СЕДОВА ЯНА АНАТОЛЬЕВНА

МОДЕЛИ И АЛГОРИТМЫ ОБРАБОТКИ КОРПУСА ДОКУМЕНТОВ НАУЧНОЙ ИНФОРМАЦИИ

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Астрахань – 2011

Работа выполнена в Федеральном государственном образовательном учреждении высшего профессионального образования «Астраханский государственный технический университет»

Научный руководитель: доктор технических наук, профессор **Квятковская Ирина Юрьевна.**

Официальные оппоненты: заслуженный деятель науки РФ, *доктор технических наук, профессор* **Камаев Валерий Анатольевич,**

кандидат технических наук, доцент **Щербатов Иван Анатольевич.**

Ведущая организация: **ГОУ ВПО «Тамбовский государственный технический университет».**

Защита состоится 14 мая 2011 г. в 12 часов 00 минут на заседании диссертационного совета Д.307.001.06 при Астраханском государственном техническом университете по адресу: 414025, г. Астрахань, ул. Татищева 16, ауд. Г. 305.

Отзывы на автореферат в двух экземплярах, заверенные гербовой печатью организации, просим направлять по адресу: 414025, г. Астрахань, ул. Татищева, 16, ученому секретарю диссертационного совета Д.307.001.06.

С диссертацией можно ознакомиться в библиотеке Астраханского государственного технического университета.

Автореферат разослан «13» апреля 2011 г.

Ученый секретарь
диссертационного совета А. А. Ханова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Современный этап развития науки характеризуется значительными темпами увеличения объема научного знания, представленного в виде диссертаций и авторефератов. Согласно статистике, в последние годы количество диссертаций, ежегодно утверждающихся Высшей аттестационной комиссией, в среднем растет на 5% в год. Часть научных знаний хранится в виде корпусов документов, содержащих монографии, публикации, эссе, диссертации и т. д. Наибольшей информативностью и достоверностью обладает автореферат диссертации, который полностью повторяет ее терминологию и позволяет представить диссертацию в сжатом виде.

Особенностью представления документальных научных знаний является их слабая структурированность, что делает невозможным их автоматическую обработку для организации эффективного доступа к знаниям.

Вопросами автоматизации анализа естественного языка занимались многие ученые как в нашей стране, так и за рубежом: в области автоматического понимания текстов – Р. Шенк, Э. В. Попов, Н. Н. Леонтьева, Э. Ф. Скороходько, в области разработки информационно-поисковых систем – П. И. Браславский, И. Е. Кураленок, И. С. Некрестьянов, Б. В. Добров, Д. В. Ланде, Н. В. Лукашевич, в области разработки семантических моделей текста – Т. А. Гаврилова, В. Ф. Хорошевский, А. Е. Ермаков, A. Maedche, E. Alfonseca, E. Agirre, в области выделения терминов из текста – Е. И. Большакова, K. Frantzi.

Работы этих авторов привели к созданию ряда методов анализа естественного языка, позволяющих в автоматизированном режиме обрабатывать неструктурированные тексты. Однако существующие модели информационного поиска обладают рядом недостатков: традиционные модели отличаются низкой эффективностью поиска, сложностью формулировки запроса, новые модели – необходимостью создания вручную хранилищ знаний, используемых для поиска.

Таким образом, в настоящее время существует актуальная научная и техническая задача, состоящая в разработке методик, позволяющих автоматизировать анализ представленного документально научного знания. Решение такой задачи позволит повысить эффективность обработки информации при анализе научного знания.

Объектом исследования является корпус документов научной полнотекстовой информации.

Предмет исследования – методы, модели и алгоритмы обработки текстовой информации.

Целью настоящей работы является повышение эффективности аналитической обработки научной информации, представленной в виде распределенных корпусов текстовых документов.

Поставленная цель достигается решением следующих **задач**:

1. Провести системный анализ процесса обработки неструктуриро-

ванной текстовой информации для выявления системных характеристик корпуса документов.

2. Разработать семантическую модель корпуса документов и алгоритм ее построения на основе латентно-семантического анализа, использующий статистические меры оценки веса терминов.

3. Разработать алгоритм уточнения поискового запроса на сгенерированной семантической модели корпуса, использующий поиск в глубину и в ширину и кластерный анализ множества терминов.

4. Модифицировать существующую информационную технологию поиска и анализа документов путем применения разработанных алгоритмов и разделения этапа семантического анализа текста на локальный и глобальный этапы.

5. Апробировать модифицированную информационную технологию обработки информации с использованием вновь разработанной автоматизированной системы.

Методы исследования. Для решения поставленной задачи применялись методы системного анализа, линейной алгебры, кластерного анализа, теории графов, теории множеств, теории информации, теории алгоритмов.

Научная новизна.

1. По результатам теоретико-множественного и теоретико-информационного анализа выделены системные характеристики корпуса документов, позволяющие расширить набор параметров информационного поиска.

2. Модифицирована информационная поисковая технология в части анализа и систематизации распределенного научного знания, позволяющая в процессе интеллектуального анализа неструктурированной текстовой информации генерировать семантические модели корпуса документов.

3. Разработан алгоритм построения трехмерной семантической модели корпуса документов, позволяющей представить его в форме графа для дальнейшей визуализации и анализа с использованием введенной системы количественных оценок свойств корпуса.

