На правах рукописи

Бодров Даниил Александрович

ДИАЛОГОВЫЕ АЛГОРИТМЫ ПОИСКА И НАВИГАЦИИ В АВТОМАТИЗИРОВАННОЙ СИСТЕМЕ ТЕКСТОВОГО ДОКУМЕНТООБОРОТА МЕТАЛЛУРГИЧЕСКОГО ПРЕДПРИЯТИЯ

Специальность 05.13.01
Системный анализ, управление и обработка информации (металлургия)

Автореферат
диссертации на соискание ученой степени кандидата технических наук

Москва — 2007

Работа выполнена на кафедре АСУ в Государственном технологическом университете «Московский институт стали и сплавов».

Научный руководитель:
кандидат технических наук, доцент Поляков Владимир Николаевич

Официальные оппоненты:
доктор технических наук, профессор Попов Игорь Иванович
кандидат технических наук, доцент Филиппович Андрей Юрьевич

Ведущая организация:
Государственное образовательное учреждение высшего профессионального образования «Казанский государственный университет им. В. И. Ульянова-Ленина»

Защита состоится «31» октября 2007 г. в 14 часов на заседании Диссертационного Совета Д.212.132.07 при Государственном технологическом университете «Московский институт стали и сплавов» по адресу: 119049, Москва, Ленинский проспект, д. 4, ауд. _____.

С диссертацией можно ознакомиться в библиотеке Государственного технологического университета «Московский институт стали и сплавов».

Автореферат разослан «___» сентября 2007 г.

Ученый секретарь
диссертационного Совета _____ к. т. н., профессор Калашников Е. А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В настоящий момент большая часть документов на современном металлургическом предприятии, а также в других областях деятельности хранится в электронном виде.

Сейчас на металлургических предприятиях активно внедряются системы автоматизации документооборота, однако в них в первую очередь поддерживается поиск по значениям полей регистрационных карточек, а также по отдельным словам и словосочетаниям, использованным в тексте документа. Использование поиска по полям регистрационных карточек, требует от пользователей высокой дисциплины в заполнении этих полей, что на практике бывает достаточно редко.

Помимо корпоративной системы документооборота, на любом крупном предприятии имеется достаточное количество текстовой информации за ее пределами: электронная почта, базы нормативно-справочной информации, файловые архивы, другие системы (например, внутренние доски объявлений).

Отсутствие современных информационно-поисковых систем для предприятий металлургического комплекса приводит к снижению эффективности работы с документами, т. к. поиск должен производиться в нескольких источниках, а грубый поиск по вхождению слов приводит к большому информационному шуму или ненахождению необходимых документов из-за неправильной подборки ключевых слов.

Проблема усугубляется высокой степенью многозначности для слов металлургической тематики, которые зачастую пересекаются как со словами общей лексики, так и с другими специальными значениями.

Очень остро стоит вопрос поиска по полнотекстовой конструкторско-технологической документации в проектных организациях металлургической отрасли, например, таких как ОАО «Гипромез». Переход на новую систему стандартизации в связи со вступлением в ВТО, ставит вопрос о кросс-лингвистическом поиске и создании двуязычной терминологических системы, обладающей современными средствами поиска и навигации..

Еще одной областью применения полнотекстовых поисковых машин является патентный поиск. Сейчас в Интернет доступен поиск по крупнейшим базам данных патентов России, США, Европы. Однако современные системы предоставляют поиск только по ключевым словам, то есть для того, чтобы убедиться в новизне своей идеи автор вынужден перебирать различные варианты описания патента, самостоятельно подбирая синонимы, комбинируя ключевые слова. Это нелегко проделать даже для родного языка, поиск же на других языках становится еще более затруднительным.

Согласно многочисленным исследованиям, более 50 процентов пользователей заканчивают сеанс работы с поисковой системой, просматривая первые 10–20 ссылок. При этом, как правило, лишь 2–3 ссылки содержат действительно полезную для пользователя информацию. По различным оценкам 50–70% навигационного трафика в Интернет приходится на просмотр ошибочно найденных

страниц.

Итак, в настоящий момент для поисковых машин очевидны следующие области применения: поиск в Интернет, поиск в системах документооборота предприятий, патентный поиск, поиск в хранилищах текстовой информации (новости, научные ресурсы). Рост объема информации, происходящий одновременно с ростом информационных потребностей пользователей, ставит проблему эффективного информационного поиска остро как никогда ранее.

Таким образом, **актуальность работы** обуславливается огромным количеством доступной информации и отсутствием адекватных инструментов ее структурирования, поиска и навигации. К настоящему моменту накоплен достаточный объем знаний в области компьютерной лингвистики, поисковых технологий, разработки поисковых машин, построения пользовательских интерфейсов, кроме того, производительность современной вычислительной техники многократно превосходит ту, которая была в момент начала разработки большинства имеющихся поисковых систем, что позволяет решать задачу поиска качественно новыми способами.

Цель работы заключается в исследовании различных диалоговых (интерфейсных) механизмов поиска, основанных на изменении поисковых запросов, разработка и апробация диалоговых моделей фокусировки и расширения поиска в системах документооборота металлургических предприятий, а также исследование возможностей применения частотных зависимостей для помощи пользователям в формулировании запросов при патентном поиске.

Для достижения поставленных целей были решены следующие **задачи**:

– проанализированы имеющиеся подходы к организации интерфейса поисковых систем, выявлены их узкие места и направления развития;

– предложены диалоговые решения для повышения эффективности поиска, основанные на методах фокусировки, расширения и переформулирования запроса;

– предложены диалоговые решения для повышения эффективности поиска, основанные на частотных моделях;

– предложена формальная постановка задач расширения и фокусировки поиска, создано программное обеспечение для их решения;

– проведена оценка эффективности предложенных методов разрешения многозначности, фокусировки поиска, навигации по онтологиям при использовании в системах документооборота металлургических предприятий, патентного поиска, сети Интернет.

Научная новизна работы заключается в:

– формальной постановке задачи расширения и фокусировки поиска в интерфейсном модуле поисковой машины, основанной на использовании лексического значения;

– интерфейсной модели поисковой машины, основанной на технологиях разрешения многозначности;

– подтверждении возможности и эффективности применения частотных показателей при работе с лексическими онтологиями;

– математическом описании различных частотных факторов для исполь-

зования в пользовательском интерфейсе.

Практическая ценность работы заключается в следующем:

– выполнена формальная постановка задачи построения пользовательского интерфейса, основанного на технологиях разрешения многозначности, и разработке диалогового алгоритма фокусировки и расширения запроса;

– создана математическая модель частотных факторов при навигации по онтологической системе, которая позволяет строить пользовательские интерфейсы для различных сфер применения;

– использование результатов исследования при построении информационно-поисковой составляющей систем полнотекстового документооборота промышленного предприятия в металлургическом комплексе, должно привести к сокращению потерь и экономии оборотных средств;

– разработаны новые интерфейсных принципы с использованием лексических онтологий, которые позволяют строить более эффективные системы патентного поиска

– использование новых интерфейсных моделей при создании информационно-поисковых систем в Интернет, имеет потенциал сокращения общего объема передаваемой информации на 10%;

– использование разрешения многозначности может повысить отдачу от рекламы при размещении платных ссылок в результатах поиска в сети Интернет в 2–3 раза за счет лучшей фокусировки.

Методы исследования.

При выполнении работы использовались методы:

– алгоритмического моделирования;

– структурного программирования;

– реляционная модель построения баз данных;

– метод частотного анализа текстов;

– метод частотного анализа запросов к поисковым системам;

– методы семантического анализа текстов, основанные на разрешении лексической многозначности;

– методы системного анализа и принятия решений.

Результаты работы были практически реализованы в виде программных прототипов пользовательских интерфейсов. Методы организации интерфейса к лексической онтологии на частотных принципах приняты к внедрению в учебном процессе МИСиС для обучения по курсу «Лингвистические основы информатики».

На защиту выносятся следующие основные научные результаты:

– математическая модель пользовательского интерфейса к поисковой системе, основанной на технологиях разрешения многозначности;

– диалоговый алгоритм решения задачи информационного поиска, основанный на технологиях разрешения многозначности;

– математические модели использования частотных факторов при навигации в лексических онтологиях.

Работа производилась по следующим направлениям специальности 05.13.01:

– теоретико-множественный и теоретико-информационный анализ сложных систем;

– методы и алгоритмы интеллектуальной поддержки при принятии управленческих решений;

– визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации.

Публикации и апробация работы. По материалам исследований опубликовано 6 печатных работ, в том числе одна работа [6] в издании, входящем в Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертации на соискание ученой степени доктора и кандидата наук (редакция июль 2007 года), утвержденный Высшей аттестационной комиссией Министерства образования и науки Российской Федерации.

