На правах рукописи

ГРИГОРЬЕВ АЛЕКСАНДР СЕРГЕЕВИЧ

РАЗРАБОТКА МЕТОДА И СОЗДАНИЕ СИСТЕМЫ ПОЛНОТЕКСТОВОГО ПОИСКА НА ОСНОВЕ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ОГРАНИЧЕННОГО КОНТЕКСТА СЛОВА

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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

Москва – 2006

Работа выполнена в Московском государственном техническом университете им. Н.Э. Баумана.

Научный руководитель: доктор технических наук, профессор
Б.Г. Трусов

Официальные оппоненты: доктор технических наук, профессор
Я.Л. Шрайберг
кандидат физико-математических наук
В.И. Шабанов

Ведущая организация: НТЦ «Информрегистр»

диссе ня 2006 года в 14:30 на заседании
ниче Московском государственном тех-
манс по адресу 107005, Москва, 2-я Бау-

овского госу-

ечатью, про-

занов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Постоянное увеличение объемов хранилищ информации повышает роль поисковых средств, используемых для определения интересующей пользователя части документов. Поиск и предоставление документов читателям чаще всего ограничены описаниями документов, как правило, включающими название издания и его авторов, которые недостаточно полно отражают содержание документа. Для более полного описания документа он снабжается списком ключевых слов и словосочетаний, которые отражают тематику документа. Выделение таких слов и словосочетаний обычно выполняется авторами или библиотекарями. С увеличением количества документов, хранящихся в информационных фондах, задача построения списков ключевых слов и словосочетаний усложняется. Поэтому в настоящее время существует потребность в создании поисковой системы, автоматически выявляющей в текстах документов ключевые слова и словосочетания для обеспечения высокой степени релевантности результатов поиска сформулированным пользователем запросам. Запрос пользователя обычно формулируется в виде строки, содержащей перечисление слов. По аналогии с выделением слов и словосочетаний в текстах документов ключевые слова и словосочетания должны выделяться и в запросах пользователя, сформулированных на естественном языке. Под естественным языком понимается язык, который используется пользователем при общении с другими людьми.

Многие существующие системы определяют ключевые слова и словосочетания по заранее подготовленным правилам обработки текста документов и запросов. Другие ограничиваются выделением ключевых слов, не учитывая взаимосвязи между ними. Актуальность данной работы следует из важности создания эффективного метода поиска информации в фондах электронных документов с учетом языка изложения и тематики каждого документа обрабатываемого фонда.

Целью диссертационной работы является создание метода, направленного на повышение качества полнотекстового поиска путем выделения повторяющихся сочетаний слов как в тексте анализируемых документов, так и в тексте поисковых запросов, сформулированных на естественном языке.

Для достижения этой цели в диссертации решены следующие **задачи**:

- систематизированы известные методы и стратегии поиска, выделены основные этапы обработки текстов на естественном языке;
- разработаны и оптимизированы структуры для хранения служебной информации, создаваемой в процессе статистического анализа текстов;
- разработан метод поиска по тексту произвольных документов на естественном языке, использующий повторяющиеся сочетания слов, автоматически выделяемые как в текстах документов, так и в тексте запросов;
- использовано группирование документов по спискам повторяющихся сочетаний слов с целью ускорения поиска;
- разработан метод автоматического обучения анализатора текста языку по динамически пополняемому информационному фонду документов за счет выявления закономерностей при статистическом анализе ассоциативных связей между словами в текстах документов;

- создан программный комплекс, реализующий разработанный метод поиска.

Методы исследования. При решении поставленных задач в данной работе использованы элементы комбинаторного анализа, теоретико-множественный аппарат, статистические методы и теория автоматической классификации. При разработке программного обеспечения применялись методы объектно-ориентированного и функционального программирования с использованием сред разработки Microsoft Visual Studio.NET и СУБД Microsoft SQL Server.

Научная новизна.

В результате выполненной работы получены новые научные результаты:

- разработан новый метод поиска информации в фондах электронных документов, основанный на статистической обработке текста, сопоставлении сочетаний слов документов и запроса, и позволяющий автоматически обучать поисковую машину естественному языку путем извлечения закономерностей статистической сочетаемости слов в текстах документов;
- разработан метод и алгоритм группирования слов на основе расширенного метода вероятностного морфологического поиска, учитывающий вхождение родственных слов в устойчивые сочетания;
- предложен подход, позволяющий устанавливать контекстно-зависимую синонимию слов в обрабатываемых текстах.