4. Разработан алгоритм уточнения поискового запроса, осуществляющий кластерный анализ множества терминов и эмулирующий движение по семантической модели корпуса документов как поиск на графе в глубину и ширину. Предложены критерии останова: достижение заданного уровня энтропии, измеряющей детализацию термина, достижение заданного порога количественных характеристик термина.

Практическая ценность работы. Результаты работы могут применяться для анализа как распределенных, так и централизованных хранилищ данных и использоваться для обработки любых документальных знаний, содержащих персоналии, названия организаций, даты и другие устойчивые выражения.

Реализация результатов работы. Результаты исследования реализованы в госбюджетных научно-исследовательских работах ФГОУ ВПО «Астраханский государственный технический университет» «Теоретический анализ и математическое моделирование информационных систем», «Теоретический анализ и математическое моделирование систем поддержки принятия управленческих решений»; внедрены в учебный процесс Астраханского государственного технического университета и в муниципальном бюджетном учреждении г. Астрахань «Информационно-аналитический центр»

На основе модифицированной информационной технологии разработана автоматизированная система «Информационно-аналитическая система интеллектуального анализа текстовых электронных ресурсов», прошедшая государственную регистрацию.

Личный вклад автора. В работах, выполненных в соавторстве, автору принадлежат формализация задачи, построение моделей, разработка алгоритмов, проектирование и реализация программного обеспечения.

Апробация научных результатов. Основные положения докладывались и обсуждались на конференциях студентов, аспирантов и молодых ученых «Технологии Microsoft в теории и практике программирования» (Нижний Новгород, 2007–2009), XIV–XVI Международных молодежных научных форумах «Ломоносов» (Москва, 2007–2009), V Всероссийской межвузовской конферен ции молодых ученых (Санкт-Петербург, 2008), XXI–XXIII Международных научных конференциях «Математические методы в технике и технологиях» (Саратов, 2008; Псков, 2009), I Международной научно-практической конференции «Эволюция системы научных коммуникаций ассоциации университетов прикаспийских государств» (Астрахань, 2008), Всероссийской конференции студентов, аспирантов и молодых ученых «Технологии Microsoft в теории и практике программирования» (Москва, 2009), V Всероссийской научно-инновационной конференции студентов, аспирантов и молодых ученых (Москва, 2009), 54-ой Научно-практической конференции профессорско-преподавательского состава Астраханского государственного технического университета (Астрахань, 2010), Международной научно-практической конференции «Фундаментальные и прикладные исследования университетов, интеграция в региональный инновационный комплекс» (Астрахань, 2010).

Публикации. Основные положения диссертационной работы отражены в 16 опубликованных научных работах, среди которых 3 статьи в журналах, рекомендованных ВАК, 1 свидетельство о регистрации программы для ЭВМ и 12 публикаций в сборниках международных, всероссийских научных конференций.

Структура и объем работы. Диссертационная работа состоит из введения, четырех глав основного текста, заключения, списка литературы из 96 наименований и 2 приложений. Общий объем работы 107 страниц

машинописного текста, который включает 38 рисунков, 16 таблиц и 39 формул.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, определены цель и задачи исследования.

Первая глава посвящена обзору и анализу результатов исследований по системному анализу и автоматизированной обработке неструктурированной текстовой информации. Рассмотрены существующие типы моделей информационного поиска, типовая архитектура современных информационно-поисковых систем.

Практика показывает слабую востребованность методов семантической обработки текстовой информации, поскольку они опираются на онтологии, тезаурусы или семантические сети, создание которых требует привлечения экспертов. Практическое применение получили методы, использующие статистическую обработку текста и не осуществляющие его семантический анализ.

Для оценки эффективности информационного поиска общепринятыми являются метрики, использующиеся конференцией по оценке систем текстового поиска Text Retrieval Conference (TREC) и Российским семинаром по оценке методов информационного поиска (РОМИП): полнота, точность, аккуратность, ошибка и F-мера. Данные характеристики взяты за основу для оценки достижения цели исследования.

Во второй главе произведен системный анализ неструктурированной текстовой информации, представленной в виде корпуса текстов научного знания, позволивший отделить ряд системных характеристик объекта исследования.

Текстовый корпус рассматривается в работе как система, а термины и документы – как системные признаки корпуса. С помощью теоретико-множественного моделирования текстовый документ представлен в виде $D = < T, W >$, где $T = \{t_i \mid i = 1...m\}$ – множество доминантных терминов документа, $W = \{w_i\}$ – множество весов терминов, показывающих важность термина t_i для документа D. Корпус текстовых документов представлен в виде матрицы C «термин-документ» вида

$$C = \begin{pmatrix} & D_1 & D_2 & \dots & D_n \\ t_1 & w_{11} & w_{12} & \dots & w_{1n} \\ t_2 & w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_m & w_{m1} & w_{m2} & \dots & w_{mn} \end{pmatrix}, \qquad (1)$$

где $t_1 ... t_m$ – доминантные (ключевые) термины всех документов корпуса $D_1 ... D_n$, w_{ij} – веса терминов в соответствующих документах.