Результаты работы докладывались на следующих научных конференциях:

– Международный семинар Диалог’2002, Протвино, 6–11 июня 2002 г.;

– Когнитивное моделирование в лингвистике’2002, п. Дивноморское, сентябрь 2002 г.;

– International Workshop Speech and Computer (SPECOM’2003), Москва, 27–29 октября 2003 г.;

– Международный семинар Диалог’2003, Протвино, 11–16 июня 2003 г.

Работа выполнялась при частичной поддержке Российского Фонда Фундаментальных Исследований (грант РФФИ № 05-07-90939, «Система онтологического типа для поиска и обработки текстовой информации»).

Структура диссертации. Работа состоит из введения, четырех глав, заключения, списка литературы, изложенных на 130 страницах машинописного текста, содержит 19 рисунков, 9 таблиц, список литературы из 81 наименования.

СОДЕРЖАНИЕ РАБОТЫ

В **первой главе** дается анализ проблем в области полнотекстового документооборота на промышленном предприятии металлургического комплекса. Формулируется проблематика поиска и навигации в массиве документов.

Производится анализ современного состояния ИПС. В главе освещаются основные критерии оценки качества поисковых машин, освещаются последние исследования в этой области.

Дается краткое введение в проблематику информационно поиска. Проводится грань между информационным поиском и поиском по базам данных. Излагается основная терминология в области информационного поиска. Далее описываются основные принципы работы классических информационно-поисковых систем.

В первой главе также указываются основные общепринятые критерии оценки качества информационного поиска, и указывается разница между двумя ключевыми показателями качества — релевантностью и пертинентностью на-

ходимых поисковой машиной документов.

Релевантность определяется как формальный признак соответствия документу поисковому запросу, а пертинентность — как соответствие документа информационной потребности пользователя. Следует отметить, что современные поисковые системы достаточно далеко продвинулись в поиске релевантных документов, однако, добиться повышения пертинентности можно, только помогая пользователю в более качественной формулировке запроса.

В главе описаны основные проблемы, которые встают перед разработчиками поисковых систем, в том числе и для металлургии, основной из которых является проблема многозначности.

После этого произведен анализ основных направлений и путей развития современных информационно-поисковых систем и исследованы различные перспективные подходы к информационному поиску.

Также в главе дано краткое введение в проблему патентного поиска и рассмотрены традиционные методы поиска патентной информации, а также ее поиск в сети Интернет.

Анализ состояния дел в сфере текстового документооборота показывает, что на металлургических предприятиях хранится и обрабатывается значительное число документов нескольких категорий. При этом зачастую отсутствуют современные единые информационно-поисковые системы, а имеются только средства поиска по ключевым словам, что приводит к невозможности эффективно получить относящиеся к решаемой задаче документы, и в свою очередь приводит к снижению эффективности работы управляющего персонала и принятию неверных решений.

Анализ проблемы патентного поиска показывает, что в настоящий момент, благодаря доступности баз данных патентных ведомств в сети Интернет, патентный поиск с одной стороны становится доступен большему числу исследователей, но с другой из-за сложности имеющихся механизмов остается уделом специалистов, особенно для поиска на иностранных языках.

Решение поставленных проблем находится на пересечении компьютерной лингвистики, психологии, информатики.

Произведенный анализ литературы позволяет сделать следующие выводы.

В основе большинства популярных ныне поисковых машин лежат несколько усовершенствованные, но в целом хорошо известные и описанные, алгоритмы и результаты достаточно старых исследований. Однако в последнее время появляется все больше и больше оригинальных поисковых инструментов, причем как коммерческих, так и свободно распространяемых.

Ни в одной из известных систем нынешнего поколения в достаточной мере не используются механизмы обратной связи или модели пользователей, хотя как показывают исследования с участием автора, применение методов интерактивной фокусировки и уточнения запросов способно существенно сократить число непертинентных документов и весьма сильно сузить область поиска.

Итак, в результате анализа имеющихся исследований и эксплуатирующихся информационно поисковых систем можно сделать следующие выводы:

– практически все широко применяемые в настоящее время поисковые машины являются развитием традиционных подходов к поиску;

– многочисленные исследования показывают, что предел качества результатов традиционных подходов к информационному поиску практически достигнут и при этом не дает желаемой эффективности поиска;

– многочисленные исследования подтверждают, что можно добиться повышения качества поиска путем изменения способа взаимодействия с пользователями.

В последние годы были достигнуты значительные успехи в следующих областях, связанных с созданием информационно-поисковых механизмов:

– вычислительная мощность современной компьютерной техники позволяет производить более сложный анализ текстов и интерактивно взаимодействовать с пользователями;

– компьютерная лингвистика достигла значительных успехов в области морфоанализа, синтаксического и семантического анализа текстов.

Все эти факторы позволяют перейти к созданию поисковых систем, выполняющих более точный анализ текстов и осуществляющих интерактивное взаимодействие с пользователем с целью уточнения его информационных потребностей.

Во **второй главе** описан предлагаемый подход к интерактивным методам фокусировки и расширения поиска в системах документооборота на металлургических предприятиях.

В главе описаны предлагаемые методы фокусировки и расширения поиска среди которых:

– фокусировка на основе тематических и коммуникативных кластеров;
– фокусировка по устойчивым словосочетаниям;
– расширение на основании списка словообразований;
– расширение по аббревиатурам;
– переформулирование запросов;
– навигация по онтологиям;

Пример:

Для слова *прибыль* формируется следующий список устойчивых словосочетаний: *бухгалтерская прибыль, прибыль изложницы, прибыль компании, прибыль на капитал, прибыль населения, прибыль от сделок*.

Пример:

Для слова *прибыль* в БД ИПС в системе документооборота металлургического предприятия было обнаружено три предметные области: *металлургия, экономика, бухгалтерия*.

Пример:

Для одного из значений слова *прибыль* можно предложить такой список слов кластера *металлургия*: *изложница, металл, отливка …*

Содержательное описание методов на естественном языке позволяет сделать формальную постановку задачи поиска с использованием механизмов фокусировки/дефокусировки запроса.

Будем использовать следующие множества:

– множество документов: $D = \{d_1, d_2, ..., d_n\}$;

– множество лексем: $L = \{l_1, l_2, ..., l_m\}$;

– множество значений: $I = \{i_{11}, i_{12}, ..., i_{m1}, ..., i_{mk}\}$.

Между этими множествами возможны следующие отношения:

– пословный индекс — отношение $R_1(D,L): (l_j, d_i) \in R_1(D,L) \Leftrightarrow$ лексема l_j содержится в документе d_i;

– толковый словарь — отношение $R_2(I,L): (i_{jl}, l_j) \in R_2(I,L) \Leftrightarrow$ значение i_{jl} относится к лексеме l_j;

– индекс по значениям — отношение $R_3(D,I): (d_i, l_j) \in R_3(D,I) \Leftrightarrow$ лексема l_j содержится в документе d_i в значении i_{jl}.

Задачу поиска можно сформулировать следующим образом.

Поисковый запрос можно представить в виде множества $Z = \{z_1, z_2, ..., z_v\} \subset L$, где z_1 — ключевая лексема.

Тогда выполнение И-запроса можно описать как:

1. Построение сечения $R(z_1) = \{z_{1z}\} \subset I$ на отношении $R_2(I,L)$, т. е. множества лексических значений ядерной лексемы.

2. Построение сечения каждого элемента множества $R(z_1)$ — множества $R(z_{1z}) = \{d_{zd}\} \subset D$ на отношении $R_3(I,L)$, т. е. построение для каждого из возможных значений ядерной лексемы множества документов, куда оно входит.

3. Построение множества документов, в которые входят слова из поискового запроса $M\{m_d\}: m_d = R(z_{1z}) \cap \bigcap_{u=2}^{v} R(z_u) \subset D$, где $R(z_u) \subset D$ — сечение отношения $R_1(D,L)$ по z_u.

4. Выбор пользователем одного из элементов множества M, обозначенного далее $M_p \subset D$, где p — номер значения из подмножества $I^{z_1} = \{i_1, ..., i_p, ..., i_r\}$, $I^{z_1} \subset I$ ..

В зависимости от мощности множества M_p возможны следующие сценарии:

– если $|M_p| > 10$, то фокусировка запроса;

– если $|M_p| \le 10$, то расширение полноты поиска.

Для фокусировки запроса используются следующие множества:

– множество тематических кластеров $K\{k_k\}$;

– множество словосочетаний $E\{e_e\}$;

– множество вопросов $Q\{q_q\}$;

– множество коммуникативных кластеров $C\{c_c\}$.