Основные положения, выносимые на защиту:

- метод статистической обработки ограниченного контекста слова;
- метод автоматического обучения анализатора текста языку по набору обучающих текстов;
- метод получения и сравнения структур, использующих сочетания слов для описания текста документов и поискового запроса;
- программное обеспечение, реализующее разработанные методы обработки текста и метод поиска текстовой информации.

Практическая значимость. Основным практическим результатом работы является создание метода автоматического обучения языку путем выявления закономерностей статистической сочетаемости слов проанализированных текстов, реализованного в виде программного комплекса, позволяющего выполнять поиск по запросам на естественном языке среди полнотекстовых электронных документов в фондах различных электронных библиотек. Это позволяет рекомендовать разработанное программное обеспечение в качестве поисковой подсистемы в программные комплексы, хранящие и обрабатывающие фонды электронных документов.

Апробация и внедрение результатов работы. Полученный в результате выполненной работы программный комплекс внедрен и используется в рамках единой Автоматизированной Библиотечной Информационной Системы МГТУ им. Н.Э. Баумана. Разработанные методы, алгоритмы и модели использованы при создании системы поддержки фонда электронных документов. Основные результаты работы докладывались и обсуждались на 6, 7, 8-м всероссийских научно-практических семинарах «Новые информационные технологии», Москва, 2003 – 2005.

Публикации по теме диссертации. По теме диссертации опубликовано 6 печатных работ.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав и списка литературы из 113 наименований. Общий объем работы составляет 158 страниц, включая 39 рисунков и 30 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** изложена проблема поиска текстовой информации в фондах электронных документов, обоснована ее актуальность, сформулирована цель исследования и необходимость разработки метода поиска, основанного на статистической обработке контекста слова, и создания программного обеспечения, реализующего данный метод.

В **первой главе** дана математическая постановка задачи, проведен анализ литературных источников, приведен обзор стратегий текстового поиска и детально рассмотрены методы обработки текстов. Рассмотренная в работе модель естественного языка $L=L(W, C)$ определяется множеством слов W и набором связей C, объединяющих слова в осмысленные обороты и предложения. Множество текстов T на естественном языке L определено как $T=T(L)$. Для каждого текста $t \in T$ строится структура M_t, описывающая слова этого текста $w_t \in W_t$ и взаимосвязи этих слов $c_t \in C_t$, причем $W_t \subseteq W$ и $C_t \subseteq C$.

Если пересечение языка запроса и языка текста пусто, множество результатов поиска пусто, поэтому множество запросов Q, адресуемых поисковой системе, также формулируются на языке L, то есть $Q=Q(L)$. Каждый запрос $q \in Q$ преобразуется в структуру M_q, описывающую связи $c_q \in C_q$ между словами $w_q \in W_q$, причем $W_q \subseteq W$ и $C_q \subseteq C$. Построенные структуры, описывающие множество документов и запросов формируют над языком специфичный для каждого метода поиска набор связей C между словами W языка L.

Задача поиска по запросу q состоит в том, чтобы построить структуры M_q и M_t и из множества текстов T выбрать такое подмножество $T_R \subseteq T$, что элементы структуры M_q входят в структуру M_t для каждого текста $t \in T_R$. Найденные таким образом документы t оцениваются по величине скалярного значения функции релевантности $R(q, t)$ поисковому запросу q. Эта функция имеет тем большее значение, чем больше элементов структур, описывающих документы и запрос, совпадают. Известные алгоритмы полнотекстового поиска различаются способами построения структур M_q и M_t, а также видом функции релевантности R.