Разработана *семантическая модель* корпуса документов:

$$A =< D, T, S^D, S^t, S^{tD} > \qquad (2)$$

где $D = \{D_j \mid j = 1...n\}$ – множество документов корпуса; $T = \{t_i\}$ – множество терминов корпуса; $S^D = (s^D_{jl})$ $(l = 1,...,n)$ – матрица, в которой элемент s^D_{jl} отражает меру сходства между документами D_j и D_l; $S^t = (s^t_{ki})$ $(k = 1,...,m)$ – матрица, в которой элемент s^t_{ki} отражает меру сходства между терминами t_k и t_i; $S^{tD} = (s^{tD}_{ij})$ – матрица, в которой элемент s^{tD}_{ij} отражает меру сходства между термином t_i и документом D_j.

Кортеж A позволяет представить корпус в виде взвешенного *графа* $G =< X, R >$, где $X =< D, T >$ – множество вершин графа, состоящее из множества документов корпуса и множества входящих в них терминов, $R =< R^D, R^t, R^{tD} >$ – множество ребер, соединяющих документы и термины между собой и друг с другом, и определена функция $w : R \to \Re$, на множестве ребер принимающая значения в действительных числах (рис. 1).

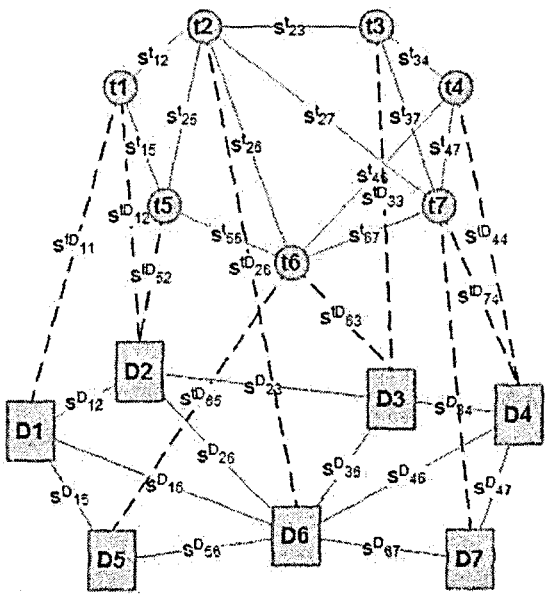

Рис. 1. Представление корпуса в виде графа

Ребра $(D_j, D_l) \in R^D$, $(t_k, t_i) \in R^t$, $(t_i, D_j) \in R^{tD}$ существуют при выполнении условий $s^D_{jl} > \varepsilon_D$, $s^t_{ki} > \varepsilon_t$, $s^{tD}_{ij} > \varepsilon_{tD}$, где ε_D, ε_t и ε_{tD} – заданные пороги.

Графовое представление корпуса документов позволяет выделить *системные характеристики корпуса и его элементов* (табл. 1).

Таблица 1.

Системные характеристики термина, документа и корпуса

Характеристика	**Формула**
Характеристики термина	
Эксцентриситет	$ecc(t_i) = \max\limits_{j=1...m} d(t_i, t_j)$
Степень	$\deg(t_i) = \lvert \{t_j \in T : (t_i, t_j) \in R^t\} \rvert$
Характеристики документа	
Эксцентриситет	$ecc(D_i) = \max\limits_{j=1...m} d(D_i, D_j)$
Степень	$\deg(D_i) = \lvert \{D_j \in D : (D_i, D_j) \in R^D\} \rvert$
Характеристики корпуса	
Радиус словаря	$rad_T(G) = \min\limits_{i=1...m} (ecc(t_i))$
Радиус корпуса	$rad(G) = \min\limits_{i=1...n} (ecc(D_i))$
Диаметр словаря	$diam_T(G) = \max\limits_{i=1...m} (ecc(t_i))$
Диаметр корпуса	$diam(G) = \max\limits_{i=1...n} (ecc(D_i))$
$d(t_i, t_j)$ – расстояние между вершинами t_i и t_j	

Теоретико-информационный анализ корпуса позволяет сформулировать новую системную характеристику, определяющую степень детализации термина t_j, – *информационную энтропию*:

$$H(t_j) = -\sum_{z=1}^{Z} \bar{r}_{jz} \log_2 \bar{r}_{jz}, \qquad (3)$$

где $\bar{r}_{jz} = \dfrac{s^t_{jz}}{\sum\limits_{m=1}^{M^t} s^t_{jm}}$, M_t – количество терминов t_m, для которых существует ребро $(t_j, t_m) \in R^t$, s^t_{jm} – длина этого ребра.

Разработан *алгоритм построения семантической модели корпуса документов* (рис. 2), включающий три этапа обработки информации:

1. Формирование списка терминов корпуса.

1.1. *Лексический, морфологический, синтаксический анализ текста документа.*

1.2. Генерация для каждого документа списка терминов.