И отношения:

– индекс по тематическим кластерам — отношение $R_K(K,D)$: $(k_k, d_i) \in R_K(K,D) \Leftrightarrow$ кластер k_k содержится в документе d_i;

– индекс по словосочетаниям — отношение $R_E(E,D)$: $(e_e, d_i) \in R_E(E,D) \Leftrightarrow$ словосочетание e_e содержится в документе d_i;

– значения в словосочетании — отношение $R_{IE}(I,E)$: $(i_{jl}, e_e) \in R_{IE}(I,E) \Leftrightarrow$ значение i_{jl} содержится в словосочетании e_e;

– индекс по вопросам — отношение $R_Q(Q,D)$: $(q_q, d_i) \in R_Q(Q,D) \Leftrightarrow$ вопрос q относится к документу d_i;

– индекс по коммуникативным кластерам — отношение $R_C(C,D)$: $(c_c, d_i) \in R_C(C,D) \Leftrightarrow$ кластер c (или его часть) содержится в документе d_i;

– онтологическая связь — отношение $R_O(I,I)$: $(i_{j_1l_1}, i_{j_2l_2}) \in R_O(I,I) \Leftrightarrow$ значение $i_{j_1l_1}$ состоит в онтологической связи со значением $i_{j_2l_2}$.

Если обозначить в отношениях фокусировки первое множество как $X\{x_x\}$, второе — как $Y\{y_y\}$, а отношение между ними как $R(X, Y)$, то задачу расширения полноты можно сформулировать в общем виде как построение сечения $R(x_x): (x_x, y_y) \in R(X,Y), R(x_x) \subset Y$ множества Y по x_x и последующее построение его пересечения с множеством M, т. е. построение множества $M_2 = M_1 \cap R(x_x),\ M_2 \subset D$.

Для расширения полноты поиска строятся отношения:

– словообразовательная парадигма — отношение $R_W(L,I)$: $(l_j, i_{jl}) \in R_W(L,I) \Leftrightarrow$ лексема l_j является морфологическим дериватом лексемы со значением i_{jl};

– синонимический ряд — отношение $R_S(I,I)$: $(i_{j_1l_1}, i_{j_2l_2}) \in R_S(I,I) \Leftrightarrow$ значение $i_{j_1l_1}$ является синонимом значения $i_{j_2l_2}$;

– аббревиатура — отношение $R_A(L,E)$: $(l_j, e_e) \in R_A(L,E) \Leftrightarrow$ лексема l_j является аббревиатурой словосочетания e_e;

– онтологическая связь — отношение $R_O(I,I)$: $(i_{j_1l_1}, i_{j_2l_2}) \in R_O(I,I) \Leftrightarrow$ значение $i_{j_1l_1}$ состоит в онтологической связи со значением $i_{j_2l_2}$.

Если обозначить в отношениях расширения полноты поиска первое множество как $X\{x_x\}$, второе — как $Y\{y_y\}$, а отношение между ними как $R(X, Y)$, то задачу расширения полноты можно сформулировать в общем виде как построение сечения $R(x_x): (x_x, y_y) \in R(X,Y), R(x_x) \subset Y$ множества Y по x_x. В результате получается подмножество элементов множества Y, которое можно использовать для новой итерации процесса поиска, т. е. для построения нового поискового запроса или объединения его с множеством $R(z_1)$.

Интерактивную поисковую систему можно рассматривать как систему управления, т. к. она обладает всеми необходимыми признаками такой системы.

В процессе взаимодействия с поисковой системой пользователь является

и лицом, принимающим решения (ЛПР) и потребителем информации, полученной в результате перевода поисковой системы в оптимальное состояние.

Результатом работы поисковой системы является список найденных документов. При этом результат поиска, как правило, оценивается, по крайней мере, по критериям суммарной пертинентности найденных документов (не сумма пертинентностей отдельных документов, а пертинентность квазидокумента, являющегося объединением множества документов) и количества найденных документов.

Теоретически, показателем результативности поиска является только первый критерий — удовлетворение информационной потребности пользователя множеством найденных документов. Однако, в силу того, что пользователь физически может ознакомиться только с ограниченным числом документов, одного этого критерия недостаточно. Тем не менее, критерий пертинентности можно считать главным.

Исходя их этого, задачу информационного поиска можно сформулировать следующим образом.

Найти такой поисковый запрос s^* (составленный с помощью множеств, описанных выше), который обеспечит выполнение следующих критериев

$$\max_{s\subset S} P\left(M\left(s\right)\right), \max_{s\subset S}\left(-\left|M\left(s\right)\right|\right); S=\left\{s: S\in\Omega, P\left(M\left(s\right)\right)>0, \left|M\left(s\right)\right|>0\right\} \qquad (1)$$

где $P(X)$ — суммарная пертинентность множества найденных документов;

$M(s)$ —множество результатов (найденных документов);

S — множество результативных запросов;

Ω — множество всех запросов.

Важно отметить, что сформулированная таким образом задача может не иметь решения, если в общем множестве документов D нет ни одного удовлетворяющего информационную потребность пользователя хоть в какой-то степени.

Так как, единственной известной характеристикой поискового запроса на первом этапе поиска является совокупность ключевых слов, целесообразно рассматривать задачу поиска, как задачу со следующими ограничениями, формируемыми пользователем в диалоговом режиме в процессе решения:

$$\begin{aligned} &E'\{e_e\}: (z_{1,1}, e_e) \subset R_{IE}(I,E), (e_e, d_m) \subset R_E(E,D) \\ &Q'\{q_q\}: (q_q, d_m) \subset R_Q(Q,D) \\ &K'\{k_k\}: (k_k, d_m) \subset R_K(K,D) \\ &C'\{c_c\}: (c_c, d_m) \subset R_C(C,D) \end{aligned} \qquad (2)$$

где d_m — элемент множества $M(s)$;

E' — множество включенных в запрос словосочетаний;

Q' — множество включенных в запрос вопросов;

K' — множество включенных в запрос тематических кластеров;

C' — множество включенных в запрос коммуникативных кластеров.

Задача информационного поиска по своей природе является нечеткой, что подтверждается достаточным количеством исследований, т. к. главный критерий оценки качества информационного поиска — пертинентность — является плохо формализуемой качественной характеристикой.

Для описания пертинентности введем следующую лингвистическую переменную:

$$\langle ПЕРТИНЕНТНОСТЬ, T(L), [0,1], G, M \rangle, \quad (3)$$

где $T(L)$ = {непертинентно, среднепертинентно, пертинентно} — терм-множество;

G — процедура образования новых термов с помощью связок и модификаторов типа «очень», «слегка», «совсем», «не» и др. Например: «малопертинентно»;

M — процедура задания на множестве [0, 1] нечетких подмножеств, выполняемая пользователем в процессе работы с поисковой системой.

Хотя мощность множества результатов вполне можно оценить количественно, точное значение пользователя не интересует, поэтому для его характеристики также можно ввести лингвистическую переменную:

$$\langle КОЛИЧЕСТВО\ РЕЗУЛЬТАТОВ, T(L), [0, |D|], G, M \rangle, \quad (4)$$

где $T(L)$ = {мало, много} — терм-множество;

G — процедура образования новых термов с помощью связок и модификаторов типа «очень», «слегка», «совсем», «не» и др. Например: «слишком много»;

$|D|$ — мощность множества всех документов;

M — процедура задания на множестве $[0, |D|]$ нечетких подмножеств, выполняемая пользователем в процессе работы с поисковой системой.

Полученную многокритериальную задачу (1) можно свести к однокритериальной путем выбора идеальной точки. Очевидно, что такой точкой будет единственный документ, полностью удовлетворяющий информационную потребность пользователя. Изменяя параметры запроса (ограничения (2)), пользователь приближается к этой идеальной точке, т. е. в процессе работы с поисковой системой он в интерактивном (диалоговом) режиме путем формулирования запроса включением в него или исключением из него слов, словосочетаний, кластеров и т. д. влияет на результат работы системы с целью получения значения «пертинентно» для лингвистической переменной *ПЕРТИНЕНТНОСТЬ* (3), на основе полученного значения и динамики его изменения, он принимает решения о целесообразности дальнейшего переформулирования запроса и спосо-

ба переформулирования.

Формулирование поискового запроса и получение результатов поиска происходит в диалоге с пользователем по следующему алгоритму.

1. Разбор поискового запроса, т. е. построение множества $Z = \{z_1, z_2, ..., z_v\} \subset L$, где z_1 — ключевая лексема.