Выполненная в работе классификация методов поиска текстовой информации делит их на булевские, формально-грамматические, статистические, алгебраические и лингвистические. Булевские методы учитывают при сравнении запросов с текстом документов слова, как независимые объекты. Документ признается соответствующим запросу, если слово запроса найдено среди слов документа. Практически данный метод реализован в алгоритмах прямого поиска (Д. Кнут, Н. Вирт, Р. Седжвик) и алгоритмах, использующих инвертированный список (Г.П. Лун, С. Брин, Л. Пейдж, И. Сегалович). В формально-грамматических методах для определения связей слов в предложениях составляются формальные пра-

вила обработки предложений языка аналогично правилам грамматик, описывающим формальные языки. Практически данный подход реализован в методе, использующем структурную схему предложения (В.А. Крищенко) и вероятностно-грамматическом методе (А.В. Брик). Статистический подход подразумевает использование функции оценки связности слов. Формально-грамматическая модель может быть дополнена использованием функции вероятностной оценки связности слов для учета ассоциативных связей между словами (В.И. Шабанов). Другой метод статистического ранжирования текстов по степени соответствия запросу основан на определении «различительной силы» слов (Т. Джойс, Р. Нидхэм, Дж. Солтон). Алгебраические методы включают в себя алгоритмы искусственного интеллекта. В частности, имитационный подход (Ф.С. Файн), адаптивное распознавание образов, построение нейронных сетей (Е. Шикута) позволяют, не акцентируя внимание на формируемых правилах поиска, моделировать обучение человека языку по принципу «черного ящика». На стыке алгебраических и статистических находится вероятностный метод обучения языку (Л.А. Растригин), получивший распространение в генетических алгоритмах. Лингвистический подход занимает особое место среди методов обработки текстов (Э.В. Попов, Н.Н. Леонтьева) в связи со сложностью применяемого в нем описания языка. Реализующие данный подход модели пока что используются только в ограниченных предметных областях а, зачастую, и не доводятся до практической реализации, как, например, уникальная модель «Смысл-Текст» (И.А. Мельчук, А. Жолковский).

На основании анализа методов и рассмотренной классификации сформулированы требования к методу анализа текста, учитывающие преимущества различных подходов к проведению поиска:

- рассматривать все слова как равноценные лексические единицы без привязки к конкретному естественному языку;
- группировать имеющие схожее написание слова для обобщения статистических данных об употреблении различных форм слова;
- извлекать взаимосвязи между словами, используя окружающие слова в областях текста, несущих самостоятельную завершенную констатацию;
- определять уровни подчинения слов в сочетаниях слов, исходя из их статистической сочетаемости друг с другом, чтобы при сравнении текстов документов и запросов учесть подчинительные отношения между составляющими их словами;
- при проведении автоматического обучения языкам обеспечивать соответствие набора правил обработки текста содержимому информационного фонда путем обновления правил при пополнении фонда новыми документами по результатам обработки этих документов.

Приведенные требования в наибольшей степени удовлетворяются при использовании статистического подхода, направленного на сбор сведений о взаимосвязях слов в текстах. Сделан вывод, что чисто статистический подход должен быть расширен и ориентирован на подготовку аналогов лингвистических словарей: морфологического, синтаксического и семантического.

Во **второй главе** предложен подход к поиску текстовой информации по запросам на естественном языке в фонде текстовых документов. Описан метод

статистической обработки текста документов для оценки схожести слов по написанию (аналог морфологического анализа), определения уровней подчинения слов в устойчивых сочетаниях (аналог синтаксического анализа) и поиска слов, встречающихся в одинаковом контексте (аналог семантического словаря).

В основе метода лежит поиск в текстах устойчивых сочетаний слов при обработке смысловых фрагментов – областей текста, несущих самостоятельную завершенную констатацию. Сочетанием называется набор слов смыслового фрагмента, причем не предъявляется требований к порядку слов в нем. Устойчивыми называются сочетания, повторяющиеся в различных смысловых фрагментах текстов всего фонда. Существенно, что это не словосочетания, а устойчивые наборы слов. При поиске таких сочетаний предварительные правила сочетаемости слов языка не используются.

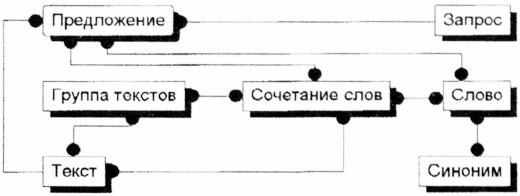

Рис. 1. Логическая диаграмма сущностей структуры хранения данных

Для сохранения обрабатываемых текстов, предложений, слов и служебной информации, создаваемой в процессе статистической обработки текстов и запросов, разработана специальная структура хранения данных (рис. 1).