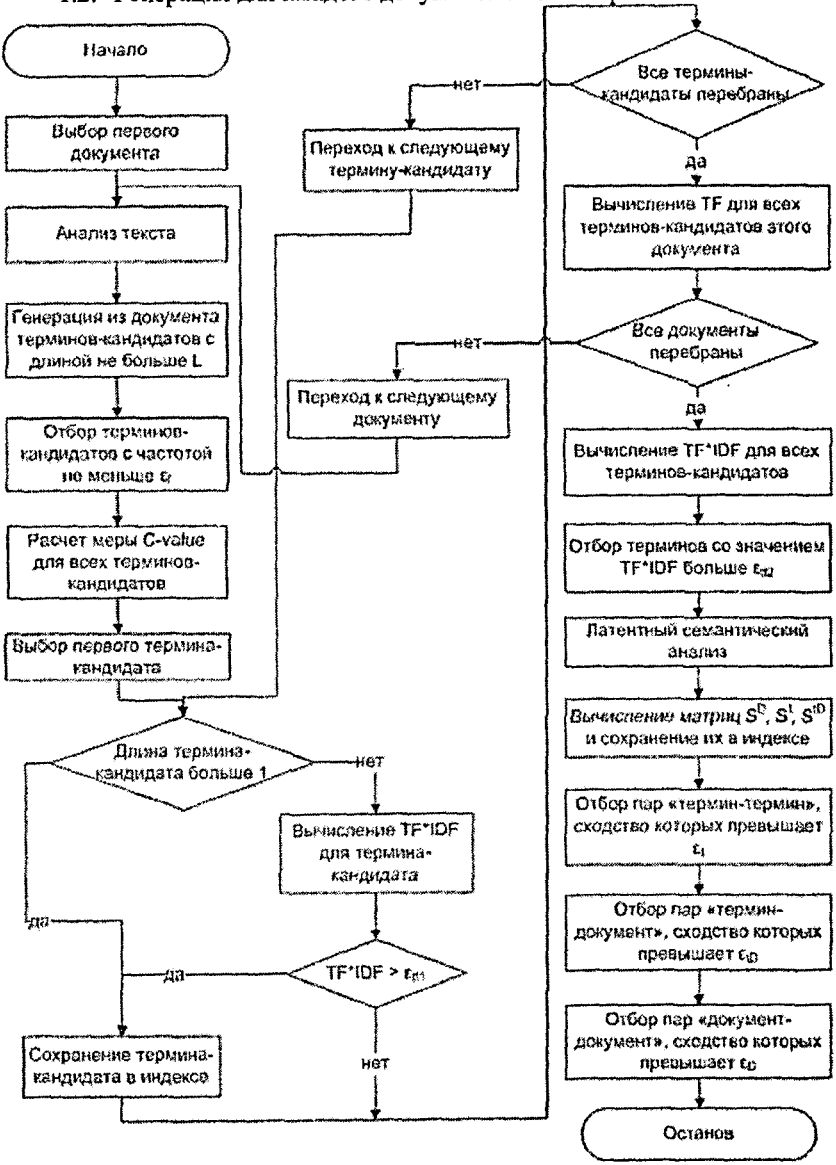

Рис. 2. Алгоритм построения семантической модели корпуса

Из текста извлекаются термины-кандидаты (словосочетания, которые соответствуют заданным грамматическим шаблонам), для которых выполняются условия:

- $1 \le |a| \le L$, где $|a|$ – количество слов, входящих в словосочетание a, L – заданный порог;
- $freq(a) > \varepsilon_f$, где $freq(a)$ – частота употребления в документе термина-кандидата a, ε_f – заданный порог.

Для каждого термина-кандидата вычисляется значение:

$$C-value(a) = \begin{cases} \log_2 |a| * freq(a), \text{если строка а не вложена} \\ \log_2 |a| - \dfrac{1}{P(T_a)} * \sum\limits_{b \in T_a} freq(b), \text{в противном случае} \end{cases}, \quad (4)$$

где T_a – множество терминов-кандидатов, содержащих строку a, $P(T_a)$ – их количество.

2. Определение для каждого документа доминантных терминов.

Осуществляется контрастный тест: с помощью меры TF*IDF терминам, которые часто встречаются в других документах корпуса, присваивается низкий вес, а терминам, которые в данном документе встречаются часто, а в других – редко, высокий вес. В общем случае мера TF*IDF вычисляется по формуле:

$$W(t) = \frac{freq(t)}{|d|} \log \frac{N_D}{N_d}, \quad (5)$$

где $|d|$ – количество слов в документе, N_D – количество документов в анализируемом корпусе, N_d – количество документов в корпусе, содержащих термин t.

В отличие от традиционного подхода в предложенном алгоритме TF*IDF вычисляется дважды. В первый раз данное значение вычисляется на этапе индексации одного документа, что позволяет сразу же сравнить полученное значение с пороговым ε_{tf1} и исключить из множества терминов-кандидатов наиболее употребительные слова, не характерные для какой-либо предметной области. Так как весь корпус еще не обработан, то вместо него рассматривается корпус общей тематики C_V и в формуле (5) пр инимается $N_D = N_V$, $N_d = freq_V(t)$, где N_V – размер корпуса C_V, $freq_V(t)$ – частота употребления термина t в корпусе C_V.

После обработки всего корпуса производится расчет TF*IDF для всех найденных терминов-кандидатов, причем для повышения точности ре-

зультатов алгоритма вместо значения $freq(t)$ используется

$$freq_C(t) = \begin{cases} freq(t), \text{если} \mid t \mid = 1 \\ C - value(t), \text{если} \mid t \mid > 1 \end{cases}.$$

Доминантными для документа терминами считаются те, значение меры TF*IDF для которых превышает заданный порог ε_{tf2}.

