На правах рукописи

Тарасов Елизар Савович

РАЗРАБОТКА ЛИНГВОСЕМАНТИЧЕСКИХ МЕТОДОВ ОБРАБОТКИ ЭКСПЕРТНОЙ ИНФОРМАЦИИ ДЛЯ СИТУАЦИОННЫХ ЦЕНТРОВ ОРГАНОВ ГОСУДАРСТВЕННОЙ ВЛАСТИ

Специальность 05.13.01 – «Системный анализ, управление и обработка информации (информационные и технические системы)»

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Краснодар – 2011

Работа выполнена в ГОУ ВПО «Кубанский государственный технологический университет»

Научный руководитель:	доктор технических наук, профессор Симаков Владимир Сергеевич
Официальные оппоненты:	доктор технических наук, профессор Ключко Владимир Игнатьевич
	кандидат технических наук, Мягкий Алексей Евгеньевич
Ведущая организация:	ГОУ ВПО «Кубанский государственный университет», г. Краснодар

Защита диссертации состоится «2» марта 2011 г. в 12.00 часов на заседании диссертационного совета Д 212.100.04 в ГОУ ВПО «Кубанский государственный технологический университет» по адресу: 350072, г. Краснодар, ул. Московская, 2, Г-251

С диссертацией можно ознакомиться в библиотеке Кубанского государственного технологического университета по адресу: 350072, г. Краснодар, ул. Московская, 2А

Автореферат разослан «31» января 2011 г.

Ученый секретарь
диссертационного совета Д 212.100.04
канд. техн. наук, доцент Власенко А.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Современный этап развития государства в условиях высокой динамики экономической ситуации и правовой базы требуют от руководителей органов государственной власти (ОГВ) постоянного внимания к различным аспектам рассматриваемых проблем в ходе принятия управленческих решений. Особую важность в данных условиях играет возможность эффективной обработки информации и принятия обоснованных решений в условиях нечеткости, неопределенности, неполноты и противоречивости исходных данных либо условий окружающей среды, специфики проблемной области.

В этой ситуации аппарат руководителя ОГВ нуждается не только в традиционных системах сбора и обработки информации, но и в аналитических моделях, позволяющих оперативно оценить реальное состояние проблемной области, предусмотреть тенденции развития и проанализировать возможные последствия управленческих решений. Этот комплекс задач позволяют решить ситуационные центры (СЦ), которые представляют собой автоматизированный информационно-аналитический комплекс для принятия стратегических решений и управления всеми аспектами деятельности ОГВ.

В современных системах обработки информации и получения знаний в условиях нечеткости, неполноты или противоречивости исходной информации о рассматриваемой проблеме, преимущественно естественно-языковых (ЕЯ) форм ее представления, а также связи со многими предметными областями, актуальным становятся ряд вопросов, связанных с привлечением к процедуре групповой экспертной оценки квалифицированных специалистов в различных предметных областях и обработкой получаемой экспертной информации.

Однако недостаточная эффективность существующих методов информационно-аналитического обеспечения в СЦ обуславливает необходимость дальнейшей разработки методологии и прикладных алгоритмов системного подхода, практической реализации процедур получения знаний и обработки разнородной информации, что повысит адекватность и обоснованность принимаемых в ОГВ решений по задачам оперативного, стратегического и ситуационного управления.

В рамках решения этих задач особый интерес представляет круг вопросов, связанный с формализацией естественно-языковых описаний проблем в рамках интересующей предметной области исследования, их последующего анализа и моделирования, а также дальнейшего использования в процедурах организации и проведения экспертизы, анализа и обобщения получаемой информации. Необходимо разработать методики и алгоритмы применения набора формальных и неформальных подходов к анализу ЕЯ- описания проблемы, ее формализации, оценке и впоследствии – к подбору специалистов в состав экспертных групп, обработке и обобщению поступающей информации по разработанным методикам.

Выбор лингвосемантического подхода в качестве платформы для разрабатываемых методик и алгоритмов обусловлен его эффективностью в обработке ЕЯ-описаний, возможностью интеграции с другими методами

получения и аналитической обработки знаний, гибким математическим и алгоритмическим аппаратами.

Целью работы является разработка методического аппарата лингвосемантического анализа и оценки экспертной информации, подходов к его применению в контуре принятия решений ситуационных центров органов государственной власти.

Объектом исследования является комплекс информационно-аналитических систем в составе ситуационных центров ОГВ.

Предмет исследования – математическое, алгоритмическое и программное обеспечение процедур лингвосемантического анализа естественно-языковых описания проблемы и экспертной информации, система соответствующих подходов, методов и моделей.

Основными задачами исследования являются следующие:

1. Разработка подходов, методик и алгоритмов лингвосемантического анализа и формализации информации, представленной на естественном языке с учетом факторов ее неопределенности, неполноты и противоречивости;
2. Разработка методик и алгоритмов формирования тезаурусных описаний экспертной информации;
3. Разработка методик построения моделирующих семантических сетей для формального представления ЕЯ-описаний и экспертной информации;
4. Разработка методики формирования проблемно-ориентированных экспертных групп в СЦ ОГВ, анализа, обобщения и формализации результатов экспертизы;
5. Программная реализация модуля с использованием архитектуры клиент-сервер и технологий интеллектуального анализа данных с поддержкой распределённых режимов работы комплекса.
6. Оценка эффективности разработанных методик, алгоритмов и программного комплекса.

Методы исследования включают: методы семантического, синтаксического, лингвистического и морфологического анализа, теории семантических сетей, кластерного анализа, теории графов, интеллектуального анализа данных (Data Mining).

Положения, выносимые на защиту.

К основным научным результатам, изложенным в диссертационной работе и выносимым на защиту, относятся:

- подходы, методики и алгоритмы лингвосемантического анализа естественно-языковых описаний проблемы и получаемой экспертной информации в контуре принятия решений СЦ ОГВ;
- методика практической реализации математических моделей и алгоритмов процедур морфологического, синтаксического и лингвосемантического анализа, построения моделирующих семантических сетей обработки экспертной информации;
- программный комплекс «Эксперт», реализующий разработанные методики, модели и алгоритмы, интегрированный в структуру СЦ и

обеспечивающий автоматизацию процедур организации и проведения групповых экспертиз;

- клиент-серверная архитектура программного комплекса, механизмы его интеграции в СЦ ОГВ, подходы и результаты оценки его эффективности, подтверждающие адекватность полученных в работе результатов.

Научная новизна работы:

- усовершенствованные математические модели и алгоритмы лингвосемантического анализа, формализации и обобщения естественно-языковых описаний проблемы и экспертной информации;
- оптимизация методик и алгоритмов формирования тезаурусных описаний, определения мер семантической близости моделей ЕЯ-информации, их кластеризации и ранжирования;
- подходы к практическому использованию разработанных методик в ситуационных центрах органов государственной власти;
- архитектура программного комплекса «Эксперт», методика его интеграции в состав ситуационных центров органов государственной власти; модель информационного взаимодействия участников контура принятия решений.
- оценка эффективности разработанных методик на примере формирования проблемно-ориентированных экспертных групп, анализа и обобщения получаемой экспертной информации;

Обоснованность и достоверность научных положений, основных выводов и результатов диссертации обеспечивается тщательным анализом состояния результатов российских и зарубежных исследований в областях теории прикладной лингвистики и семантического анализа, организации и проведения экспертиз, проектирования и реализации ситуационных центров.

Практическая значимость. Разработана совокупность теоретических положений и реализован специализированный программный комплекс, позволяющий осуществлять лингвосемантический анализ, формализацию и построение тезаурусов экспертной информации, представленной в естественно-языковой форме с учетом специфики решаемых задач, ограничений и условий внешней среды, сформирована методика его интеграции и использования в составе ситуационных центров органов государственной власти.