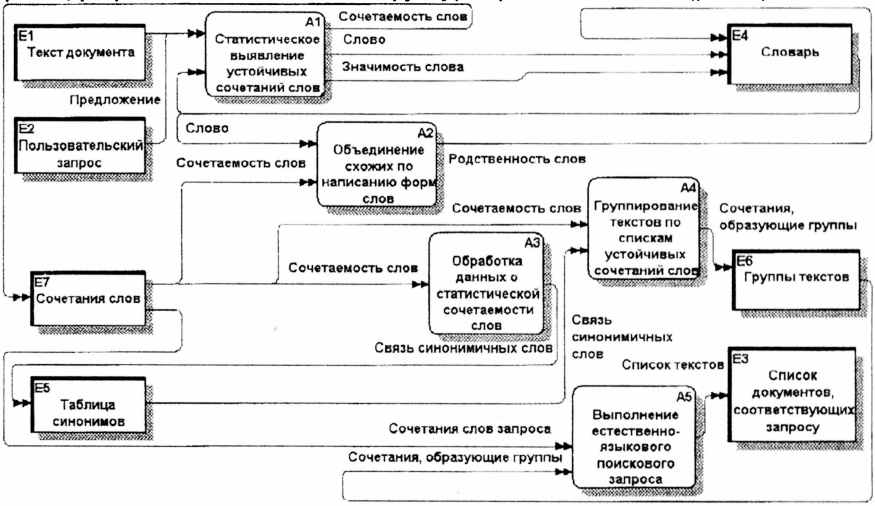

Рис. 2. Общая функциональная схема системы обработки текста и поиска

Проанализированные тексты и запросы представляются в виде списков

предложений (рис. 2). Предложения образуют списки слов. Слова заполняют словарь системы. Слова предложений используются для составления всех возможных сочетаний этих слов. Отобранные из них устойчивые сочетания слов позволяют определить степень сочетаемости слов в предложениях и обнаружить контекстно-зависимые синонимы. Контекстно-зависимыми синонимами признаются главные слова в устойчивых сочетаниях, отличающихся в одном слове. Главным называется слово, которое при анализе текстов чаще других слов появляется в этом сочетании и реже других слов появляется в других сочетаниях.

Статистическая обработка текста

При выполнении статистического анализа добавляемых в фонд документов все слова обработанных текстов заносятся в словарь с сохранением количества их повторений. При этом новым словом считается любая уникальная комбинация алфавитно-цифровых символов, исключая символы разделителей слов и знаки препинания. Все слова ранжируются по степени значимости. Для количественной оценки значимости слова w введена функция $S^*(w)$:

$$S^*(w) = F_{\max} * \log_b \frac{N_T}{N}, \qquad (1)$$

где w - слово в словаре; $F_{\max}$ - наибольшее значение частоты повторения слова w (то есть отношение числа появлений слова w в документе к количеству слов в этом документе) среди всех документов фонда T; N_T - число документов в фонде T; N - число документов фонда, в которых встретилось слово w; b – настроечный коэффициент, позволяющий регулировать влияние распространенности слова в текстах фонда на значение функции значимости.

Практика показала, что частота появления слова в документе находится в интервале [0; 0,1]. Доля документов, в которых встречаются те или иные слова, лежит в интервале (0; 1] (рис. 3).

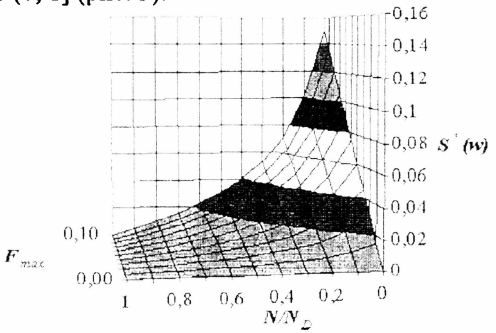

Рис. 3. График функции значимости слова

С ростом количества повторений слова в фонде рост значимости слов, встречающихся в малом количестве текстов, намного интенсивнее роста значимости слов, встречающихся в большинстве текстов фонда.