2. Построение множества лексических значений ядерной лексемы, т. е. сечения $R(z_1) = \{z_{1,1} ... z_{1,z}\} \subset I$ на отношении $R_2(I, L)$.

3. Построение для каждого из возможных значений ядерной лексемы множества документов, куда оно входит, т. е. построение сечения каждого элемента множества $R(z_1)$ — множества $R(z_{1,z}) = \{d_{z,d}\} \subset D$ на отношении $R_3(I, L)$. В результате получается z множеств документов $M'\{m'_d\} : m'_d = R(z_{1,z}) \subset D$.

4. Построение множества документов, куда входят остальные слова поискового запроса, т. е. множества $D_Z = \bigcap_{u=2}^{v} R(z_u) \subset D$, где $R(z_u) \subset D$ — сечение отношения $R_1(D, L)$ по z_u.

5. Для каждого множества, полученного на шаге **Ошибка! Источник ссылки не найден.**, строим его пересечение с множеством D_Z, т. е. $M\{m_d\} : m_d = m'_d \cap D_Z \subset D$.

6. Выбор пользователем одного из элементов множества M, обозначенного далее $M_p \subset D$, где p — номер значения из подмножества $I^{z_1} = \{i_1, ..., i_p, ..., i_r\}$, $I^{z_1} \subset I$, соответствующее лексическое значение обозначено как i_p.

7. Если $|M_p| > 10$, то переход к шагу 8, если $|M_p| \le 10$, то переход к шагу 17.

8. Построение множества устойчивых словосочетаний, в которые входит значение i_p, т. е. $E_{z_{1,1}}\{e_e\} : (z_{1,1}, e_e) \in R_{IE}(I, E)$ — сечение отношения $R_{IE}(I, E)$ по i_p.

9. Построение множества устойчивых словосочетаний, входящих в документы множества M_p, т. е. $E_{M_1}\{e_e\} : (e_e, d_i) \in R_E(E, M_ç)$ — сечение отношения $R_E(E, D)$.

10. Построение множества словосочетаний для использования в запросе, т. е. множества $E_1 = E_{i_p} \cap E_{M_p}$.

11. Построение множества вопросов, на которые отвечают документы множества M_p, т. е. $Q_1\{q_q\} : (q_q, d_i) \in R_Q(Q, M_p)$— сечение отношения $R_Q(Q, D)$.

12. Построение множества тематических кластеров, в которые входят документы множества M_p, т. е. $K_1\{k_k\} : (k_k, d_i) \in R_K(K, M_p)$— сечение отношения $R_K(K, D)$.

13. Построение множества коммуникативных кластеров, в которые входят документы множества M_p, т. е. $C_{1_1}\{C_c\} : (c_c, d_i) \in R_C(C, M_p)$— сечение отношения $R_C(C, D)$.

14. Построение множества гипонимов значения i_p, т. е. множества

$O_{1_1}\{i_{jl}\}:(i_{jl},i_p)\in R_O(I,I)$.

15. Если пользователь удовлетворен результатами поиска, то переход к 17, если нет, то к 16.

16. Выбор пользователем одного из элементов множеств E_1, Q_1, K_1, C_1 или O_1. Если пользователь выбирает один из элементов E_1, Q_1, K_1, C_1, то строится множество $M'_{p_1}\{d_i\}: d_i \in M_p, (x,d_i)\in R_X(X,D)$, где x — выбранный элемент, X — соответствующее множество. Если пользователь выбирает элемент множества O_1, то строится сечение множества $R(i_{jl})=\{d_i\}\subset D$ на отношении $R_3(I,L)$ и множество $M'_{p_1}=R(i_{jl})\cap D_Z$, выбранный элемент обозначается как i_p. Множество M'_{p_1} обозначается как M_p и осуществляется переход к шагу 7.

17. Построение словообразовательной парадигмы для i_p, т. е. $W_1\{l_j\}:(l_j,i_p)\in R_W(L,I)$.

18. Построение синонимического ряда для i_p, т. е. $S_1\{i_{jl}\}:(i_{jl},i_p)\in R_S(I,I)$.

19. Построение списка словосочетаний, расшифровывающих аббревиатуру z_1, т. е. $A_1\{e_e\}:(z_1,e_e)\in R_A(L,E)$.

20. Построение множества когипонимов и гиперонимов значения i_p, т. е. множества $O_{1_1}\{i_{jl}\}:(i_{jl},z_{1,1})\in R_O(I,I)$.

21. Выбор пользователем одного из элементов множеств W_1, S_1, A_1 или O_1. Построение сечения $M'_{p_1}\subset D$ по выбранному элементу с помощью отношений R_1, R_3, R_E и R_3.

22. Построение нового множества $M'=M'_p\cup M_p$.

23. Обозначение M' как M_p и переход к шагу 7.

24. Завершение поиска.

В главе показано, на основе каких законов сделаны выводы об эффективности предлагаемых методов.

Считается, что человек способен эффективно обрабатывать одновременно от 5 до 9 объектов в зависимости от индивидуальных особенностей и некоторых других факторов.

Также экспериментально установленная зависимость времени реакции выбора от числа альтернативных сигналов, известная как Закон Хика, аппроксимируется логарифмической функцией следующего вида:

$$T = b\log_2(n+1), \tag{5}$$

где T — время реакции;

b — эмпирически устанавливаемая константа, получаемая аппроксимацией измерений;

n — число объектов.

Этими закономерностями часто руководствуются при промышленном дизайне, в том числе и при построении пользовательских интерфейсов, стараясь помещать объекты в небольшие группы по 5–7 объектов, однако в традицион-

ных поисковых системах в ответ на поисковый запрос пользователю возвращается не менее 10 результатов на страницу, что превышает указанный порог эффективного восприятия. Снижение числа результатов, отображаемых на странице только увеличивает сложность восприятия, т. к. разбиение не имеет какого-либо логического обоснования.

Для запроса по слову «стан» традиционная ПМ в сети Интернет формирует страницу результатов поиска, на которой перечислены подряд все найденные страницы без учета смыслового значения слова. В результате в проведенном эксперименте из первых десяти результатов только два имели отношение к металлургии.

ИПС в системе документооборота металлургического предприятия на том же входном множестве сформирует список результатов, сгруппированный по значениям, как показано на рис. 1.

Фрагмент страницы с результатами поиска ИПС

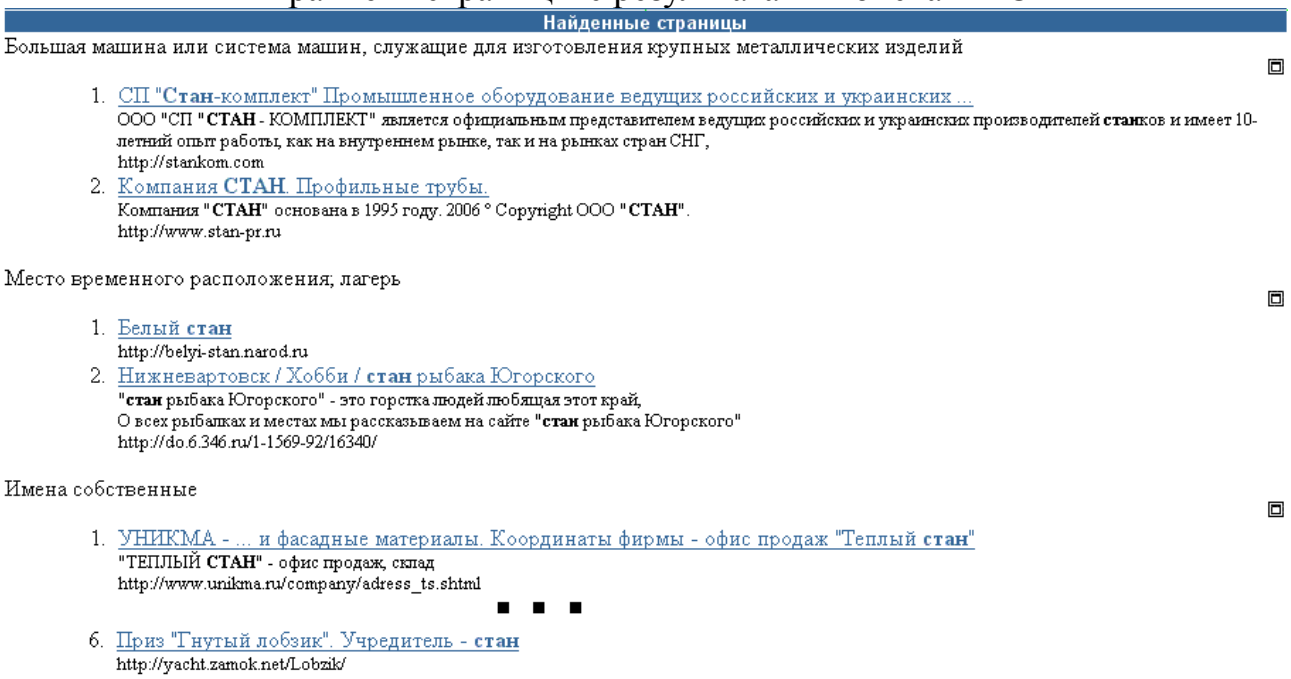

Рис. 1

Очевидно, что при таком разбиении, объем информации, которую должен проанализировать пользователь системы, эффективно понижается: вместо 10 неструктурированных информационных единиц, ЛПР получает три группы не более, чем по 4 единицы информации. Помимо снижения количества информации, происходит ее качественное изменение: информация делится на четко очерченные смысловые группы, внутри которых информация однородна. Это приводит к уменьшению константы b в законе Хика.