Усовершенствование научно-методического аппарата информационно-аналитического обеспечения и частичной автоматизации процедур экспертного принятия решений в СЦ ОГВ дает возможность повысить функциональность и оперативность процедур управления.

Публикация результатов и апробация работы. По результатам диссертации опубликовано 10 печатных работ, из них 5 статей (2 статьи в издании из Перечня ВАК для публикации научных результатов диссертаций на соискание ученой степени доктора и кандидата наук), 6 тезисов докладов в материалах Международных, Всероссийских и внутривузовских конференций, а также 1 свидетельство о государственной регистрации программы для ЭВМ. Восемь работ выполнены в соавторстве; личный вклад соавтора (научного руководителя) заключался в постановке задач и общем руководстве.

Основные результаты работы обсуждались на следующих Международных, Всероссийских и внутривузовских конференциях: международная научно-практическая конференция «Информационная безопасность» (Таганрог, 2005); международная заочная научно-практическая конференция «Прогрессивные технологии развития» (Томск, 2008); конференция получателей грантов регионального конкурса «ЮГ» Российского фонда фундаментальных исследований» (Краснодар, 2008); всероссийская конференция с элементами научной школы для молодёжи «Проведение научных исследований в области обработки, хранения, передачи и защиты информации» (Ульяновск, 2009); научно-практическая конференция «Научно-техническое творчество молодежи – путь к обществу, основанному на знаниях» (Москва, 2009); научно-практическая конференция «Ситуационные центры 2009» (Москва, РАГС 2010); международная научно-практическая конференция «Молодёжь и наука: реальность и будущее» (2010 г, Невинномысск).

Реализация и внедрение результатов работы.

Проведение исследований, отражённых в диссертации, было поддержано в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009-2013 гг., ГК № П742 «Разработка теоретических основ и построение интеллектуальной информационно-аналитической системы как основы региональных ситуационных центров органов государственной власти».

Часть результатов была использована при выполнении работ по ГК № П2026 "Разработка подходов к анализу и практической реализации интеллектуальных информационно-аналитических систем органов власти на основе ситуационного моделирования"; ГК №П2378 «Разработка теоретических основ и построение интеллектуальной информационно-аналитической системы как платформы поддержки принятия решений в органах государственной власти"; проекта РФФИ № 08-07-99030, «Разработка теоретических основ и построение интеллектуальных систем мониторинга, анализа и поддержки принятия политических, социально-экономических и технологических решений регионального уровня для ситуационных центров органов власти».

Объем и структура работы. Диссертация включает в себя введение, 5 глав, заключение, список используемых источников из 108 наименований. Работа изложена на 198 страницах, содержит 42 рисунка и 12 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, сформулированы цели и задачи диссертационного исследования, основные положения, выносимые на защиту, определена научная новизна и практическая значимость, содержание и методы выполнения работы, кратко изложены основные результаты.

В первой главе «Аналитический обзор проблем информационно-аналитического обеспечения деятельности ситуационных центров органов государственной власти» приведено краткое описание состояния и результатов российских и зарубежных исследований в области проектирования и реализации ситуационных центров, роли и места информационно-аналитического обеспечения в контуре принятия решений, а также проблем автоматизированной обработки естественно-языковой информации,

существующих методов и подходов лингвистического, семантического и других направлений анализа, теории прикладной лингвистики и семантики.

Обоснована необходимость и актуальность разработки методик повышения эффективности подбора экспертов с учетом специфики решаемых проблем, целесообразность частичной автоматизации этих процессов на основе методов обработки естественно-языковых (ЕЯ) описаний, предложены обобщенные структурно-функциональные и информационные модели контуров взаимодействия участников организации и проведения экспертиз.

Показано, что процедура экспертного оценивания обладает рядом специфических черт: слабая формализуемость, противоречивость, значительная нечеткость, неполнота, неопределенность исходных данных и получаемых рекомендаций, необходимость их обобщения, согласования с учетом как требований регламента, так и специфики решаемой проблемы. Указанные особенности налагают ряд ограничений и требований на подходы и методики автоматизации процедур экспертизы, обуславливая необходимость использования методов системного анализа, нечеткой логики и обработки, формализации ЕЯ-описаний объектов. В этой связи предложено использование лингвосемантических подходов к обработке информации, обоснована их адекватность и эффективность для решения сформулированных задач.

Взаимосвязь решаемых задач в ходе функционировании СЦ на примере контура экспертного оценивания приведена на рис. 1.

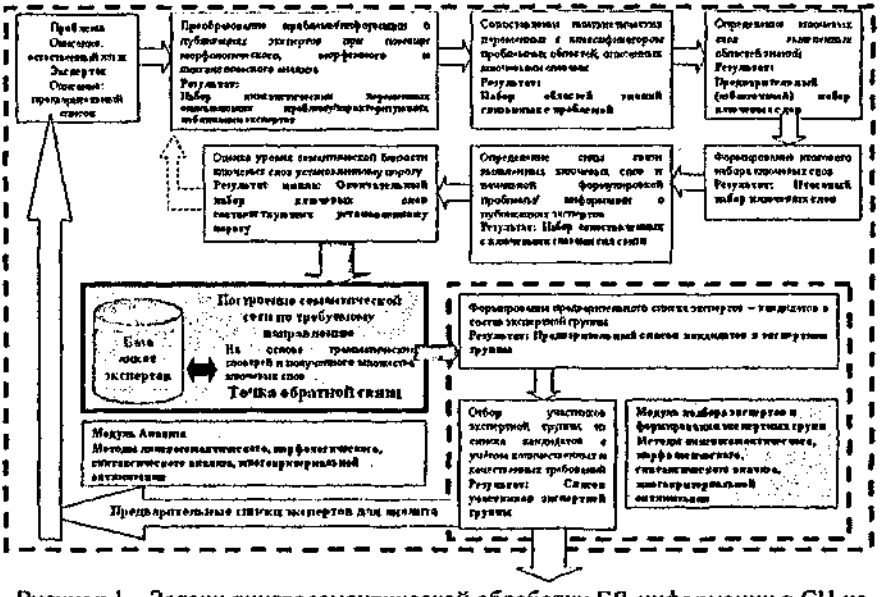


Рисунок 1 – Задачи лингвосемантической обработки ЕЯ-информации в СЦ на примере контура экспертного оценивания

Проведен анализ методов, существующего математического аппарата и алгоритмического обеспечения лингвосемантического подхода к анализу и формализации ЕЯ-объектов. Формально поставлен ряд задач:

- лингвосемантический анализ разнородной информации, представленной на естественном языке, построение моделирующих семантических сетей;
- выделение ключевых слов, словосочетаний и семантических ареалов из полученных описаний (модели «Онтология - Тезаурус»);
- определение мер семантический близости, ранжирование, кластеризация модельных и ЕЯ-описаний (модели «Semantic Similarity/Clustering»);
- формирование итоговых обобщений имеющихся описаний и получаемой экспертной информации;
- разработка «обобщенного» лингвосемантического алгоритма анализа, формализации и обработки ЕЯ-информации с учетом факторов неполноты, нечеткости и противоречивости;

Во второй главе «Разработка методов, моделей и алгоритмов лингвосемантического анализа и обработки естественно-языковой информации» исследованы теоретические аспекты обработки знаний в ИАС; особенности архитектуры и функционирования ситуационных центров ОГВ; разработаны модели потоков данных, исследованы особенности представления и использования естественно-языковой информации в рамках информационно-аналитического обеспечения деятельности СЦ, которая рассмотрена как объект моделирования, управления и автоматизации. В результате определен ряд существенных недостатков традиционно используемых подходов, сформулированы предложения по частичной автоматизации и повышению его эффективности на базе лингвосемантического подхода к анализу и обработке информации об объектах управления и окружающей среде.