В отличие от существующих методов определения значимости слов введенная функция позволяет оценить значимость для фонда в целом, а не для от-

дельных документов.

Слова, имеющие высокую значимость, используются для составления сочетаний слов. Выделенные из них устойчивые сочетания заносятся в таблицу устойчивых сочетаний слов (M). Для всех слов w устойчивых сочетаний m вычисляется степень их статистической сочетаемости $A(w, m)$, используя функцию сопряженности Чупрова.

$$A(w,m) = \frac{2p_{wm}}{2p_{wm} + p_w + p_m}, \qquad (2)$$

где p_{wm} - частота совместного использования слова w и сочетания m (отношение количества сочетаний m к общему количеству всех сочетаний); p_w - частота использования слова w в сочетаниях, отличных от m (отношение количества появлений этих сочетаний в текстах к общему количеству всех сочетаний); p_m - частота использования в тексте сочетания m (отношение количества повторений этого сочетания в текстах к общему количеству всех сочетаний).

Функция статистической сочетаемости $A(w, m)$ слов использована для построения иерархии подчинения слов в содержащих их сочетаниях. Это позволяет различать контексты слов, используемых в разных предметных областях и различных языках.

Объединение схожих по написанию слов

Аналогично статьям лингвистических словарей различные грамматические формы слова, имеющие схожее написание, объединяются в логическую единицу. Эти слова называются родственными, а слово, используемое в дальнейшем вместо всех родственных слов, называется заглавным. Чтобы дать численную оценку схожести пары слов, вводится понятие расстояния редактирования, которое равно минимальному числу операций вставки, замены, удаления или транспозиции символов, необходимых для преобразования одной строки в другую. В основу разработанного метода объединения схожих форм слов положен механизм вероятностного поиска, практически реализованный как метод расширения выборки.

Учет вхождения родственных слов в устойчивые сочетания, отличающиеся одним словом, существенно повышает качество традиционного вероятностного морфологического поиска.

Предложенный алгоритм поиска родственных слов отличается от традиционного, формально-грамматического алгоритма объединения грамматических форм слова. Он обнаруживает не все возможные модификации исходного заглавного слова, а только наиболее близкие к нему по написанию. При этом традиционная группа грамматических форм слова представляется несколькими полученными группами. Поэтому создаваемый словарь содержит больше заглавных слов, чем соответствующий такому же тексту лингвистический словарь.

Определение схожести текстов по спискам устойчивых сочетаний слов

В предлагаемом методе использован кластерный анализ для автоматического группирования документов. Для каждого документа $t \in T$ строится вектор, называемый профилем документа $P(t)$, в котором каждый элемент хранит значение функции $K(t, m)$, выражающей степень связности текста t с каждым сочe-

танием слов m из множества всех устойчивых сочетаний M. Значение функции $K(t, m)$ определено как доля предложений текста t, в которых встретилось сочетание слов m.

$$K(t,m) = \frac{\sum_{i=1}^{|H_t|} v_i}{|M_t|}, \qquad (3)$$

где H_t – множество предложений текста t, M_t – множество устойчивых сочетаний, встретившихся в тексте t. Если i-е предложение текста t содержит сочетание m, то v_i=1. В противном случае – v_i=0.

Документы, вектора профилей которых совпадают во многих элементах, образуют группы.

Обработка естественно-языкового поискового запроса

Из слов текста запроса q формируется список сочетаний слов M_q. Алгоритм составления сочетаний для запроса совпадает с алгоритмом, описанным для смыслового фрагмента текста. Полученный из слов запроса набор сочетаний расширяется, используя словарь синонимов. Для каждого сочетания $m \in M_q$, в M_q добавляются сочетания, образующие с ним контекстно-зависимые синонимы.

На первом этапе поиска по сочетаниям определяется список групп, содержащих эти сочетания целиком или частично. На втором этапе для всех текстов документов t полученных групп вычисляется значение функции релевантности $R(q, t)$, зависящей от числа совпадений в наборах сочетаний слов, полученных из текста запроса (M_q) и текста документов (M_t). Полученный список документов, упорядоченных по значению функции релевантности, является результатом поиска.

