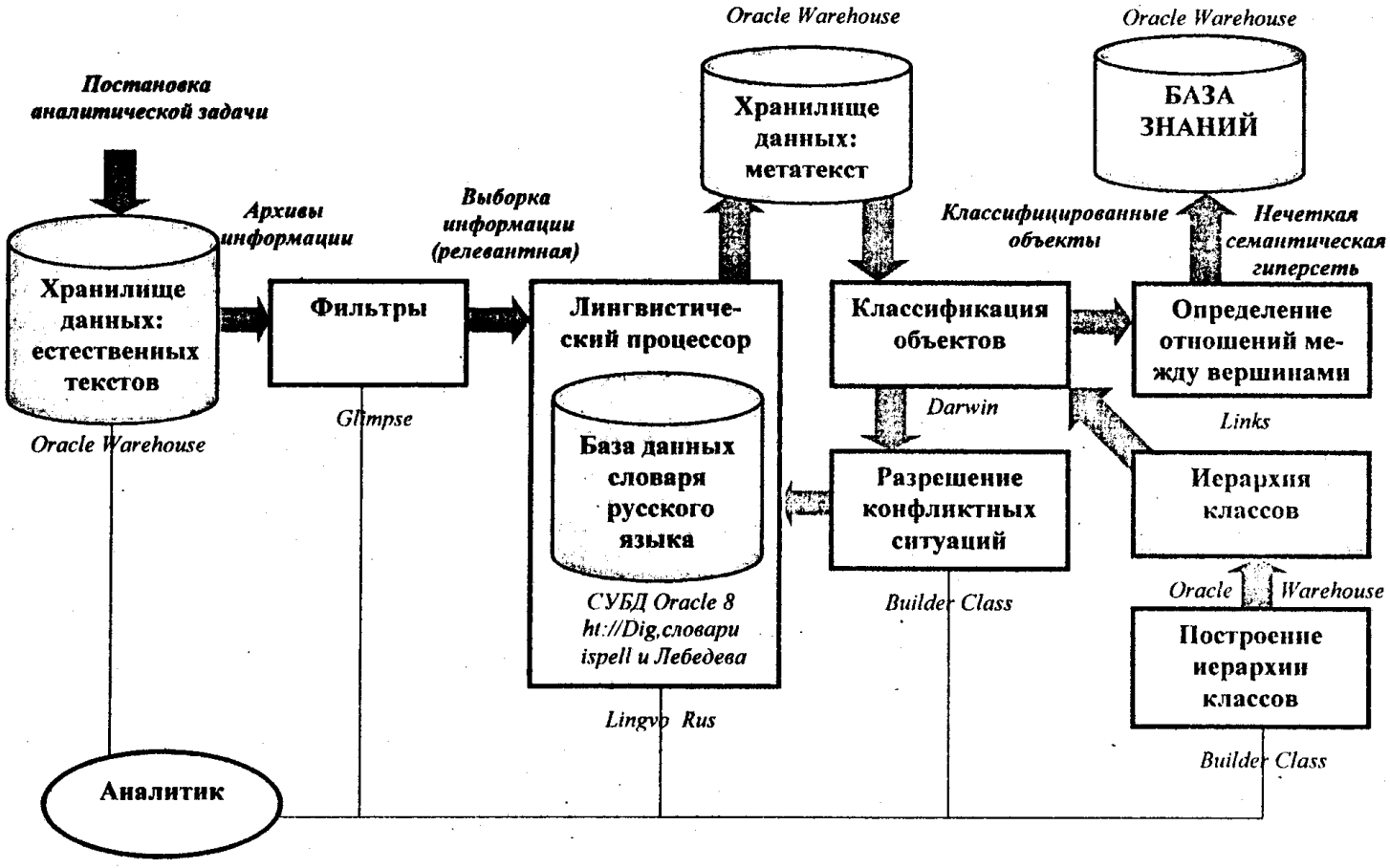

Рис.4. Информационная схема построения базы знаний

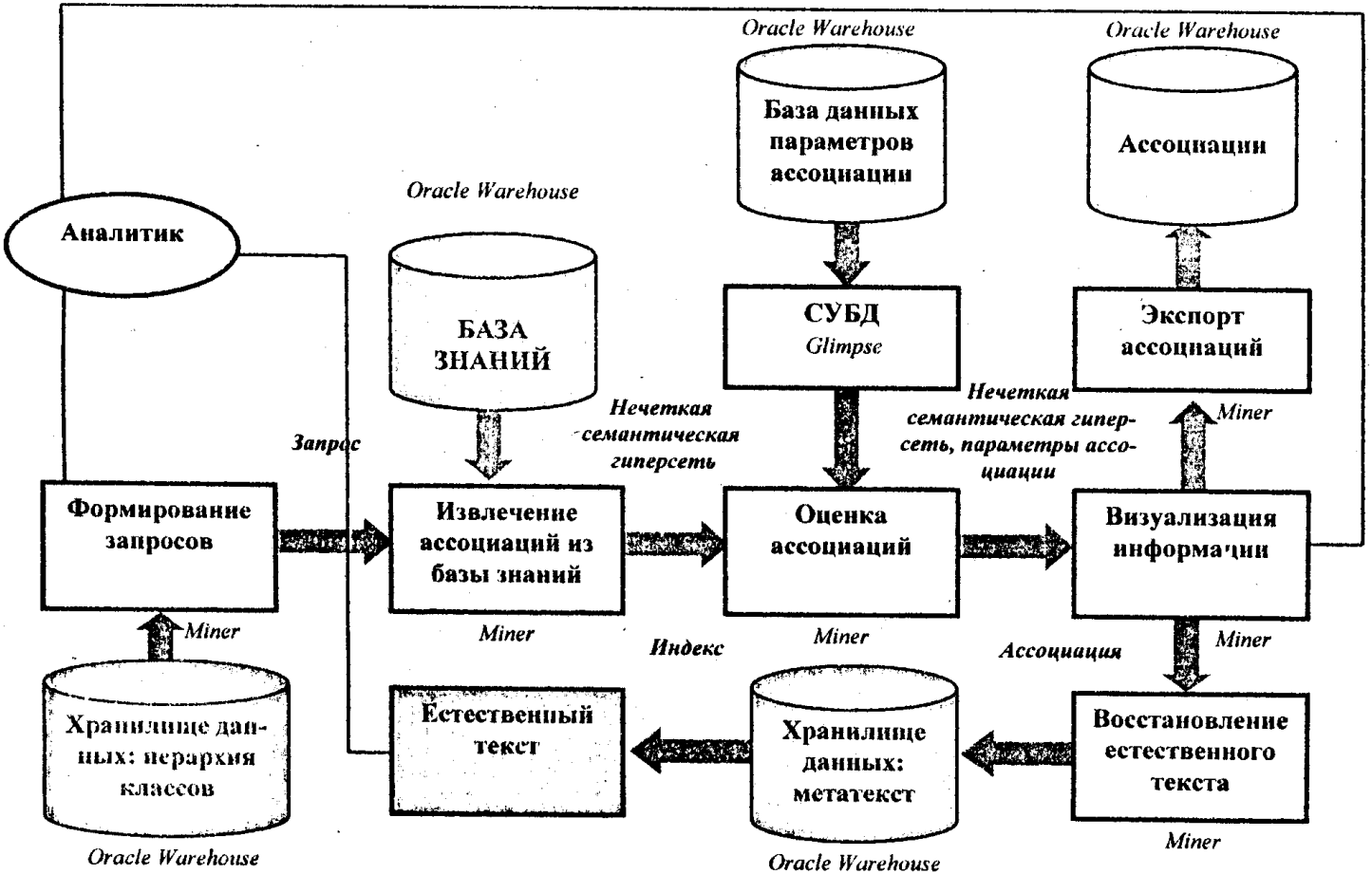

Рис.5. Информационная схема извлечения ассоциаций

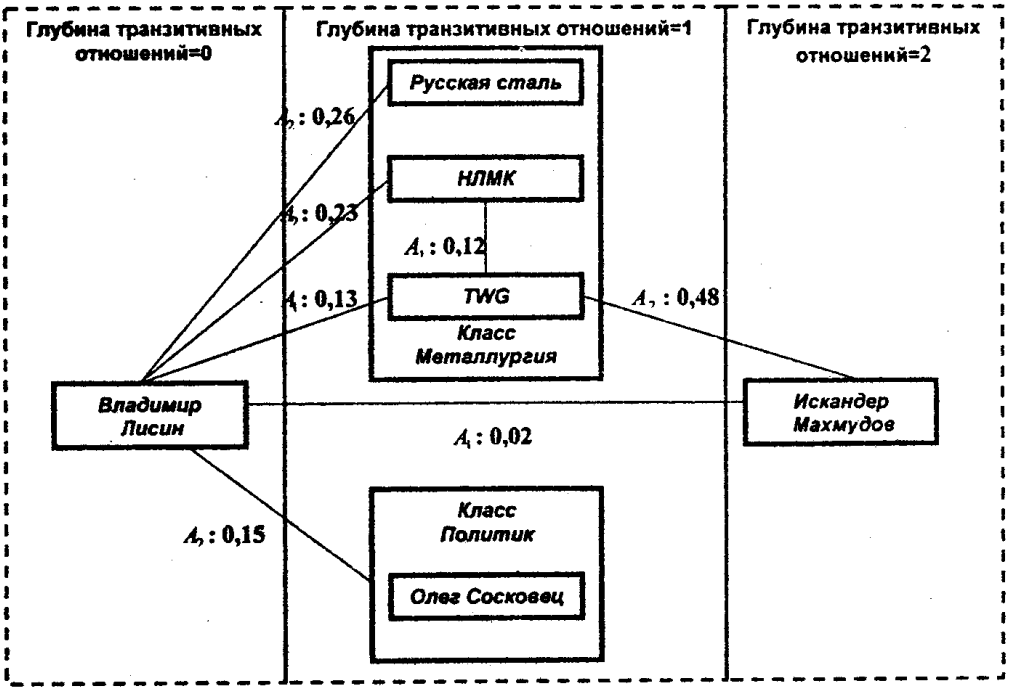

Рис.6. Нечеткая семантическая гиперсеть – Черная металлургия

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ

Проведенные в диссертационной работе исследования позволили сформулировать следующие основные выводы и получить практические результаты:

1. Анализ современных концепций, методик, технологий и систем, связанных с интеллектуальным анализом естественных текстов, позволил сделать вывод о совершенно недостаточном уровне развития существующих моделей знаний, методов анализа информации, средств представления информации и соответствующего информационно-аналитического обеспечения. Особенно остро эти недостатки проявляются в области интеллектуального анализа естественных текстов на русском языке.
2. В диссертационной работе предлагается модель знаний в виде нечеткой семантической гиперсети, которая является основой для построения методов представления и анализа структурной информации с точки зрения иерархии классов и типов