3. Для построения семантической модели корпуса применен метод латентного семантического анализа (LSA), который заключается в сингулярном разложении матрицы C

$$\begin{matrix} & C & & & U & & & Z & & & V^T & \\ \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{bmatrix} & = & \begin{bmatrix} \begin{bmatrix} \\ u_1 \\ \\ \end{bmatrix} & \cdots & \begin{bmatrix} \\ u_l \\ \\ \end{bmatrix} \end{bmatrix} & \cdot & \begin{bmatrix} z_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & z_l \end{bmatrix} & \cdot & \begin{bmatrix} \begin{bmatrix} & v_1 & \end{bmatrix} \\ \vdots \\ \begin{bmatrix} & v_l & \end{bmatrix} \end{bmatrix} \end{matrix} \qquad (6)$$

где $z_1,...,z_l$ – сингулярные числа, $u_1,...,u_l$ и $v_1,...,v_l$ – левый и правый сингулярные вектора, и аппроксимации ее матрицей $C_k = U_k Z_k V_k^T$ меньшего ранга k. Сходство между двумя терминами определяется по какой-либо метрике сходства (например, косинус угла) между соответствующими векторами, представляющими собой строки матрицы $U_k Z_k$. Аналогично, сходство между двумя документами определяется с помощью матрицы $V_k Z_k$, а сходство между терминами и документами – с помощью матрицы $U_k \sqrt{Z_k}$ и $\sqrt{Z_k} V_k$.

Ребра между соответствующими вершинами графа корпуса определяются путем отбора пар, значение сходства для которых превышает заданные пороговые значения ε_D, ε_t и ε_{tD}.

Разработанная семантическая модель позволяет применить к анализу корпуса документов алгоритмы поиска на графе, а также расширить набор параметров информационного поиска выделенными системными характеристиками.

В **третьей главе** описан разработанный алгоритм уточнения запроса пользователя. Описана модифицированная информационная поисковая технология.

Сформулирована задача информационного поиска на вновь разработанной семантической модели:

Для заданного запроса $T^q = \{t_1^q,...,t_{m^q}^q\}$ необходимо построить подграф $G^q =< X^q, R^q >$ графа G, где $X^q =< D^q, T^q >$ – множество вершин, а R^q – множество ребер, соединяющих документы и термины между собой и друг с другом, причем $D^q = \{D_i \in D \mid \exists t_j^q \in T^q : (t_j^q, D_i) \in R^{tD}\}$.

Решена задача уточнения поискового запроса путем добавления к нему новых терминов, семантически связанных с терминами $t_1^q,...,t_{m^q}^q$.

Алгоритм уточнения запроса пользователя – человеко-машинная интерактивная процедура, в процессе которой существующий запрос дополняется новыми терминами по мере вербализации информационной потребности пользователя (рис. 3).

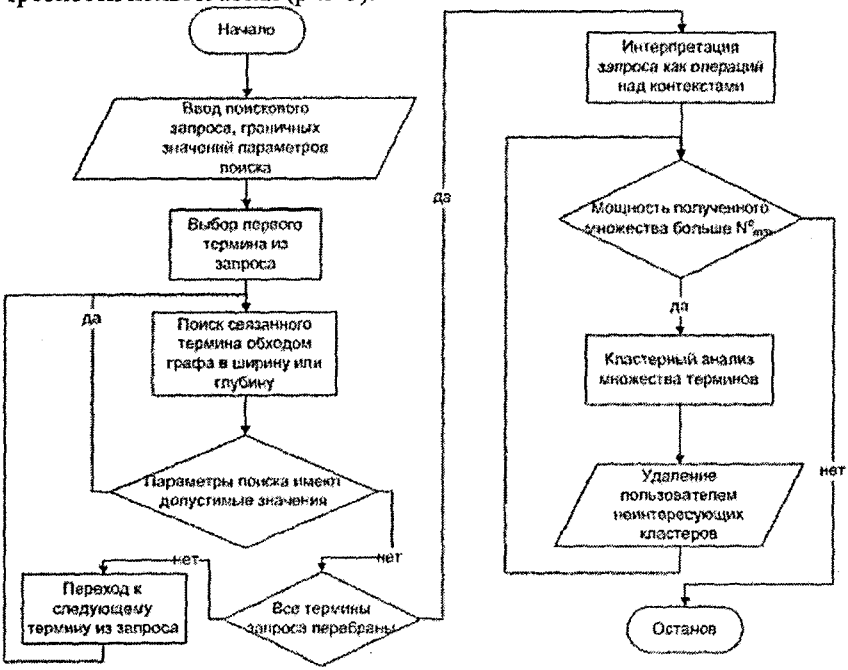

Рис. 3. Алгоритм уточнения поискового запроса

Для уточнения поискового запроса введем понятия:

m -ый уровень детализации термина t – множество T_m^t, такое что:

$$T_m^t = \begin{cases} \{t_i \in T \mid (t_i,t) \in R^t\}, m=1 \\ \{t_i \in T \mid t_i \notin T_k^t, k=1,...,m-1, \exists t_j \in T_{m-1}^t : (t_i,t_j) \in R^t\}, m>1 \end{cases} \quad (7)$$

Контекст термина t – множество $E = T_1^t \cup ... \cup T_m^t \cup ... \cup T_{n^E}^t$, состоящее из n^E уровней.

Параметры алгоритма уточнения поискового запроса приведены в табл. 2.

Таблица 2.