В ходе дальнейшего диалога с ИПС пользователь с помощью предлагаемых средств фокусировки постепенно уменьшает количество предлагаемой к анализу информации, одновременно повышая ее качество, оставляя результаты, только относящиеся к выбранной пользователем области.

На основе предложенного алгоритма разработан следующий обобщен-

ный алгоритм работы пользователя с интерфейсным (диалоговым) блоком ИПС в системе документооборота металлургического предприятия, представленный на рисунке 2

Алгоритм сценария поиска

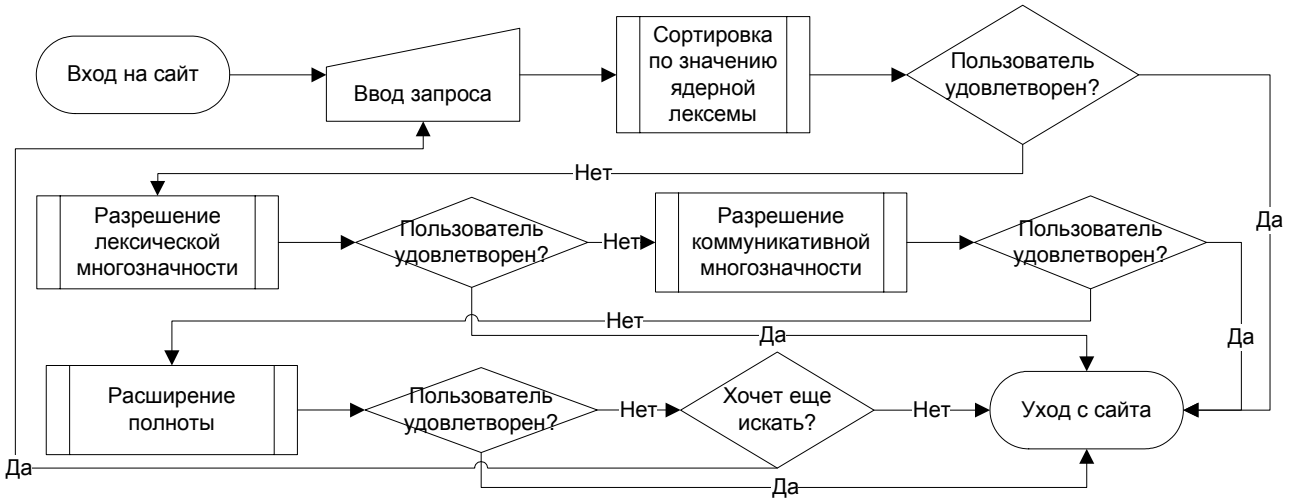

Рис. 2

Алгоритм обработки поискового запроса на первом этапе поиска (сортировка по значению ядерной лексемы) представлен на рисунке 3.

Алгоритм обработки запроса

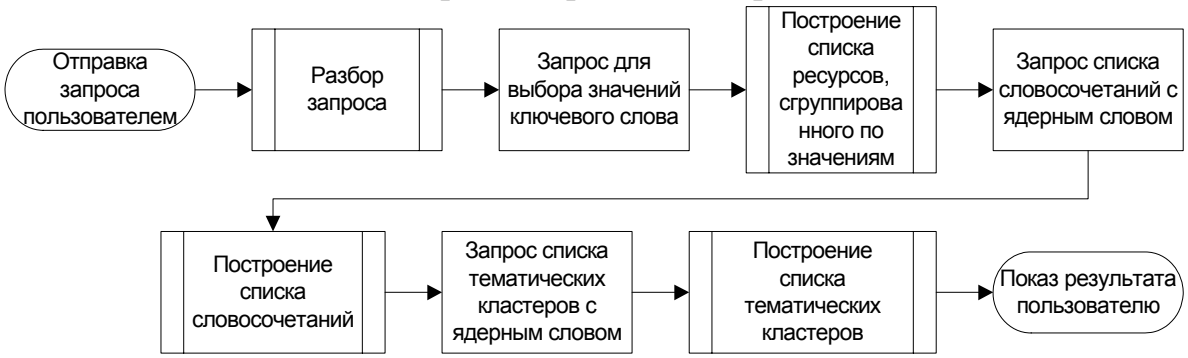

Рис. 3

ИПС в системе документооборота металлургического предприятия включает несколько программных компонентов, схема взаимодействия которых представлена на рисунке 4.

Объектом данной работы являются интерфейсный блок и база данных ИПС в системе документооборота металлургического предприятия.

Программно-технический комплекс «Интерфейсный блок» (в дальнейшем ИБ) предназначен для использования в качестве составной части интеллектуальной поисковой системы и отвечает за выполнение следующих функций:

– ввод поискового образа запроса пользователем с экрана;

– предварительный анализ запроса, включая когнитивный морфологический анализ, выявление типа запроса, генерация сценария обработки запроса, формирование запросов к БД;

– расчет рейтинга значений слов и ссылок для каждого значения;

– настройку на профиль пользователя и выбор оптимальных режимов отображения;

– синтез страницы результатов выполнения запроса;

– активный диалог с пользователем по выбору лексических и коммуникативных значений, навигации в результатах выполнения запроса, переформулирование запроса к БД;

– сервисные функции.

Схема взаимодействия программных блоков ИПС в системе документооборота металлургического предприятия

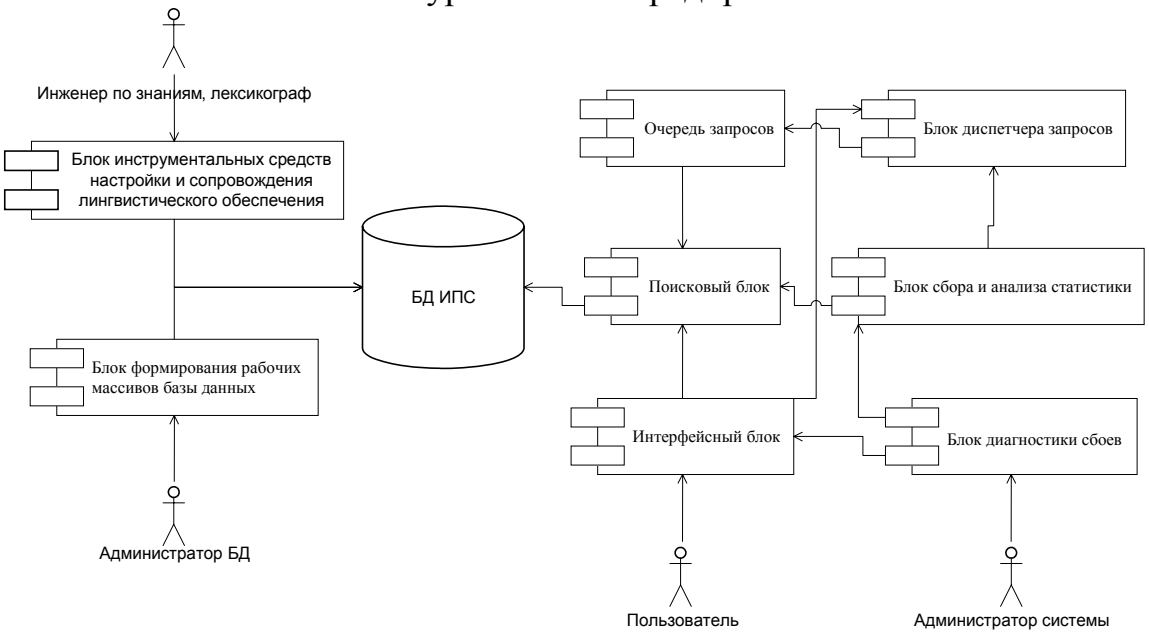

Рис. 4

В рамках работы был изготовлен прототип, который можно рассматривать как ядро промышленной версии, которая будет содержать ряд функциональных дополнений и технических решений, обеспечивающих распространение на рынке поисковых технологий.