В отличие от большинства существующих подходов использование в разработанном методе поиска автоматически полученного словаря контекстно-зависимых синонимов позволяет расширить результаты поиска документами, не содержащими слов запроса.

В **третьей главе** разработана и обоснована алгоритмическая структура, соответствующая описанному во второй главе методу. Эти алгоритмы используются при обработке каждого текста пополняемого фонда документов.

Деление слов на значимые и незначимые

Область значений функции значимости $S^*(w)$ находится в области неотрицательных действительных чисел. Для разделения слов на значимые и незначимые введено пороговое значение функции значимости $S_{порог}$. Все слова со значением функции значимости $S^*(w)$ (см. 1), превышающим $S_{порог}$ являются значимыми и используются при составлении сочетаний слов.

$$S^*(w) > S_{порог}. \qquad (4)$$

Выявление устойчивых сочетаний слов

При поиске устойчивых сочетаний формируются все возможные сочетания слов каждого смыслового фрагмента текста. В данной работе единицей смысловой фрагментации текста является предложение. Из последовательности слов предложения $H=<w_1, w_2..., w_l, ... w_N>$ длиной N составляются все сочетания длиной от 1 до K. Длиной сочетания называется количество слов в сочета-

нии. Все полученные сочетания сохраняются в специальной таблице с подсчетом количества повторений каждого сочетания. Повторяющиеся в рамках фонда в целом сочетания признаются устойчивыми и используются при обработке данных о статистической сочетаемости слов.

Группирование документов по их профилям

Для определения степени связности всех текстов из T с каждым сочетанием из M строится таблица P, строки которой проиндексированы номерами текстов, а столбцы – сочетаниями слов из M. Таким образом, k-я строка P_k таблицы P является профилем документа $t_k \in T$. На пересечении строки профиля P_k и столбца некоторого сочетания слов $m \in M$ записывается значение функции $K(t_k, m)$ связности текста t_k и сочетания m (см. 3).

Для вычисления степени схожести пары текстов t_k, $t_n \in T$ при k, $n \in [1; N_T]$ их профили P_k и P_n сравниваются с использованием функции Жаккара, значения которой образуют матрицу I, строка и столбец которой проиндексированы номерами текстов. Эта функция содержит операции пересечения и объединения векторов профилей документов. В программной реализации пересечение вычисляется как покомпонентное произведение векторов профилей, а объединение – как сумма этих векторов. Формула для заполнения ячейки таблицы I, находящейся на пересечении k–ой строки и n-го столбца, имеет вид:

$$I_{kn}^2 = \frac{\sum_{i=1}^{|M|}(P_{ki} * P_{ni})}{\sum_{j=1}^{|M|}(P_{kj} + P_{nj})}. \qquad (5)$$

Матрица I коэффициентов связности текстов является симметричной относительно главной диагонали, поэтому рассматривается только ее часть, расположенная выше главной диагонали. Для группирования текстов последовательно рассматриваются все строки таблицы I. Все тексты t_n ($n>1$), для которых указанный в этой строке коэффициент связности с первым текстом t_1 больше порогового значения I_ε, образуют с этим текстом группу. Аналогично рассматриваются все последующие (k-е) строки и указанные в них коэффициенты связности с текстом t_k остальных текстов t_n, $n>k$. Если ни один текст t_n из строки не образует группу с главным текстом строки, и главный текст до этого не был добавлен ни в одну группу, то он образует группу, состоящую только из самого себя. Так как операция создания группы выполняется для каждой строки таблицы I, максимальное количество групп текстов не превышает количество текстов. При этом текст может входить в несколько групп.

Язык изложения в тематически близких документах схож и содержит схожие слова, их сочетания и обороты, что позволяет рассматривать автоматически созданные группы как предметно-ориентированные.

Обработка данных о статистической сочетаемости слов

В соответствии с алгоритмом поиска родственных слов по очереди рассматриваются все слова в словаре. Первое необработанное слово образует группу и считается заглавным словом группы. Для заглавного слова проводится поиск родственных ему форм по всему словарю. Такими являются слова, имеющие

равное 1 расстояние редактирования с главным словом. Слова, имеющие расстояние редактирования, не превышающее ρ, признаются родственными, если входят в сочетания, различающиеся в одном слове. Существенно, что редактированию подвергаются только символы, занимающие в слове w позиции в интервале $(l_w/2+\Delta_l; l_w]$, где l_w - длина анализируемого слова, а Δ_l - определяет длину варьируемой суффиксной части слова. Все найденные слова включаются в эту группу. Если какое-то из найденных слов уже входит в некоторую группу, все слова той группы также перемещаются в формируемую группу.