Параметры алгоритма уточнения запроса

Название	**Граничное значение**
Глубина поиска – количество уровней в контексте термина	$L_{\max}$
Количество терминов в контексте	$N_{\max}$
Количество терминов в запросе	$N^q_{\max}$
Степень вершины	$\deg_{\min}$
Эксцентриситет вершины	$ecc_{\max}$
Энтропия	$H_{\min}$

Вариант алгоритма, использующий *поиск в ширину*, сводит участие пользователя в процессе построения расширенного запроса к минимуму, реализуя автоматическое построение контекста. При этом контексты терминов могут пересекаться. Вариант алгоритма, использующий *поиск в глубину*, требует участия пользователя уже на первом этапе, однако позволяет точнее подобрать множество терминов. При этом контексты терминов пересекаться не могут.

Если поисковый запрос задан в виде

$$q = \vee_i(\theta_i), \theta_i \in \{t_i^q, \neg t_i^q\}, \tag{8}$$

то может быть использован как поиск в глубину, так и поиск в ширину, а если поисковый запрос задан в виде

$$q = \wedge(\vee_i(\theta_i)), \theta_i \in \{t_i^q, \neg t_i^q\}, \tag{9}$$

то используется только поиск в ширину, т.к. при поиске в глубину отсутствие пересечений между контекстами дало бы в результате пустое множество.

После построения множества контекстов $E_1,...,E_{m^q}$, соответствующих терминам запроса, запрос интерпретируется как операции над множествами $\{t_1^q\} \cup E_1,...,\{t_{m^q}^q\} \cup E_{m^q}$ по следующим правилам:

$$t_i^q \vee t_j^q \to \{t_i^q, t_j^q\} \cup E_i \cup E_j \tag{10}$$

$$t_i^q \wedge t_j^q \to \{t_i^q, t_j^q\} \cap E_i \cap E_j \tag{11}$$

$$t_i^q \vee \neg t_j^q \to \{t_i^q\} \cup (E_i \setminus E_j) \tag{12}$$

$$t_i^q \wedge \neg t_j^q \to \{t_i^q\} \cup (E_i \setminus E_j) \tag{13}$$

Если мощность получившегося в результате множества терминов превышает $N^q_{\max}$, то производится кластерный анализ данного множества и пользователь удаляет не интересующие его кластеры.

Для полученного в результате выполнения алгоритма множества терминов решается классическая задача информационного поиска.

В целях экономии вычислительных ресурсов при обработке корпуса по описанным выше принципам предлагается использовать распределенную обработку данных и проводить индексацию документов непосредственно на тех веб-серверах, на которых они находятся.

При обработке распределенного корпуса документов используется набор текстовых корпусов $D^1, D^2, ..., D^n$, хранящихся на n веб-серверах, причем каждый корпус представлен в виде графа $G^k =< X^k, R^k >$, где $X^k =< D^k, T^k >$, $R =< R^{Dk}, R^{tk}, R^{tDk} >$. Граф распределенного корпуса имеет вид $G =< X, R >$, где $X =< D, T >$, $D = D^1 \cup D^2 \cup ... \cup D^n$, $T = T^1 \cup T^2 \cup ... \cup T^n$, а ребра графа определяются следующим образом:

При заданных новых пороговых значений ε_D, ε_t и ε_{tD} для пары (D_i, D_j) рассчитывается

$$s^D_{ij} = \begin{cases} s^{Dk}_{pq}, \text{если } \exists k : D_i \in D^k \text{ и } D_j \in D^k \\ -1, \text{в противном случае} \end{cases}, \qquad (14)$$

где p – номер документа D_i в корпусе D^k, q – номер документа D_j в корпусе D^k. При $s^D_{ij} > \varepsilon_D$ ребро $(D_i, D_j) \in R^D$ добавляется к графу G.

Для пары (t_i, t_j) рассчитывается

$$s^t_{ij} = a^n_{ij}, \qquad (15)$$

где $a^1_{ij} = s^{t1}_{xy}$, $a^k_{ij} = \begin{cases} \dfrac{a^{k-1}_{ij} + s^{tk}_{xy}}{2}, \text{если } t_i \in T^k \text{ и } t_j \in T^k \\ a^{k-1}_{ij}, \text{в противном случае} \end{cases}$ при $k = 2,...,n$, x – номер термина t_i в наборе терминов T^k, y – номер термина t_j в наборе терминов T^k. При $s^t_{ij} > \varepsilon_t$ ребро $(t_i, t_j) \in R^t$ добавляется к графу G.

Для пары (t_i, D_j) рассчитывается

$$s^t_{ij} = \begin{cases} s^{tk}_{xq}, \text{если } \exists k : t_i \in T^k \text{ и } D_j \in D^k \\ -1, \text{в противном случае} \end{cases} \qquad (16)$$

и при $s^{tD}_{ij} > \varepsilon_{tD}$ ребро $(t_i, D_j) \in R^{tD}$ добавляется к графу G.

Распределенная обработка данных и разработанные алгоритмы были использованы для модификации существующей информационной поисковой технологии (рис. 4) путем разделения этапа семантического анализа текста на локальный и глобальный этапы и применения алгоритмов: построения семантической модели – для семантического анализа, расширения поискового запроса – на этапе анализа запроса.