Так как проект носит исследовательский характер в ИПС в системе документооборота металлургического предприятия позволено менять большинство настроек. Это сделано для того, чтобы иметь возможность исследовать разные режимы поиска, с одной стороны, и, чтобы удовлетворить самым разнообразным вкусам пользователя, с другой стороны.

В **третьей главе** описана Навигация в пределах лексической онтологии с учетом частотных факторов в задачах патентного поиска

Обычная методика организации лексической информации по алфавитному принципу группирует слова, которые сходны по написанию, но распыляет по всему списку слова со сходными или связанными значениями. К сожалению этому нет доступной альтернативы, которая позволяла бы с такой же легкостью пользователям находить слова, а лексикографам работать. Однако поиск по алфавитно-организованному словарю, как правило, достаточно скучен и отнимает

много времени.

Проект WordNet предлагает более эффективную комбинацию традиционных лексикографических механизмов и возможностей современной техники. WordNet — БД лексических (семантических) связей, созданная в соответствии с теорией человеческой лексической памяти. Английские слова в ней организованы в синонимические ряды, представляющие некоторое понятие, а они, в свою очередь, связаны различными семантическими связями.

Предложенные интерфейсы для навигации по онтологиям не являются до конца интуитивными и легкими в использовании. Особенно большие затруднения могут возникать при использовании онтологической сети пользователями, имеющими ограниченный словарный запас. Это в первую очередь подростки, люди без специального образования. Связано это со спецификой организации онтологической сети.

Онтологические сети можно использовать для навигации в системах информационного поиска (в том числе в системах документооборота металлургических предприятий) и в системах патентного поиска.

В работе для целей навигации предложено использовать три частотных фактора: частотная функция встречаемости слова-узла, вес поддерева, число подчиненных узлов (лексических термов). Предложено несколько моделей учета частотных факторов при организации интерфейса. Результаты работы планируется использовать в рамках проекта Интеллектуальная поисковая машина.

Базовая посылка настоящего исследования заключается в том, что в силу частотных закономерностей большинство пользователей поисковых систем интересует частотная лексика. Можно выделить следующие четыре подхода к маркированию лексики, позволяющие организовать частотно-зависимый онтологический интерфейс для различных категорий и информационных потребностей пользователей:

1. Маркирование лексики с максимальной частотой использования (для пользователей которых не интересует специфическая, а вполне удовлетворяет общеупотребительная в рамках данного онтологического класса лексика).

2. Маркирование лексики с минимальной частотой использования (для пользователей со специфическими интересами).

3. Маркирование лексики, наиболее часто встречающейся в запросах других пользователей к поисковой системе.

4. Маркирование часто используемых путей в лексической онтологии.

В зависимости от конечной цели построения частотного интерфейса можно сконструировать различные частотные функции, которые должны обладать следующим основным свойством:

$$x >= y \Rightarrow F(x) >= F(y)\,. \qquad (6)$$

и дополнительными свойствами (при их отсутствии требуется дополнительная нормировка значений частотных функций):

$$x \in [0,1] \Rightarrow F(x) \in [0,1] \qquad (7)$$

$$F(0)=0, F(1)=1 \tag{8}$$

В эксперименте были использованы описанные ниже частотные функции.

Линейная шкала:

$$F_i = f_i, \tag{9}$$

где F_i — частотная функция узла, используемая для индикации;

f_i — нормированная частота узла.

Квадратичная «усиливающая» шкала:

$$F_i = f_i^2 \tag{10}$$

Корневая «сглаживающая» шкала:

$$F_i = \sqrt{f_i} \tag{11}$$

На рисунке 5 проиллюстрировано усиливающее и сглаживающее действие степенных шкал по сравнению с линейной. Интерфейс отображает значения линейной шкалы, средняя кривая — аппроксимацию рассчитанных значений линейной частотной функции гиперболической функцией, нижняя — квадратичной функции, верхняя — корневой функции.

Влияние вида частотной функции на интерфейс

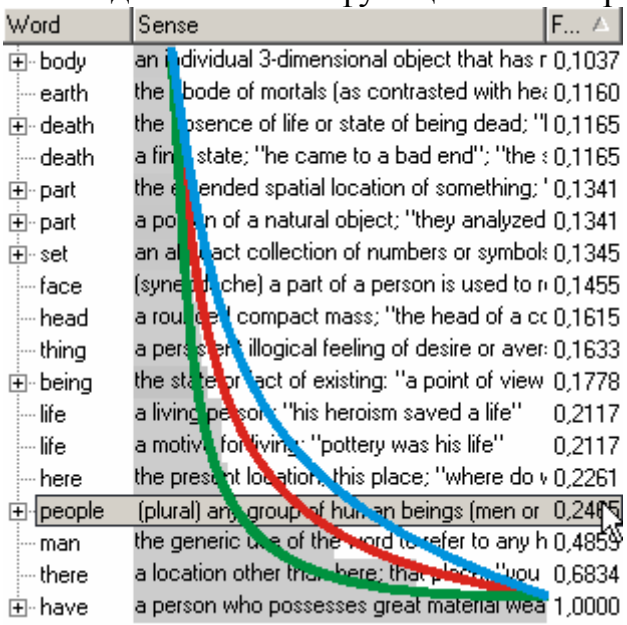

Рис. 5

Под весом поддерева подразумевается величина, зависящая от частотных факторов входящих в него дочерних узлов (гипонимов). Подобно частотным функциям, можно сконструировать множество методов вычисления весов. Единственным жестким требование к этим методам, является прямая (а не об-

ратная) зависимость от частот входящих узлов.

Далее приводятся несколько вариантов методов.

Усредненный:

$$W_j = \frac{\sum_{i=1}^{n} F_{ji}}{n}, \tag{12}$$

где W_j — вес поддерева;

n — число дочерних узлов;

F_{ji} — частотная функция дочернего узла i.

Основной недостаток этого метода заключается в «размывании» результата за счет служебных (технических) уровней и слабочастотных слов.

Усредненный с отсечением:

$$W_j = \frac{\sum_{i=1,n,F_{ji}>F_0} F_{ji}}{n}, \tag{13}$$

где F_0 — заданный уровень отсечения.

Данный метод избавлен от недостатка предыдущего за счет исключения низкочастотных узлов, однако может искажать результат.

Эти методы (усредненный и усредненный с отсечением) можно легко модифицировать для получения методов второго типа, добавив в расчет корневой узел. Учитывая тривиальность изменений, формула здесь не приводится.

Максимальный из подчиненных:

$$W_j = \max_{i=1,n}(F_{ji}) \tag{11}$$

Данный метод не вполне точно характеризует вес поддерева, т. к. не учитывает число узлов, однако в этом случае частотная функция высокочастотного узла служит своего рода маяком при навигации.

Максимальный из подчиненных уровнем ниже:

$$W_j = \max_{i=1,n,l_{ji}=1} (F_{ji}), \tag{14}$$

где l_{ji} — уровень узла (расстояние от узла до вершины поддерева).

Метод позволяет экономить ресурсы, ограничивая количество просматриваемых узлов, однако к недостатку предыдущего метода добавляет опасность получить неоправданно низкий вес поддерева в случае большого числа низкочастотных или технических узлов на следующем уровне.

Комбинированный:

$$W^*_i = F_i + W_i, \tag{15}$$

где F_i — вес узла, порождающего поддерево;

W_i — вес поддерева, рассчитанный одним из предыдущих методов.

В отличие от предыдущих методов, где нормирование желательно для получения более наглядных индикаторов, для последнего метода нормирование обязательно, т. к. существует вероятность получения значения, не лежащего в диапазоне [0; 1].

Вычисление показателей веса поддерева и предоставление этой информации пользователю позволяет оценивать перспективность направлений навигации по онтологической сети без углубления на каждом из узлов, что также сокращает время переформулирования запроса.

Число подчиненных частотных узлов N_i показывает число «перспективных» узлов в поддереве.

$$N_i = \left|\{L_i, W_i > W_0\}\right|, \tag{16}$$

N_i определяется как мощность множества L лексических термов онтологии, для которых вес поддерева W_i превышает пороговое значение W_0.

Аналогичный частотный фактор можно сконструировать на базе частотной функции узла:

$$N_i = \left|\{L_i, F_i > F_0\}\right|. \tag{17}$$

Этот метод не учитывает общего количества узлов или величину отклонения от порогового значения.