На этапе предварительной обработки и предметной классификации будем рассматривать текст как «набор слов», используя численные характеристики употребления тех или иных терминов, вне зависимости от порядка их употребления. Тогда вероятность того, что термин w, принадлежащий формируемому тезаурусу W, встречается в описании проблемы или корпусе анкет экспертов d (множества D тематического классификатор), т.е. принадлежит той или иной предметной области t:

$$P(w|d) = \sum_{t \in T} P(w|t)P(t|d), \qquad (1)$$

где t – элемент множества T предметных областей.

Для оценки максимального правдоподобия параметров модели, зависящей от скрытых переменных, используем EM-алгоритм. Параметры предварительного семантического анализа $P(w|t)$ и $P(t|d)$ определим следующим образом. Пусть r – число итераций. На E-шаге вычислим $P(t \mid w,d)^{(r)}$:

$$P(t \mid w,d)^{(r)} = \frac{P(w|t)^{(r-1)}P(t|d)^{(r-1)}}{\sum_{t' \in T} P(w|t')^{(r-1)}P(t'|d)^{(r-1)}} \qquad (2)$$

На М-шаге оценим параметры:

$$P(w|t)^{(r)} = \frac{\sum_{d \in D} N(w,d) P(t|w,d)^{(r)}}{\sum_{w' \in W} \sum_{d \in D} N(w',d) P(t|w',d)^{(r)}} \quad (3), \qquad P(t|d)^{(r)} = \frac{\sum_{w \in W} N(w,d) P(t|w,d)^{(r)}}{\sum_{t' \in T} \sum_{w \in W} N(w,d) P(t'|w,d)^{(r)}} \quad (4)$$

где $N(w,d)$ – число вхождения элемента тезауруса w в рассматриваемый текст d. Процесс обучения повторяется до сходимости параметров. Однако параметры часто попадают в область локального оптимума, эффективность не улучшается в результате обучения. Введен параметр $0<\beta\leq 1$ для управления скоростью обучения. Выражение для М-шага примет вид:

$$P(t|w,d)^{(r)} = \frac{\left(P(w|t)^{(r-1)} P(t|d)^{(r-1)}\right)^{\beta}}{\sum_{t' \in T} \left(P(w|t')^{(r-1)} P(t'|d)^{(r-1)}\right)^{\beta}} \quad (5)$$

Для достижения глобального оптимума изначально принимаем β=1 с уменьшением умножением на $0<\eta<1$, пока оценки не улучшатся.

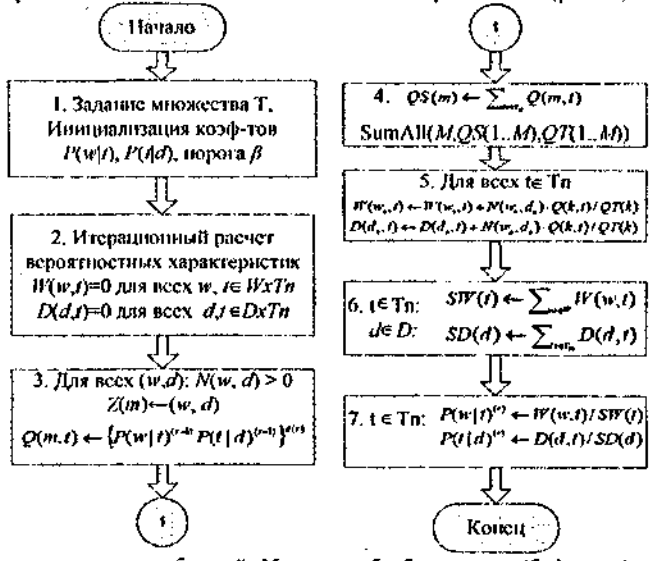
Определим суммарные вероятности $W(w,t)$ и $D(d,t)$ следующим образом:

$$W(w,t)^{(r)} = \sum_{d \in D} N(w,d) P(t|w,d)^{(r)} \quad (6) \qquad D(d,t)^{(r)} = \sum_{w \in W} N(w,d) P(t|w,d)^{(r)} \quad (7)$$

По формуле (5) получим:

$$W(w,t)^{(r)} = \sum_{d \in D} \frac{N(w,d)\left(P(w|t)^{(r-1)} P(t|d)^{(r-1)}\right)^{\beta}}{\sum_{t' \in T} \left(P(w|t')^{(r-1)} P(t'|d)^{(r-1)}\right)^{\beta}} \quad (8) \qquad D(d,t)^{(r)} = \sum_{w \in W} \frac{N(w,d)\left(P(w|t)^{(r-1)} P(t|d)^{(r-1)}\right)^{\beta}}{\sum_{t' \in T} \left(P(w|t')^{(r-1)} P(t'|d)^{(r-1)}\right)^{\beta}} \quad (9)$$

Алгоритм лингвосемантического анализа примет вид (рис. 2).



T – множество предметных областей; *M* – число обрабатываемых (буферных) документов;
Z – массив размера M с парами *(w, d)* «номер термина – номер документа»;
Q(m,t) – массив для m-х промежуточных значений рассматриваемой t-области
SumAll(m,QS,QT) – коммуникационная процедура, получает массив *QS*, передает для вычисления суммы всех значений ото всех процессов, и возвращает их в массив *QT*.

Рисунок 2 – Оптимизированный алгоритм лингвосемантического анализа с ЕМ-алгоритмом параллельного обучения

Для формирования ребер семантической сети и оценки меры семантической близости выделенных понятий (элементов тезауруса) в настоящее время используются четыре распространенных оценки: меры Jaccard, Overlap, Dice и PMI (point-wise mutual information).

Эти метрики исходят из предположения, что высокие частоты совместной встречаемости терминов в тексте указывают на значительную степень ассоциации, что в свою очередь обуславливает наличие семантических связей между ними.

В зависимости от лингвистических, стилистических и иных особенностей рассматриваемого ЕЯ-описания (объем, наличие выделенных модератором ключевых слов, изначально указанной предметной области экспертизы) будем использовать следующий набор метрик:

– Нормализованное расстояние Google и его модификацию:

$$G(w_1, w_2) = \frac{\max\{A\} - \log|D|(w_1, w_2)}{\log|D| - \min\{A\}} \quad (10), \qquad G'(w_1, w_2) = e^{-2G(w_1, w_2)} \quad (11)$$

– Индекс Jaccard (*Jaccard index*) – статистическая величина, используемая для сравнения подобия и различия анализируемого набора ЕЯ-описаний:

$$J(w_1, w_2) = \frac{|K|w_1| \cap K|w_2||}{|K|w_1| + K|w_2| - K|w_1| \cap K|w_2||} \quad (12)$$

– Коэффициент Dice (*Dice's coefficient*), совместно с индексом Jaccard, определяет меру семантической близости терминов X и Y:

$$D(w_1, w_2) = 2\frac{|K|w_1| \cap K|w_2||}{|K|w_1| + K|w_2||} \quad (13)$$

– Коэффициент Overlap (*overlap coefficient*) – мера подобия, связанная с индексом Jaccard, которая вычисляет степень совпадения между двумя тезаурусами:

$$O(w_1, w_2) = \frac{|K|w_1| \cap K|w_2||}{\min(|K|w_1||, |K|w_2||)} \quad (14)$$

– Косинусный коэффициент подобия (*The Cosine similarity*) – мера подобия между двумя n-мерными векторами по углу между ними:

$$C(w_1, w_2) = \frac{|K|w_1| \cap K|w_2||}{sqrt(|K|w_1||) \times sqrt(|K|w_2||)} \quad (15)$$

– Коэффициент простого соответствия (*Simple matching coefficient*) – по числу общих терминов, без учета размеров наборов:

$$S(w_1, w_2) = |K|w_1| \cap K|w_2|| \quad (16)$$

где: мера | . | – объем набора ключевых терминов (тезауруса); K|w_i| – набор связанных с w_i других терминов, полученных из анализируемого документа.