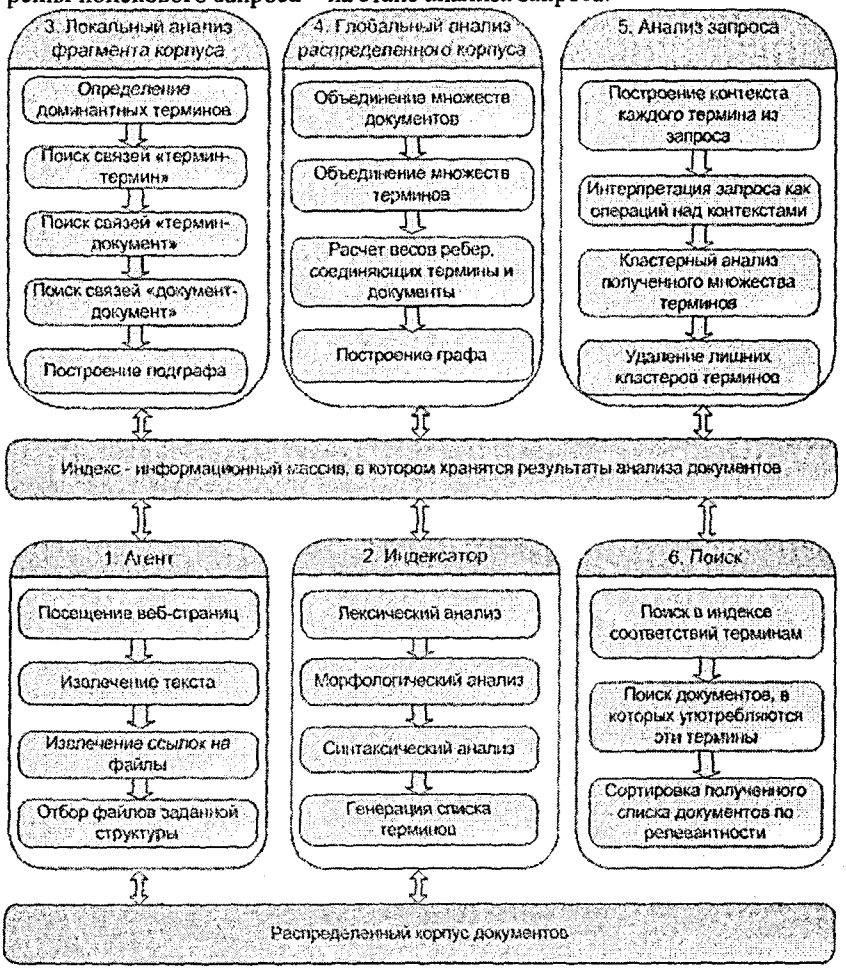

Рис. 4. Модифицированная информационная технология

В **четвертой главе** рассмотрена архитектура разработанной автоматизированной системы и описан эксперимент по сравнению разработанной системы с одним из популярных программных аналогов.

Эксперимент показал, что по ряду параметров разработанная автоматизированная система опережает «Персональный поиск Яндекса», а по остальным параметрам несущественно отстает от него (табл. 3).

Таблица 3.

Сравнение разработанной системы с системой «Персональный поиск Яндекса»

	Полнота	Точность	Аккуратность	Ошибка	F-мера
Разработанная система	0,97	0,76	0,99	0,008	0,81
«Персональный поиск Яндекса»	0,35	0,82	0,91	0,09	0,33

В **приложениях** приведены свидетельство о государственной регистрации программы для ЭВМ и акты о внедрении результатов научной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. На основе теоретико-множественного и теоретико-информационного анализа выявлены характеристики корпуса документов, используемые в задачах информационного поиска.

2. Разработана трехмерная семантическая модель корпуса документов, позволяющая структурировать содержащиеся в нем научные знания в виде графа для последующего анализа. На основе латентно-семантического анализа разработан алгоритм построения семантической модели.

3. Сформирован алгоритм расширения поискового запроса, уточняющий информационные потребности пользователя с использованием пороговых значений количественных характеристик терминов и их энтропии.

4. Модифицирована информационная технология поиска и анализа данных в части систематизации распределенного научного знания, позволяющая в процессе интеллектуального анализа неструктурированной текстовой информации генерировать семантические модели корпуса документов.

5. На основе модифицированной информационной технологии разработано программное обеспечение, апробация которого продемонстрировала повышение характеристик информационного поиска: полноты – на 62%, аккуратности – на 8% по сравнению с известными программными аналогами.

6. Результаты работы внедрены в муниципальном бюджетном учреждении г. Астрахань «Информационно-аналитический центр» и использованы при выполнении госбюджетных научно-исследовательских работ

Астраханского государственного технического университета. Учебный вариант программного обеспечения используется в Астраханском государственном техническом университете.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в периодических изданиях, включенных в список ВАК РФ

1. Седова, Я. А. Интеллектуальный анализ корпуса документов научной информации / Я. А. Седова, И. Ю. Квятковская // Вестник Астраханского государственного технического университета. Серия «Управление, вычислительная техника и информатика». – 2011. – №1. – С. 128–136.

2. Седова, Я. А. Системный анализ корпуса текстов научного знания / Я. А. Седова, И. Ю. Квятковская // Вестник Саратовского государственного технического университета. – 2011. – №4 (50). Выпуск 2. – С. 197–206.

3. Седова, Я. А. Применение стохастических фракталов к некоторым задачам информационного поиска // Научно-технический вестник Санкт-Петербургского государственного университета информационных технологий, механики и оптики. Выпуск 46. – 2008. – С. 19–22.

Статьи в межвузовских научных сборниках, сборниках трудов международных, всероссийских конференций

4. Седова, Я. А. Анализ несловарных слов русского языка как элемент семантического анализа текста // Вестник Астраханского государственного технического университета. – 2007. – №2(37). – С. 170–172.