Пороговую частоту веса узла можно использовать как частоту отсечения малочастотных узлов в случае принятия гипотезы о том, что пользователей интересует более употребительная лексика.

Выбор способов визуализации частотных факторов опирался на следующие соображения:

– метод должен быть интуитивно-понятным и не требующим дополнительных разъяснений;

– необходимо компактное и в тоже время полное отображение информации;

– необходимо органичное сочетание способов визуализации с характером информации, представленной в лексических онтологиях и между собой.

В итоге были рассмотрены и опробованы на практике различные средства визуализации частотных факторов. В результате экспериментирования был сделан вывод, что оптимальным сочетанием визуальных факторов будет следующее:

– для отображения частотного фактора лучше всего подходит линейный индикатор (длина индикатора отображает относительную частоту узла, насы-

щенность тона — вес поддерева);

– для отображения порога отсечения лучше подходит сочетание насыщенности и размера шрифта;

– для отображения числа узлов выше уровня отсечения лучше всего подходит цифровой индикатор.

Апробация предложенной интерфейсной когнитивной модели проводилось в приложении OntoBrowser. Общий вид приложения представлен на рисунке 6. В левом окне отображается онтологическое дерево с визуализаций частотного фактора, в правом — частотные узлы (с частотой выше частоты отсечения) выбранного поддерева и их количество.

Общий вид приложения OntoBrowser

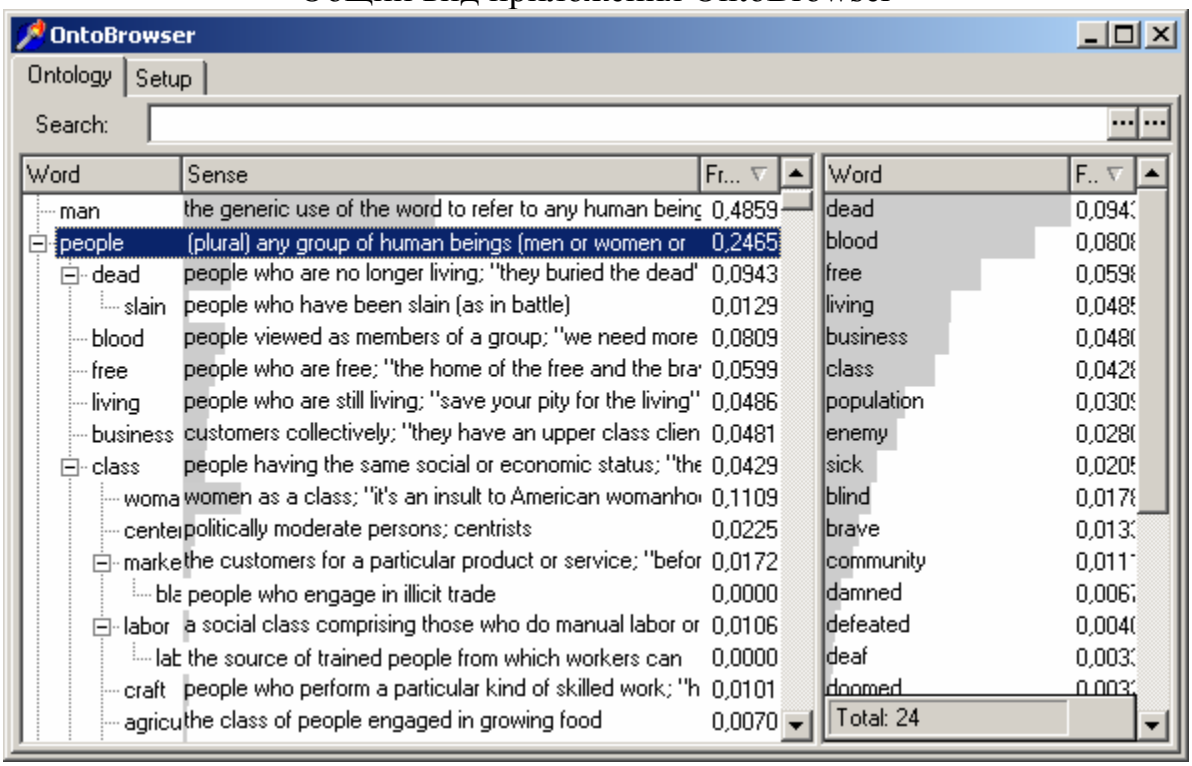

Рис. 6

В **четвертой главе** делается оценка эффективности предложенных диалоговых методов.

Хотя разрабатываемые в рамках проекта ИПМ технологии могут достаточно эффективно применяться в различных средах, в том числе и в сети Интернет, наибольшая результативность ожидается при их применении в рамках систем, ориентированных на специфические группы пользователей, такие как сотрудники предприятий или участники образовательного процесса. Обусловлено это несколькими основными причинами.

Использование технологий разрешения многозначности дает наилучшие результаты, если пользователи хорошо знакомы с терминологией предметной области и имеют четкие информационные потребности. Тем не менее, это не означает неэффективности применения технологий, основанных на использовании лексических значений, для пользователей не вполне хорошо знакомых с предметной областью.

Для эффективной работы с ИПС, построенной на предложенных принципах, пользователи должны быть нацелены на нахождение всех документов, отвечающих их информационной потребности. Однако для поисковых систем в сети Интернет большинству пользователей достаточно получить только какую-то часть документов, соответствующих их запросу.

Стоимость неэффективного поиска в системах документооборота, патентных системах и т. п. гораздо выше, чем при поиске в Интернет. Как правило, на основании только поиска в публичных сетях не принимается важных для деятельности предприятия решений, однако такие решения могут приниматься по результатам поиска в корпоративных информационных системах или базах данных патентных ведомств.

Использование новых поисковых технологий предполагает некоторое изменение стиля работы пользователей с поисковой системой по сравнению с традиционным, на что многие пользователи пойдут неохотно, если их информационные потребности недостаточно важны для них самих.

И, наконец, внедрение сложных поисковых технологий дороже, чем использование поисковых систем, основанных исключительно на поиске ключевых слов. Это вызвано более высокими требованиями к вычислительной мощности (из-за большего числа индексируемых признаков), а также необходимостью привлечения лингвистов на этапе построения семантических связей, формирования кластеров и т. д.

Экономический эффект от внедрения системы электронного документооборота определяется различными составляющими для различных категорий работников. Было бы заблуждением считать, что эффект определяется главным образом экономией на заработной плате за счет экономии времени на рутинных операциях. Автоматизация документооборота несет организации следующие выгоды:

– уменьшение количества претензий, не обработанных в срок, приводящих к выплате неустоек контрагентам или штрафов государству;

– сокращение сроков вывода новой продукции на рынок;

– сокращение времени на рутинную работу с документами сотрудников подразделений, формирующих доход организации, что высвобождает время лиц, принимающих решения, для работы над выработкой решений;

– повышение исполнительской дисциплины — своевременная реализация распоряжений руководителей, как следствие, сокращение времени выпуска новой продукции на рынок.

Можно утверждать, что положительный эффект от системы электронного документооборота можно оценить по следующей формуле (носящей иллюстративный характер и не претендующей на точный учет всех факторов):

$$R = F \cdot K_F + C \cdot N \cdot E + P \cdot K_P, \qquad (18)$$

где R — ожидаемая отдача от внедрения системы электронного документооборота в год;

F — сумма штрафов и неустоек, выплачиваемая до внедрения системы документооборота в год;

K_F — ожидаемое снижение суммы штрафов и неустоек, в долях;

C — средние затраты (зарплата, налоги, накладные расходы) на одного сотрудника в год;

N — количество сотрудников;

E — ожидаемая экономия времени на обработку документов, в долях;

P — прибыль от выпуска новой продукции в год;

K_P — ожидаемое повышение прибыли за счет уменьшения срока выпуска новой продукции, в долях.

Таким образом, для того, чтобы система электронного документооборота окупилась в первый год, общая стоимость владения системой в первый год не должна превышать R.

Основной негативный эффект для крупных предприятий (как металлургических, так и других) от недостаточно тщательного патентного поиска, в том числе и в патентных ведомствах иностранных государств заключается в непреднамеренном использовании защищаемых патентом технологий. Следствием этого являются длительные судебные разбирательства, по результатам которых выплачиваются крупные штрафы.

Второй составляющей выгоды от проведения эффективного патентного поиска может являться прибыль от вывода на рынок инновационного продукта. В том случае, если патентный поиск перед началом разработки и производства нового продукта производится недостаточно тщательно, к моменту выхода на рынок может оказаться, что продукт с похожими характеристиками уже предложен конкурентом, что приводит к снижению прибыли и может сделать продукт убыточным.