отношений. Целесообразность использования теории нечетких множеств определяется трудной формализацией исходной информации и свойствами мышления человека. Модель знаний в виде нечеткой семантической гиперсети позволяет перейти на более высокий уровень представления информации (естественный для мышления человека), который связан с введением качественных категорий отношений между объектами, определяемых множеством лингвистических переменных.

3. Разработана модель метатекста русского языка на основе теории нечетких множеств, которая позволяет формализовать процесс обработки естественного текста за счет отказа от глубинного семантического анализа естественного текста.
4. На основе проведенного анализа методов Data Mining обоснована целесообразность использования деревьев решений для классификации объектов в базе знаний. Показано, что для повышения производительности деревьев решений их необходимо модифицировать введением несмещенного критерия разбиения и метода кросс-проверочного отсечения по минимальной цене.
5. Разработаны процедуры построения базы знаний основе метатекста. Переход с уровня метатекста на уровень базы знаний осуществляется на основе функции семантики, которая определяется на основе: классов объектов в отношении, типа отношения, частоты появления отношения и лингвистической характеристики отношения. Такой подход в определении функции семантики позволяет адекватно представлять силу отношения между вершинами НСГС, потому что учитывает влияние на силу отношения предметной области анализа и неравномерность распределения частоты появления объектов и отношений в естественных текстах.
6. Разработаны методы извлечения ассоциаций из базы знаний, которые позволяет существенно сократить объем анализируемой информации, как на основе формальных методов ("глубина транзитивных отношений" и "сила отношений"), так и на основе предметно-ориентированных методов ("принадлежность объекта к классу" и "тип отношения"). Полученный в результате применения методов извлечения ассоциаций подграф НСГС легко поддается анализу, так как отношения между вершинами НСГС наглядно представимы в виде матрицы отношений или в виде графа.
7. Практическим результатом исследования является разработка ИАС "Analyst Wizard" на основе нечеткой семантической гиперсети для анализа естественных текстов на русском языке, которая позволяет:

- существенно увеличить эффективность работы информационно-аналитических служб за счет автоматизации времяемких процессов аналитической обработки информации;
- извлекать ассоциации, которые ранее были неизвестны аналитику и носят не универсальный характер, из информации в виде естественных текстов;
- организовывать предметный поиск по сверхбольшим архивам информации;
- проводить глубинное аналитическое исследование отношений между объектами в информации;
- обеспечивать оперативную аналитическую обработку информации.

По теме диссертации опубликованы следующие работы

1. Вохминцев А.В. Технология Data Mining. Особенности реализации технологии Data Mining для поиска информации в естественных текстах //Научные труды "Информационно-аналитические компьютерные системы и технологии в региональном и муниципальном управлении".–Челябинск: ЮУрГУ, ЦНТИ, РАЕН, 2001.–С.72–77.

2. Вохминцев А.В. Data Mining и анализ естественных текстов: реальность или перспектива, искусственный интеллект или многомерный анализ? Организация ассоциативного поиска по наборам документов на естественном языке //Интеллектика. Логистика. Системология: Сб. науч. тр.–Челябинск: ЧНЦ РАЕН, РУО МАИ, ЧРО МАНПО, 2001.– Вып.4–5.–С.22–25.

3. Вохминцев А.В., Мельников А.В. Методы извлечения закономерностей из объектно-ориентированной модели знаний (ООМЗ). Процедуры построения семантической сети ООМЗ //Интеллектика. Логистика. Системология: Сб. науч. тр.–Челябинск: ЧНЦ РАЕН, РУО МАИ, ЧРО МАНПО, 2001.– Вып.6.– С.30–34.

4. Вохминцев А.В. Модель естественного текста на русском языке //Интеллектика. Логистика. Системология: Сб. науч. тр.–Челябинск: ЧНЦ РАЕН, РУО МАИ, ЧРО МАНПО, 2001.– Вып.6.– С.35–38.

5. Вохминцев А.В., Мельников А.В. Методика извлечения структурных знаний из естественных текстов //Известия Челябинского научного центра.–Челябинск: ЧНЦ Уро РАН, РФЯЦ – ВНИИТФ, ЮУрГУ, 2002.– Вып.15.–С.10–15.

6. Вохминцев А.В., Мельников А.В. Модель знаний на основе нечетких семантических гиперсетей для представления отношений между объектами в естественном тексте //Интеллектика. Логистика. Системология: Сб. науч. тр.–Челябинск: ЧНЦ РАЕН, РУО МАИ, ЧРО МАНПО, 2002.– Вып.7.–С.21–33.

7. Вохминцев А.В., Мельников А.В. Архитектура информационно-аналитической системы (ИАС) на основе нечетких семантических гиперсетей //Системная интеграция в управленческой деятельности: Сб. науч. тр.–Екатеринбург: УГТУ-УПИ, 2002–С.26–30.

8. Мельников А.В., Вохминцев А.В. Новые технологии text mining: извлечение ассоциаций из естественных текстов на основе нечетких семантических гиперсетей //Третья Международная конференция "Актуальные проблемы современной науки": Тезисы докладов. – Самара: СГТУ, 2002–С.26.

9. Вохминцев А.В., Мельников А.В. Применение объектно-ориентированной модели знаний для представления знаний из естественных текстов //Научные труды "Компьютерные системы поддержки принятия решения руководителей".–Челябинск: ЮУрГУ, 2002–С.33–42.