Предложена *реляционная* модель вычисления семантической близости, использующая набор отношений $R(a;b)$, связывающих термины *a* и *b*:

$$\mathrm{sim}(a, b) = \Xi(R(a, b)) \quad (17)$$

Здесь sim(*a; b*) – семантическая близость между терминами a и b,

Ξ – весовая функция, определенная над множеством семантических отношений $R(a; b)$, выражающая силу семантической связи между a и b.

В работе предложено использование результатов работы модуля лексического анализа – автоматически извлекаемые лексические образцы (lexical patterns), которые позволяют успешно представлять различные типы семантических отношений между терминами (порождение, наследование).

Следуя этому подходу, отношение $R(a; b)$ представляется набором лексических образцов. Обозначим частоту встречаемости лексического образца для пары $(a; b)$ как $f(r; a; b)$. Наиболее простой подход к определению Ξ, заключающийся в использовании линейно-взвешенной комбинации отношений:

$$\Xi(R(a,b)) = \sum_{r_i \in R(a,b)} w_i \times f(r_i, a, b), \qquad (18)$$

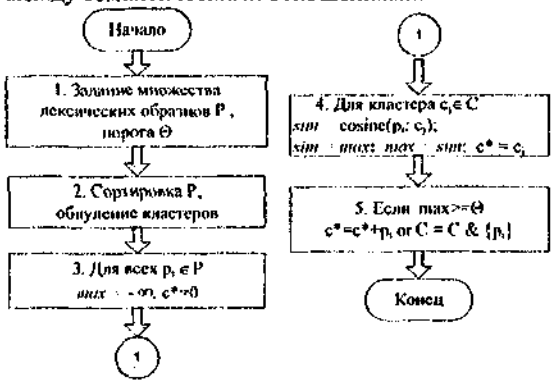
где w_i – вес, связанный с r_i и определяемый с использованием обучающей выборки (по описанному выше EM-алгоритму) – обладает рядом недостатков: ростом числа параметров при повышении сложности модели, предположением о взаимной независимости параметров линейной модели, что не соответствует природе ЕЯ-описаний.

Для преодоления указанных ограничений разработан алгоритм кластеризации лексических образцов для определения семантически связанных терминов (рис. 3). Используя результаты кластеризации, определим Ξ следующим образом:

$$\Xi(R(a,b)) = \mathbf{x}_{ab}^{t} \Lambda \mathbf{x}_{ab} \qquad (19)$$

здесь x_{ab} – вектор, описывающий термины a и b. j-й элемент x_{ab} равен сумме частот всех образов кластера c_j, т.е. $\sum_{r \in C_j} f(r, a, b))$,

Λ – межкластерная корреляционная матрица, (i,j)-й элемент матрицы описывает корреляцию между кластерами c_i и c_j; вводится для учета зависимостей между семантическими отношениями.



P – вектор частот пар (a, b), $f(a_i; b_i; p)$, в лексическом образце p; Θ – порог подобия (задается пользователем); $SORT$ – функция сортировки образцов по общей встречаемости в парах (a, b); Вычисление подобия между p_i и центроидом кластера c_j ведется по косинусному коэффициенту.

Рисунок 3 – Алгоритм кластеризации для меры семантической близости

Предложенная модель отличается от подобных ей (например, contrast model of similarity) тем, что определена над множеством семантических связей, существующих между двумя терминами, а не набором свойств каждого термина, реализуя реляционный подход к оценке семантической близости.

Для формирования итоговых обобщений имеющихся описаний и получаемой экспертной информации предложен подход, заключающийся в формировании семантических пространств (ареалов) максимальной близости на основе применения ЕА-алгоритма к результатам лингвосемантического анализа. Обозначим $\theta_1,...,\theta_k$ – формализованная модель текста с k различными предметными областями полученной семантической сети и θ_B – модель набора текстов С. Термин w в тексте d оценивается следующей величиной:

$$p_d(w) = \lambda_B p(w|\theta_B) + (1 - \lambda_B)\sum_{j=1}^{k}(\pi_{d,j}p(w|\theta_j)) \qquad (20)$$

где w – термин в тексте d, $\pi_{d,j}$ – вес текста d для выбора j-й предметной области θ_B ($\sum_{j=1}^{k}\pi_{d,j} = 1$), и λ_B – вес θ_B.

Использование модели θ_B направлено на большее разделение моделей предметных областей, т.к. θ_B присваивает высокие вероятности незначимым и неинформативным словам, снижая их влияние на модели предметных областей. θ_B оценивается на наборе текстов С и не меняется в ходе дальнейших оценок:

$$p(w|\theta_B) = \frac{\sum_{d\in C} c(w,d)}{\sum_{w\in V}\sum_{d\in C} c(w,d)} \qquad (21)$$

Введем дополнительный параметр оценки $\Lambda = \{\theta_j,\ \pi_{d,j} | d \in C,\ 1 \le j \le k\}$. Логарифмическая оценка правдоподобия C:

$$\log p(C|\Lambda) = \sum_{d\in C}\sum_{w\in V}[c(w,d) \times \log(\lambda_B p(w|\theta_B) + (1 - \lambda_B)\sum_{j=1}^{k}(\pi_{d,j}p(w|\theta_j)))] \qquad (22)$$

где $c(w; d)$ – число терминов w в тексте d.

Возникает задача найти такое значение параметра оценки Λ, которое максимизирует (22). Другими словами,

$$\begin{aligned}\hat{\Lambda} &= \arg\max_{\Lambda} \log p(C|\Lambda) \\ &= \arg\max_{\Lambda} \sum_{d\in C}\sum_{w\in V}[c(w,d) \times \log(\lambda_B p(w|\theta_B) + (1 - \lambda_B)\sum_{j=1}^{k}(\pi_{d,j}p(w|\theta_j)))]\end{aligned} \qquad (23)$$

Введем «скрытые переменные», характеризующие термины: $\{z_{d,w}\}$ и $p(z_{d,w}=B)$ – вероятность того, что термин w в тексте d подчиняется выбранному фоновому распределению (модель набора текстов θ_B). $p(z_{d,w}=j)$ означает, что термин w в тексте d встречается в контексте предметной области j, и не учитывается притом общей моделью текста (не является незначимым). Получим выражения для шагов EM-алгоритма. E-шаг:

$$p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)}{\sum_{j'=1}^{k}\pi_{d,j'}^{(n)} p^{(n)}(w|\theta_{j'})} \qquad (24)$$

$$p(z_{d,w} = B) = \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B) + (1 - \lambda_B)\sum_{j=1}^{k}\pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)} \qquad (25)$$

М-шаг:

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B))p(z_{d,w} = j)}{\sum_{j'=1}^{k} \sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B))p(z_{d,w} = j')} \quad (26)$$

$$p^{(n+1)}(w|\theta_j) = \frac{\sum_{d \in C_i} c(w,d)(1 - p(z_{d,w} = B))p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{d \in C_i} c(w',d)(1 - p(z_{d,w'} = B))p(z_{d,w'} = j)} \quad (27)$$

Зная оценочные параметры каждого термина, группы терминов (семантические ареалы), принадлежащих предметной области j условно будем считать "псевдотекстом", итоговым обобщением по j-й предметной области текста. Используя модель (27), мы агрегируем все семантические ареалы термина w, принадлежащего предметной области j (по всем текстам, рис. 4), и нормализуем выражение $\{p(w|\theta_j)\}_{w \in V}$ для достижения $\sum_{w \in V} p(w|\theta_j)=1$.