При внедрении механизма разрешения лексической многозначности в ИПС можно рассматривать два фактора влияния на рекламные технологии и экономические характеристики функционирования Web-ресурсов:

1. Повышение таргетинга (попадания к целевой группе) рекламы.
2. Повышение эффективности поиска и, как следствие, снижение трафика (объема загружаемой информации) на сайте и во всей сети.

Для оценки изменения пертинентности можно принять следующие положения:

– первая страница результатов поиска — некая стандартная пертинентность;

– вторая и последующие страницы — пертинентность увеличивается за счет разрешения многозначности на коэффициент k (со второй страницы начинают работать механизмы фокусировки).

Среднее повышение пертинентности можно вычислить по следующей формуле:

$$K_P = \left(U_1 \cdot P + U_N \cdot P \cdot k\right) / P, \qquad (19)$$

где K_P — коэффициент повышения пертинентности за счет разреше-

ния многозначности;

U_1 — доля пользователей, просматривающих только первую страницу результатов поиска;

P — исходная пертинентность;

U_N — доля пользователей, просматривающих не только первую страницу результатов поиска.

Тогда с учетом собранной статистики[1]:

$K_P = (0{,}58 \cdot P + 0{,}42 \cdot P \cdot k)/P = 0{,}58 + 0{,}42 \cdot 5{,}53 \approx 2{,}9$

Известно, что средняя эффективность рекламы (оцененная как число пользователей перешедших по ссылке по отношению к общему числу) составляет для традиционной поисковой системы 1%. За счет улучшения таргетинга рекламы среднее число заинтересованных пользователей увеличится пропорционально повышению пертинентности и составит:

$CTR' = K_P \cdot 1\% = 2{,}9\%$

То есть, с внедрением механизма разрешения лексической многозначности ключевых слов в ИПС, эффективность баннерной рекламы в среднем возрастет в 2,9 раз.

Коэффициент сокращения обращений поисковых систем равен 2,9 (равен среднему повышению пертинентности).

Можно рассчитать общий эффект сокращения трафика в пределах всей сети:

$$K_{ñîêð} = K_{íàâ} \cdot K_{ÏÑ} \cdot K_{çàïð} - K_{íàâ} \cdot K_{ÏÑ} \cdot K_{çàïð} / K_P, \qquad (20)$$

где $K_{ñîêð}$ — общий эффект сокращения трафика;

$K_{íàâ}$ — доля навигационных сервисов в общем трафике;

$K_{ÏÑ}$ — доля поисковых систем в трафике навигационных сервисов;

$K_{çàïð}$ — доля многозначных запросов по одному слову среди общего числа запросов к поисковым системам.

Таким образом, с учетом имеющейся статистики по данным показателям, в результате применения технологий разрешения многозначности в крупнейших поисковых системах общее сокращение трафика в пределах всей сети Интернет может составить:

---

[1] Следует отметить, что многозначность рассчитывалась по именам существительным, вошедшим в данную выборку, и учитывала, в том числе, оттенки значений. Реальная многозначность всех слов языка может оказаться ниже (особенно, если в расчет принимать не только имена существительные).

$$K_{\text{ñîêð}} = 0{,}40 \cdot 0{,}80 \cdot 0{,}42 - 0{,}40 \cdot 0{,}80 \cdot 0{,}42 / 2{,}9 = 0{,}0881 = 8{,}81\%$$

ЗАКЛЮЧЕНИЕ

В настоящей работе произведено исследование методов повышения качества поисковой составляющей в системах документооборота в металлургии и в системах патентного поиска.

В ходе выполнения работы выяснилось, что большинство опубликованных исследований ориентировано на развитие традиционных подходов к созданию ИПС, в то же время исследования, направленные на использование механизмов разрешения многозначности или построения интерактивных поисковых систем практически отсутствуют.

Данная работа, в свою очередь, была направлена на исследование различных диалоговых методов повышения качества поиска, основанных на интерактивном изменении и уточнении поисковых запросов, а также проведена апробация диалоговых моделей фокусировки, расширения поиска и моделей частотного интерфейса формулирования запросов с использованием онтологий.

Созданные в результате работы модели позволили убедиться в правильности сделанных предположений о применимости интерактивного взаимодействия с пользователями для повышения качества поиска в системах документооборота металлургических предприятий. Созданное приложение-прототип продемонстрировало применимость онтологий в совокупности с частотными оценками для формулирования поисковых запросов при патентном поиске.

Проделанная работа привела к следующим результатам и выводам:

1. Выполнена формальная постановка задачи повышения качества поиска путем переформулирования запросов. Формализация охватывает следующие способы фокусировки запроса: тематические кластеры, словосочетания, вопросы, коммуникативные кластеры, и следующие способы расширения полноты поиска: тематические кластеры, словосочетания, вопросы, коммуникативные кластеры, и следующие способы расширения полноты поиска: словообразовательная парадигма, синонимы, аббревиатуры и онтологии. Введены формальные критерии качества поиска на основе понятий пертинентности и мощности множества результатов.

2. На основе выполненной формальной постановки задачи поиска предложен алгоритм построения диалоговой поисковой системы, использующей технологии разрешения многозначности.

3. Разработана реляционная модель данных, позволяющая описывать структуру текстов с целью улучшения возможности поиска, и прототип поисковой системы, в пользовательском интерфейсе которой были использованы механизмы разрешения многозначности.

4. Разработаны общие требования к частотным функциям, используемым в частотном интерфейсе. Разработано несколько частотных функций, проанализированы отношения между ними. Разработан метод оценки перспективности навигации по древовидной структуре на основе веса поддерева. Введены несколько вариантов таких методов и проанализированы их относительные пре-

имущества и недостатки.

5. Проведено исследование возможностей использования частотных факторов в организации диалога информационной системы с пользователем для облегчения использования сложных механизмов поиска. Сформулированы подходы, позволяющие организовать частотно-зависимый онтологический интерфейс для различных категорий и информационных потребностей пользователей. Разработано специализированное приложение-прототип для оценки возможностей применения различных частотных факторов и способов их визуализации.

6. Сделан вывод о наибольшей эффективности предлагаемых механизмов при применении их в информационных системах промышленных предприятий (в том числе металлургических), а также в системах патентного поиска, т. е. при использовании специалистами в предметной области (хорошо владеющих специальной лексикой).

7. Результаты исследований применяются в проекте Интеллектуальной поисковой машины, в учебном процессе МИСиС по курсу «Лингвистические основы информатики», могут быть применены в системах документооборота металлургического предприятия, в системах патентного поиска, а также других информационных системах.

По теме диссертации опубликованы следующие работы:

1. Поляков В.Н., Бодров Д. А., Точин А. В. Интерактивные методы фокусировки и Расширения поиска в интеллектуальной поисковой машине // Компьютерная лингвистика и интеллектуальные технологии: Тр. Международного семинара Диалог’2002. (Протвино, 6–11 июня 2002 г.): В 2 т. / Под ред. А. С. Нариньяни. — М.: Наука, 2002. Т. 2: Прикладные проблемы. Стр. 438–449.

2. Бодров Д. А., Поляков В.Н. Проблемы создания эффективных поисковых машин (обзорная статья) // Обработка текста и когнитивные технологии: Сборник (Вып. 7) / Под ред. Соловьева В. Д. — Казань: 2002. Стр. 8–55.

3. Поляков В. Н., Бодров Д. А. Навигация в пределах лексической онтологии с учетом частотных факторов // Компьютерная лингвистика и интеллектуальные технологии: Тр. Международного семинара Диалог’2003. (Протвино, 11–16 июня 2003 г.) / Под ред. И. М. Кобозевой, Н. И. Лауфер, В. П. Селегея. — М.: Наука, 2003. Стр. 554–568.

4. Bodrov D. A., Polyakov V. N. Frequency Factors For Navigation through Lexical Ontology // Proceedings of the International Workshop Speech and Computer (SPECOM’2003), Moscow, Russia, October 2003. — M: 2003. Стр. 77–87.

5. Бодров Д. А., Кожитов С. Л., Поляков В. Н. Автоматизация текстового оборота на металлургическом предприятии и новые поисковые технологии // Перспективные технологии и оборудование для материаловедения и наноэлектроники: Материалы семинара / Под ред. проф. Л. В. Кожитова, проф. В. К. Карпасюка. — М.: МГИУ, 2006 — 741 с.

6. Бодров Д. А., Кожитов С. Л., Поляков В. Н. Задачи интерактивной обработки поисковых запросов в теоретико-множественной постановке. // Известия Саратовского университета. Новая серия. Серия «Математика. Механика. Информатика» — Саратов: 2007. Том 7. Выпуск 1. Стр. 78-83.