При построении иерархии подчинения слов в сочетании рассматриваются устойчивые сочетания максимальной длины K. Для каждого слова w устойчивого сочетания m уровень подчинения определяется по значению функции $A(w, m)$. Слово, имеющее наибольшее значение этой функции – главное в сочетании. Остальные слова аналогично рассматриваются в рамках его подсочетания длины $(K-1)$, полученного из исходного удалением главного слова. Значение $A(w, m)$ сохраняется в таблице сочетаний.

При поиске контекстно-зависимых синонимов для каждого сочетания слов длины $(n-1)$ ищутся все содержащие его n-словные устойчивые сочетания. Эти n-словные сочетания содержат различающиеся слова, эквивалентные в контексте исходного сочетания. Если различаются главные слова, они признаются синонимами в данном контексте. Чтобы при поиске учитывать контекст синонимии, в таблице синонимов для каждой пары синонимичных слов хранятся связи с сочетаниями, в которых была обнаружена их синонимия. Это позволяет применять контекстно-зависимые синонимы только для текстов тех групп, в которые входят тексты, породившие эти синонимы.

Определение степени соответствия текста поисковому запросу

Все документы t из списка групп, полученного на первом этапе обработки поискового запроса, упорядочиваются по убыванию значения функции релевантности $R(q, t)$ запросу q. Значение функции релевантности $R(q, t)$ вычисляется как сумма относительного числа совпадений сочетаний различной длины $l \in [1; K]$ для предложений текста $M_t(l)$ и запроса $M_q(l)$, умноженного на весовую функцию $\theta(l)$, значение которой возрастает с увеличением длины сочетания.

$$R(q,t) = \sum_{l=1}^{K} \left(\frac{|M_q(l) \cap M_t(l)|}{|M_q(l)|} * \theta(l) \right). \qquad (6)$$

В **четвертой главе** описана архитектура и модульная структура разработанного программного обеспечения, приведены результаты экспериментальных исследований разработанного метода поиска, получены оценки эффективности алгоритмов, реализующих метод, а также исследованы условия применимости программного комплекса с использованием наращиваемого фонда документов.

Описание тестового набора документов

Разработанная система предназначена для использования в постоянно пополняемых фондах документов. Поэтому использованный при испытаниях информационный фонд постепенно наращивался и достиг 572 397 документов общим объемом 334 Мбайта, что соответствует примерно 300 000 страниц печатного текста. В этот набор включены технические статьи (10 статей), художе-

чатного текста. В этот набор включены технические статьи (10 статей), художественные произведения (21 документ) и реферативные статьи ВИНИТИ. Количественно рефераты распределены по предметным областям следующим образом: автоматика –77675, биология – 69729, сварка – 18447, экономика промышленности – 33424, физика – 70773, информатика – 19804, коррозия и защита от коррозии – 40640, математика – 26179, машиностроение – 100887, механика – 84645, охрана окружающей среды – 22146, вычислительные науки – 8017. Основную часть тестового информационного фонда составили рефераты, так как использование относительно коротких текстов (1-2 страницы) облегчает экспертам задачу определения релевантности документов запросам.

Определение эмпирических пороговых значений и коэффициентов

Для определения порогового значения функции значимости $S_{порог}$ все слова словаря, полученного при обработке тестового набора текстов, были упорядочены по возрастанию ее значения, вычисленного для каждого слова. Экспертом проанализирован полученный список, в результате чего определена граница, ниже которой располагаются слова, способные нести самостоятельный смысл. Для этих слов, являющихся значимыми, значение функции значимости превышает 0,05. Поэтому в данной работе $S_{порог}$ принято равным 0,05 ($S_{порог}$=0,05).