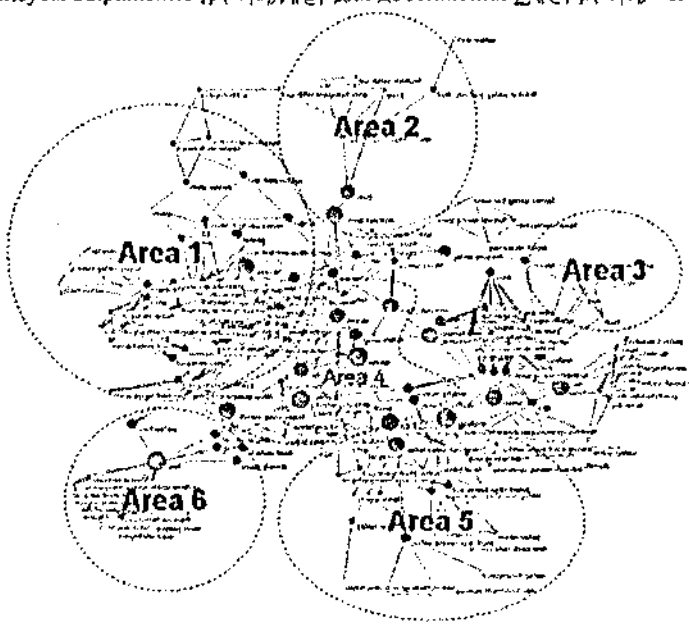


Рисунок 4 – Семантические ареалы и формирование итоговых обобщений

В третьей главе «Реализация разработанных методов и алгоритмов в составе ситуационных центров органов государственной власти» разработаны принципы и предложена методика интеграции разработанных моделей и алгоритмов в состав ситуационных центров органов государственной власти, сформированы требования к архитектуре, видам обеспечения, подходы к разработке и реализации программного модуля «Эксперт» в составе СЦ.

Показано, что создание информационно-аналитической подсистемы СЦ, реализующей разработанные методики и алгоритмы лингвосемантического анализа ЕЯ-описаний с учетом специфики предметной области в контуре принятия решений, позволит:

- уменьшить стоимость и время процедур принятия решений;
- повысить качество и эффективность принимаемых решений;

- сократить долю рутинных работ, связанных со сбором, редактированием и анализом исходных и экспертных данных;
- учесть неполноту, противоречивость и нечеткость информации о предметной области и/или о проблеме;
- обеспечить более четкое понимание поставленных целей и задач, во многом типизировать процесс;
- снизить ресурсные затраты.

Разработана модель информационного взаимодействия участников принятия управленческих решений в СЦ ОГВ (рис. 5).

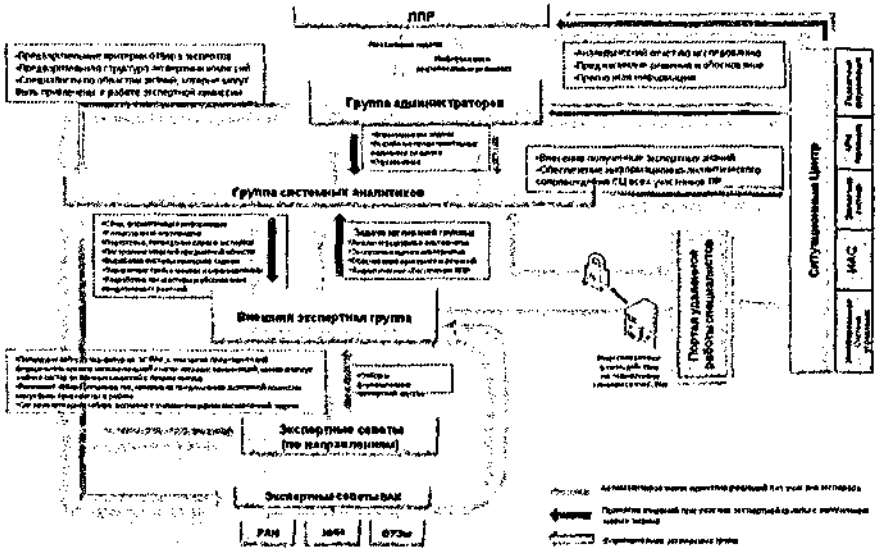


Рисунок 5 – Модель информационного взаимодействия участников принятия решений в СЦ

Описываемая ИАС в условиях нечеткости и слабой структурированности исходной информации обеспечит возможность учета экспертных знаний в дальнейшем принятии решений, частичную автоматизацию этих процессов, механизмы ситуационного управления ими. Разработана система информационных, структурно-функциональных и DFD-моделей процедур разработки ИИАС СЦ, определены требования к ее математическому и алгоритмическому обеспечению.

Для практической реализации моделей предложено выделить ряд подсистем в блоке экспертного принятия решений СЦ (рис. 6):

- Подсистема визуализации и представления данных (интерактивное представление данных, построение и функционирование когнитивных моделей, формализация результатов, интерпретация информации);
- Подсистема формирования проблемно-ориентированных экспертных групп (подбор кандидатур с учётом специфики проблемной области на

основе разработанных методик и алгоритмов анализа и формализации проблем, формализации данных об экспертах для формирования группы);

- Подсистема организации и проведения экспертиз (процессы функционирования экспертной группы в части организации и проведения экспертизы, в том числе формирования списка вопросов к обсуждению, сбор, обработку и анализ получаемых экспертных знаний с их последующей формализацией).

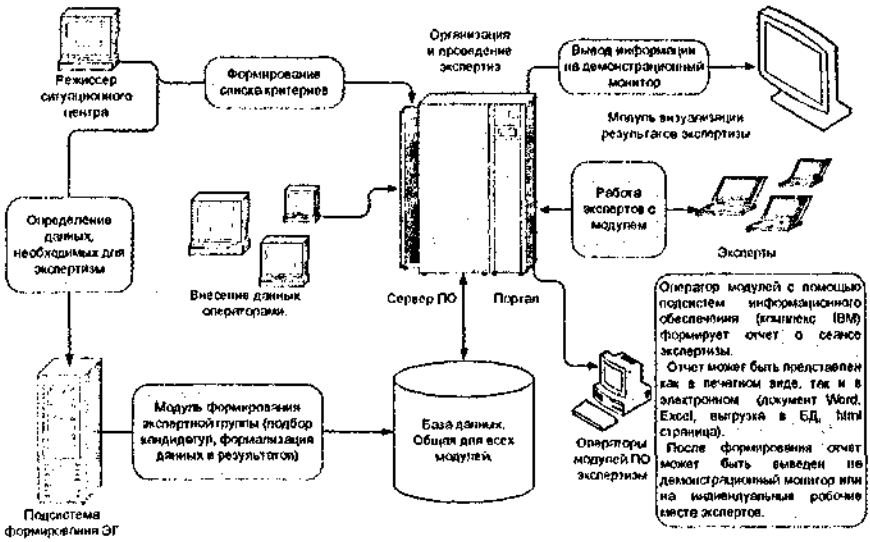


Рисунок 6 – Схема взаимодействия подсистем модуля экспертизы

В четвертой главе «Разработка программного комплекса «Эксперт» исследованы подходы к созданию программного комплекса, требования к инструментальным средствам, архитектуре, структуре и режимам работы.