5. Седова, Я. А. Принципы и методы построения словаря русского языка для алгоритмов морфемного и морфологического анализа текста на естественном языке // Материалы докладов XIV Международной конференции студентов, аспирантов и молодых ученых «Ломоносов» [Электронный ресурс] – М. : Издател. центр Факультета журналистики МГУ им. М.В. Ломоносова, 2007. – 1 электрон. опт. диск (CD-ROM); 12 см. – Систем. требования: ПК с процессором 486 +; Windows 95; дисковод CD-ROM; Adobe Acrobat Reader.

6. Седова, Я. А. Разбор текста на русском языке на основе усовершенствованного алгоритма стемминга Портера // Инновационные технологии в управлении, образовании, промышленности «Астинтех–2007»: матер. Всерос. конф. 18–20 апреля 2007 г. в 2 ч. / Сост. И. Ю. Петрова. – Астрахань: Издательский дом «Астраханский университет», 2007. – Ч. 2. – С. 139–141.

7. Седова, Я. А. Построение поисковых роботов в рамках системы фрактального анализа Web-пространства // Технологии Microsoft в теории и практике программирования. Материалы конференции / Под ред. проф. Р. Г. Стронгина. – Нижний Новгород: Изд-во Нижегородского госуниверситета, 2008. – С. 310–313.

8. Седова, Я. А. Интеллектуальная система кластерного анализа электронных текстовых ресурсов // Материалы докладов XV Междуна-

родной конференции студентов, аспирантов и молодых ученых «Ломоносов» / Отв. ред. И. А. Алешковский, П. Н. Костылев, А. И. Андреев. [Электронный ресурс] – М.: Изд–во МГУ; СП МЫСЛЬ, 2008. – 1 электрон. опт. диск (CD-ROM); 12 см. – Систем. требования: ПК с процессором 486 +; Windows 95; дисковод CD-ROM; Adobe Acrobat Reader.

9. Седова, Я. А. Применение фрактального подхода к некоторым задачам информационного поиска / Я. А. Седова, И. Ю. Квятковская // Математические методы в технике и технологиях (ММТТ–21): сб. трудов XXI Междунар. науч. конф.: в 10 т. / Под общ. ред. В. С. Балакирева. – Саратов: Сарат. гос. техн. ун-т, 2008. – Т. 8. – Секция 8. – С. 220–221.

10. Седова, Я. А. Система эффективного поиска в объединенном информационном пространстве Ассоциации университетов Прикаспийских государств // Сб. матер. Междунар. науч.–практ. конф. «Эволюция системы научных коммуникаций Ассоциации университетов Прикаспийских государств». – Астрахань: РГНФ, АГТУ, 2008. – С. 22–24.

11. Седова, Я. А. Автоматизация сбора данных для построения онтологий // Технологии Microsoft в теории и практике программирования. Материалы конференции / Под ред. проф. В. П. Гергеля. – Нижний Новгород: Изд-во Нижегородского госуниверситета, 2009. – С. 396–400.

12. Седова, Я. А. Автоматизация анализа данных для построения онтологий // Технологии Microsoft в теории и практике программирования: тр. VI Всерос. конф. студентов, аспирантов и молодых ученых. Центральный регион. Москва, 1–2 апреля 2009 г. – М. : Вузовская книга, 2009. – С. 99–100.

13. Седова, Я. А. LSPL-шаблоны для решения задачи автоматизированного построения онтологий // Материалы докладов XVI Международной конференции студентов, аспирантов и молодых ученых «Ломоносов» / Отв. ред. И. А. Алешковский, П. Н. Костылев, А. И. Андреев. [Электронный ресурс] – М. : МАКС Пресс, 2009. – 1 электрон. опт. диск (CD-ROM); 12 см. – Систем. требования: ПК с процессором 486 +; Windows 95; дисковод CD-ROM; Adobe Acrobat Reader.

14. Седова, Я. А. Автоматизация проектирования предметных онтологий с использованием интеллектуальных агентов // Сборник трудов конференции молодых ученых. Выпуск 6. Информационные технологии / Главный редактор д.т.н., проф. В. Л. Ткалич. – СПб: СПбГУ ИТМО, 2009. – С. 429–432.

15. Седова, Я. А. Архитектура автоматизированной системы построения предметных онтологий / Я. А. Седова, И. Ю. Квятковская // Математические методы в технике и технологиях (ММТТ–22): сб. тр. XXII Междунар. науч. конф.: в 10 т. / Под общ. ред. В. С. Балакирева. – Псков: Псков. гос. политехн. ин-т, 2009. – Т. 7. – Секция 8. – С. 134–135.

Свидетельство о государственной регистрации программы для ЭВМ

16. Автоматизированная система «Информационно-аналитическая система интеллектуального анализа текстовых электронных ресурсов. Св. о гос. рег. прогр. для ЭВМ №2009610640. / Квятковская И. Ю., Седова Я. А., Филандыш Н. И. Зарег. 28.01.2009.

Подписано в печать 11.04.11г. Формат 60x90/16. Гарнитура Times New Roman.
Усл. печ. л. 1,0. Тираж 100 экз. Заказ № 159

Отпечатано в типографии издательства ФГОУ ВПО «АГТУ».
414025, Астрахань, Татищева, 16.