Приведено описание программной реализации разработанных методик, алгоритмов и моделей информационного взаимодействия участников принятия управленческих решений, интерфейсы и регламент взаимодействия с подсистемами СЦ. В составе программного комплекса выделен лингвосемантический модуль (рис. 7), реализующий разработанные методики и алгоритмы, который использует морфологические, лексические и лингвистические словари на этапах предварительной обработки ЕЯ-объектов, в терминах которых формируется образ текста описания анкеты эксперта или проблемы.

Программное обеспечение реализует набор методов семантического анализа: лингвистическая обработка и семантическая интерпретация, выполняемые соответственно лингвистическим и семантическим модулями.

Лингвистический модуль объединяет этапы целостного [illegible] текста на естественном языке

графематического анализа выделяются текстовые единицы (слова, предложения и абзацы), выполняется исключения незначимых слов и конструкций. На этапе морфологического анализа определяются грамматические значения слов.

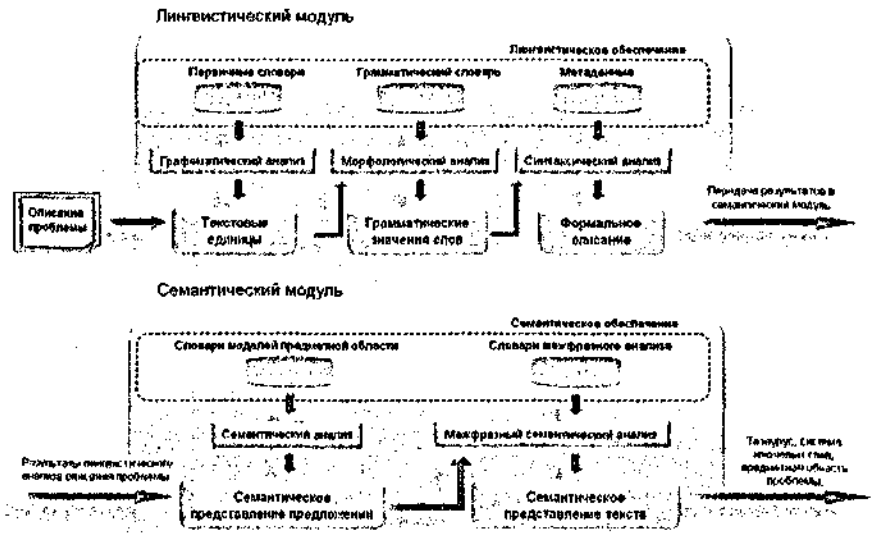
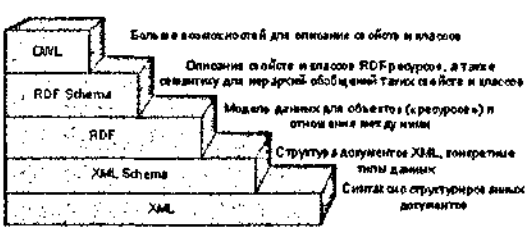


Рисунок 7 – Схема реализации в ПО методик лингвосемантического анализа

На этапе синтаксического анализа определяется синтаксическая структура предложения, описываемая формулами формального языка.

Семантический модуль выполняет смысловую обработку текста, входные данные представлены результатами обработки, полученными лингвистическим модулем. На этапе межфразового семантического анализа производится объединение семантических представлений отдельных предложений в единую семантическую сеть, описывающую смысл всего текста.

В результате лингвистического анализа поставленной проблемы производится ее структуризация в виде набора моделей проблемных областей, также формируется тезаурус и набор ключевых слов, описывающих проблему и предложения по критериям системы выбора экспертной группы (рис. 8).



семантической сети

Разработка комплекса велась с учетом требований к функциональности:

- Разработка на основе технологий Win32-приложения с учетом требований эргономики, а также программно-аппаратной совместимости;
- Работа с подсистемой формирования знаний и их формализации путем создания единой БД, ее администрированию на основе Microsoft SQL;
- Работа с единой, унифицированной формой анкет;
- Поддержка методики проблемно-ориентированного отбора и ранжирования на основе многокритериального поиска;
- Учет возможности отбора по географическому принципу – с учетом региона и поддерживаемым областям знаний.
- Обеспечение возможности формирования экспертных групп на основе ввода количественных критериев отбора и требований к группе;
- Возможности отладки, верификации этапов функционирования ИАС;
- Возможности формирования и выгрузки отчетных данных.
- Фрагмент структуры БД программного комплекса приведен на рис. 9.

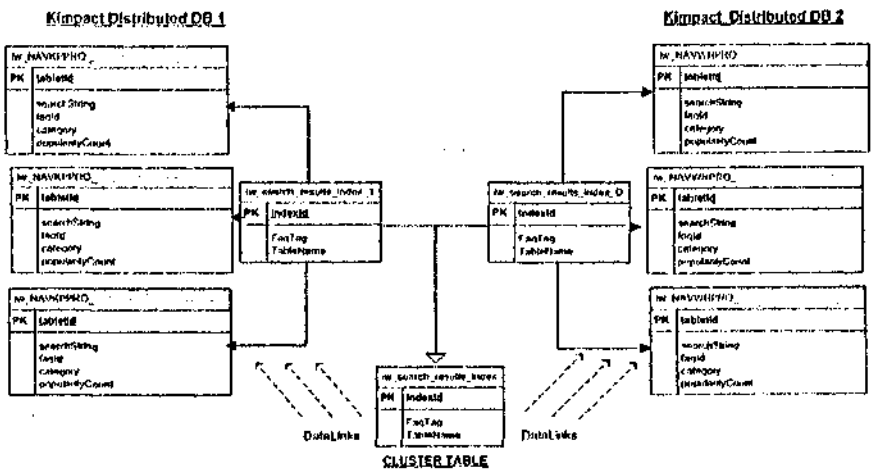


Рисунок 9 – Фрагмент БД семантической сети программного комплекса

Программный комплекс обеспечивает информационную поддержку организации и проведения экспертизы по следующим направлениям (рис. 10):

1. Работа с данными по экспертам.
2. Формулировка проблемы и ввод в систему.
3. Ввод в систему ключевых слов по проблеме
4. Отображение результатов экспертизы: графика, текст.
5. Рассылка сообщений экспертам, контроль получения ответов.
6. Контроль времени проведения экспертизы, сбор, обобщение и хранение результатов.

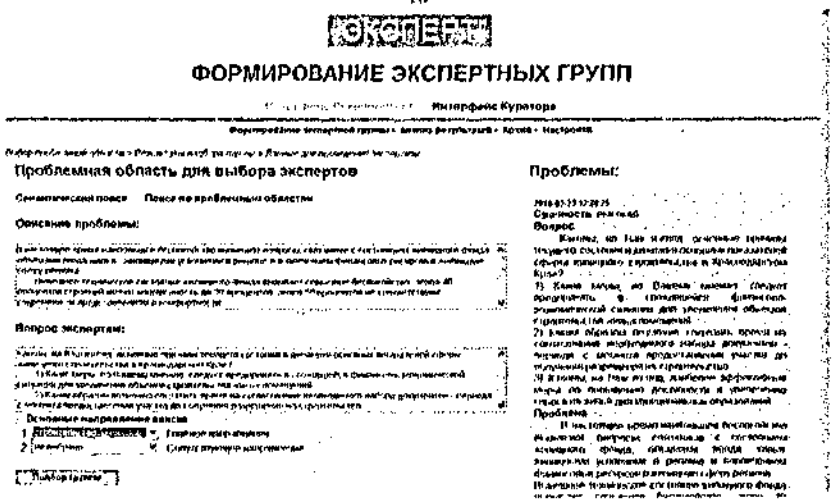


Рисунок 10 - Интерфейс модуля «Эксперт»

Следуя представленной схеме, возникает следующий контур функционирования модуля «Эксперт»:

- Ввод описания проблемы на естественном языке
- Формализация полученного описания, классификация проблемы по рубрикатору, формирование множества ключевых слов (словоформ);
- Процедуры подбора кандидатур специалистов в состав формируемой экспертной группы, ранжирование, коррекция получаемого списка;
- Рассылки участникам экспертной группы специальных сообщений с указанием адреса генерируемой анкеты-опросника для сбора мнений;
- Мониторинг хода проведения экспертизы в режиме реального времени, сбор заполненных анкет, обработка и обобщение получаемых данных;
- Визуализация результатов аналитической обработки экспертных данных

В пятой главе «Оценка эффективности использования программного комплекса» разработана методика оценки эффективности разработанных алгоритмов и программного модуля, приведено описание контрольного примера и полученных результатов (в части подбора экспертов и обобщения материалов экспертизы), сделан вывод о степени эффективности и адекватности разработанных методик и ПО.

Разработано несколько показателей численной оценки получаемых мер семантической близости и формальных представлений ЕЯ-информации в виде семантической сети: SER, ER и коэффициент Спирмена. SER (Strong Error Rate) – число ключевых слов (в процентах от общего числа найденных), которые не имеют отношения к проблемной области (определяется либо вручную – экспертом, либо автоматически – по наличию/отсутствию в тезаурусе).

$$SER(w) = \frac{\left| Words^{w}_{ISA} \setminus \left(Words^{w}_{WordNet} \cup Words^{w}_{Tes} \right) \right|}{\left| Words^{w}_{ISA} \cup Words^{w}_{WordNet} \cup Words^{w}_{Tes} \right|} \cdot 100 \qquad (28)$$

ER (Error Rate) – число слов (в процентах), не являются синонимами (определяется вручную экспертом либо автоматически – по наличию/отсутствию в тезаурусе WordNet).

$$ER(w) = \frac{\left| Words^{w}_{LSA} \setminus Words^{w}_{WordNet} \right|}{\left| Words^{w}_{LSA} \cup Words^{w}_{WordNet} \right|} \cdot 100 \qquad (29)$$

Где $Words^{w}_{LSA}$ – множество слов, найденных с помощью предложенного алгоритма для слова w, $Words^{w}_{WordNet}$ – список синонимов из тезауруса для слова w, $Words^{w}_{Ten}$ – список слов близких по значению из тезауруса.

Результаты сравнения абсолютных величин и процентного соотношения ошибок определения мер близости с помощью различных алгоритмов приведены на рис. 12.

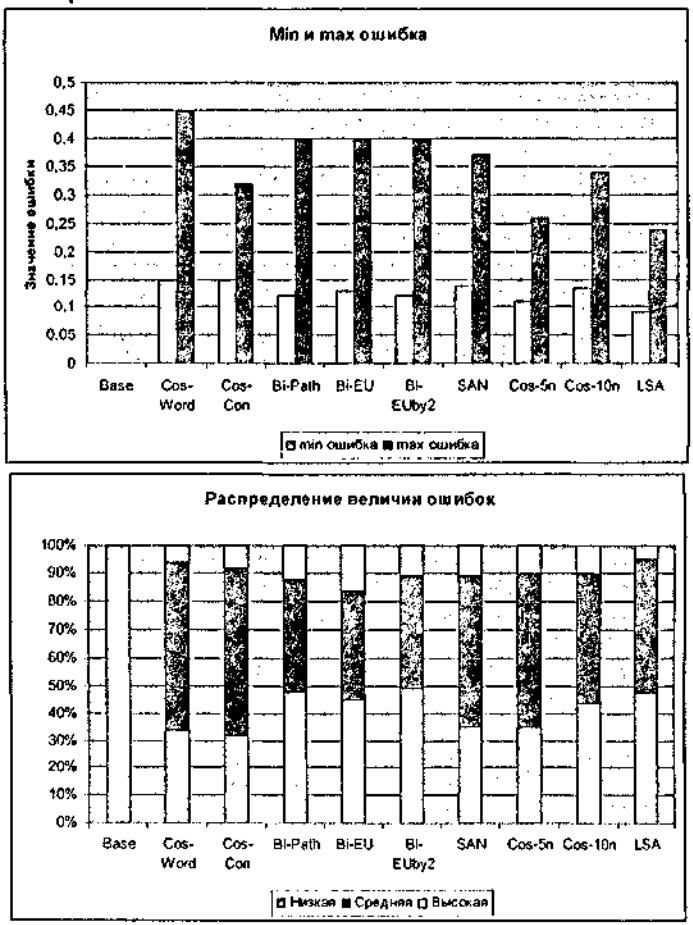


Рисунок 12 – Результаты оценки вычисления мер семантической близости различными алгоритмами

Ввиду трудности объективной оценки семантической близости для сравнительного анализа использовался набор данных Миллера-Чарльза (Miller-Charles dataset), который содержит 30 совокупностей пар слов, изначально оцененных группой экспертов от 0 (отсутствие подобия) до 4 (идентичность). Результаты вычисления степень корреляции между экспертными оценками и полученными автоматически различными алгоритмами, приведены в табл. 1.

Таблица 1 – Корреляционные коэффициенты методов на Miller-Charles set

Название и описание метода	Корреляция на Miller-Charles set
Индекс Jaccard (*Jaccard index*) (12)	0,260
Коэффициент Dice (*Dice's coefficient*) (13)	0,267
Коэффициент Overlap (*overlap coefficient*) (14)	0,382
Метод PMI (point-wise mutual information)	0,549
Нормализованное расстояние Google (11)	0,205
Метод SII (Sahami, 2006)	0,580
Мера CODC (Buckley, Salton, 1994)	0,694
Метод Chen (Chen, Lin, Wei, 2006)	0,834
Предлагаемый алгоритм ЛСА	0,867

Для оценки релевантности полученных кандидатур проведен экспертный анализ предлагаемого ранжированного списка на предмет соответствия их анкетных данных специфике поставленной проблемы (рис. 13), а также анализ полученных обобщенных заключений по результатам экспертизы. Точность отбора при этом составила от 65 до 80 процентов в зависимости от предметной области, полноты предоставленного описания проблемы и наличия в БД анкет специалистов по требуемому направлению.

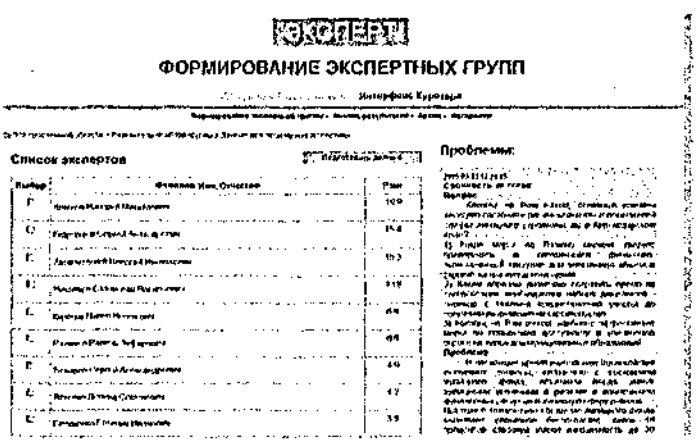


Рисунок 13 – Результаты подбора экспертов во внешнюю группу

Сравнив полученные результаты с данными методов анализа, основанных на Word-Net/таксономии и работе со специализированными предметными

областями (табл. 2), можно сделать вывод о достаточной адекватности, надежности и эффективности разработанных методик и алгоритмов.

Таблица 2 – Результаты сравнения с методами, основанными на таксономии

Название метода	Корреляция
Экспертное оценивание (эталон)	0,9015
Resnik (1995)	0,7450
Lin (1998)	0,8224
Li (2003)	0,8914
Мера Edge-counting	0,664
Мера Information-content	0,745
Jiang & Conrath (1998)	0,8484
Предлагаемый алгоритм ЛСА	0,8129

Таким образом, получаемые в ходе функционирования программного модуля результаты (с предусмотренной возможностью их корректировки модератором или администраторами экспертизы) позволяют осуществлять эффективный подбор специалистов с учетом специфики конкретных проблем, формируемых на естественном языке, в режиме реального времени обеспечивать проведение экспертизы и аналитическую обработку получаемых результатов.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В диссертационной работе формально поставлены и решены основные задачи лингвосемантического анализа естественно-языковых описаний с учетом факторов их неопределенности, неполноты и противоречивости. Разработаны подходы, методики и алгоритмы анализа и формализации информации, представленной на естественном языке.

При этом получен ряд новых результатов, к числу которых относятся:

1. Усовершенствованные математические модели и алгоритмы лингвосемантического анализа, формализации и обобщения естественно-языковых описаний, основанные на комплексном использовании результатов предварительного лингвистического, лексического и синтаксического видов анализа, что повышает эффективность формализации ЕЯ-описаний и адекватность используемого ЕМ-алгоритма в контуре обучения;
2. Методика, модели и алгоритм построения моделирующих семантических сетей для формального представления ЕЯ-описаний и экспертной информации, отличающийся от традиционно используемых метрик включением модуля кластеризации лексических образцов для определения семантически связанных терминов, что повышает эффективность и понижает вычислительную сложность алгоритма с ростом рассматриваемых параметров;
3. Подход к формированию итоговых обобщений ЕЯ-описаний и получаемой экспертной информации, заключающийся в формировании семантических пространств максимальной близости на основе применения ЕМ-алгоритма к результатам лингвосемантического анализа и дающий возможность исключать из рассмотрения неинформативных

либо незначимых терминов, а также управлять скоростью обучения с помощью задания величины порога близости.

4. Принципы и методика интеграции разработанных моделей и алгоритмов в состав ситуационных центров органов государственной власти с использованием инструментальных средств Data Mining. Сформированы требования к распределенной клиент-серверной архитектуре комплекса «Эксперт», видам обеспечения, подходы к его разработке и реализации в составе СЦ. Выделен ряд функциональных подсистем, обеспечивающих эффективную деятельность разработанного модуля: визуализации и представления данных; формирования проблемно-ориентированных экспертных групп; организации и проведения экспертиз.
5. Модель и регламент информационного взаимодействия участников процедур принятия решений, в которых выделены и описаны автоматизированный и «экспертный» контуры. Предложена структура программного комплекса со включением модулей лингвистического, морфологического, синтаксического и семантического видов анализа, реализующих разработанные алгоритмы и методики в применении к формированию проблемно-ориентированных экспертных групп в СЦ ОГВ, анализу, обобщению и формализации результатов экспертизы;
6. Самостоятельный практический интерес представляет программная реализация комплекса «Эксперт» на основе полученных теоретических результатов, с использованием архитектуры клиент-сервер и технологий интеллектуального анализа данных с учетом сформулированных требований к функциональности, режимам работы, программно-аппаратной совместимости, интегрируемости и управлению.
7. Методика оценка эффективности разработанных подходов, алгоритмов и их практической реализации, проведен сравнительный анализ эффективности и адекватности теоретического аппарата и разработанного программного комплекса с имеющимися метриками, алгоритмами и подходами – на основе коэффициентов корреляции, Спирмена, Пирсона и обработки эталонных наборов данных. Результаты оценки подтвердили вывод о достаточной адекватности, надежности и эффективности разработанных методик и алгоритмов.

Основные публикации по теме диссертации

Статьи в журналах, рекомендованных ВАК для публикации результатов диссертаций на соискание ученой степени доктора и кандидата наук:

1. Симанков В.С., Тарасов Е.С., Путято М.М., «Методологические основы принятия решений с использованием автоматизации неформальных процедур», Журнал «Естественные и технические науки», №4, 2010 г.
2. В. С. Симанков, Е. С. Тарасов. Методический подход к анализу и выработке приемов противодействия использованию нетрадиционных информационных каналов - Известия ТРТУ, №4. Информационная безопасность – Таганрог, 2005.

Другие издания:

3. Симанков В.С., Тарасов Е.С., «О проблемах управления проектированием информационных систем с учетом требований безопасности» – 4-я Международная заочная научно-практическая конференция «Прогрессивные технологии развития», Томск, 2008
4. Симанков В.С., Редько А.П., Тарасов Е.С., Котесников Д.А. и др. «Разработка теоретических основ и построение интеллектуальных систем мониторинга, анализа и поддержки принятия политических, социально-экономических и технологических решений регионального уровня для ситуационных центров органов власти. Конференция получателей грантов регионального конкурса «ЮГ» Российского фонда фундаментальных исследований». – Краснодар, 2008. С. 176-177
5. Симанков В.С., Тарасов Е.С. «Интеллектуальная подсистема лингвосемантического анализа для подбора экспертов по проблемным областям в рамках проведения психофизиологических исследований». Сборник трудов Юбилейной Десятой междунар. научно-практич. конференции. – Краснодар: изд-во КубГТУ, 2009. – 164 с. С. 121.
6. Симанков В.С., Тарасов Е.С. «Интеллектуальная подсистема лингвосемантического подбора экспертов с учётом специфики проблемной области». Всероссийская конференция с элементами научной школы для молодёжи «Проведение научных исследований в области обработки, хранения, передачи и защиты информации», Ульяновск, 2009.
7. Е.С. Тарасов. Разработка и реализация процедур функционирования экспертных групп в рамках ситуационных центров органов государственной власти - «Научно-практическая конференция *«Научно-техническое творчество молодежи* - путь к обществу, основанному на знаниях», Москва 2009
8. Симанков В.С., Тарасов Е.С. «Применение внешних проблемно-ориентированных экспертных групп в работе информационно-аналитических систем ситуационного центра» – Сборник трудов научно - практической конференции «Ситуационные центры 2009», Москва, РАГС, 2010.
9. Симанков В.С., Тарасов Е.С., Путято М.М. «О применении лингвосемантического подхода к подбору экспертов с учетом специфики проблемной области». III Международная научно-практическая конференция «Молодёжь и наука: реальность и будущее», том 5, Естественные и прикладные науки. 2010 г. Невинномысский институт экономики, управления и права.
10. Симанков В.С., Тарасов Е.С, Путято М.М. «Использование когнитивной графики для формализованного представления знаний экспертов и принятия решений», Международная научно-практическая конференция «Молодёжь и наука: реальность и будущее», том 5, Естественные и прикладные науки, 2010 г, Невинномысский институт экономики, управления и права.
11. Свидетельство об официальной регистрации программы для ЭВМ «Подсистема мониторинга и оценки эффективности деятельности органов государственной власти», (Симанков В.С., Черкасов А.Н., Путято М.М., Тарасов Е.С.) №2010614836 от 23.08.2010 г.

Подписано в печать 31.01.2011. Печать трафаретная.
Формат 60x84 1/16. Усл. печ. л. 1,35 Тираж 100 экз. Заказ № 438.
ООО «Издательский Дом-Юг»
350072, г. Краснодар, ул. Московская 2, корп. «В», оф. В-120
тел. 8-918-41-50-571
e-mail: olfomenko@yandex.ru Сайт: http://id-yug.narod2.ru