Для определения значения параметра Δ_l, ограничивающего длину варьируемой суффиксной части слова, было проведено группирование форм родственных слов с расстоянием редактирования равным 1 при Δ_l=1, 2, 3, 4. По результатам проведенных экспериментов определено, что с увеличением значения Δ_l (при Δ_l>2) качество группирования различных форм слова растет незначительно, а количество сгруппированных слов при этом стремительно снижается (рис. 4). Поэтому в данной работе Δ_l принято равным 2 (Δ_l=2).

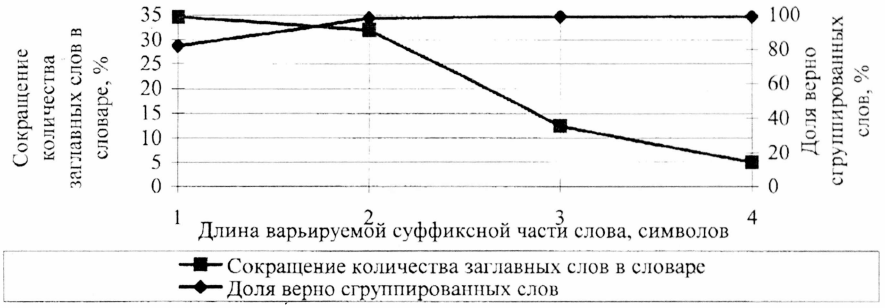

Рис. 4. Определение длины варьируемой суффиксной части слова

Выбор допустимого расстояния редактирования ρ, позволяющего признать слова родственными, обусловлен результатами эксперимента, в котором оценивалось качество поиска родственных слов с учетом их статистической сочетаемости с одинаковыми словами при значениях параметра ρ=2, 3, 4, 5. В результате эксперимента (рис. 5) определено, что увеличение ρ вызывает резкое снижение качества группирования слов. Поэтому для расстояния редактирования ρ выбрано минимальное значение 2 (ρ=2).

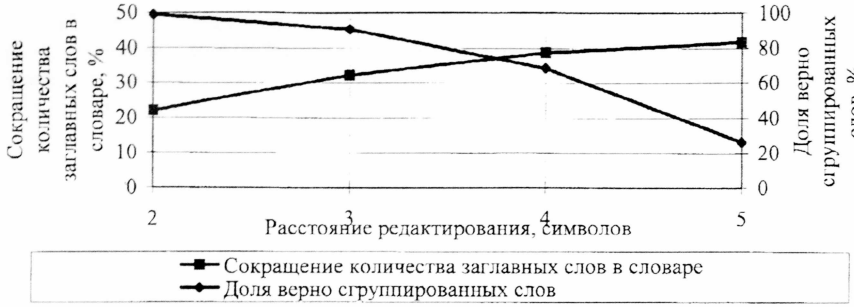

Рис. 5. Определение ограничения расстояния редактирования

Для определения ограничения на длину K составляемых сочетаний слов проведены эксперименты по составлению словаря контекстно-зависимых синонимов, используя таблицы сочетаний слов различной длины K (для K=2, 3, 4, 5). Из приведенных на рис. 6 графиков следует, что при K>3 точность определения синонимов почти прекращает рост. При этом количество полученных контекстно-зависимых синонимов снижается, поэтому принято ограничение на минимальную длину сочетаний, используемых при поиске синонимов, в 4 слова (K>3).

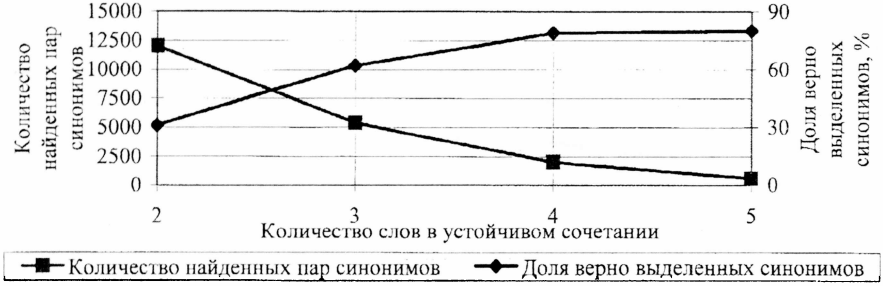

Рис. 6. Определение длины сочетаний слов

С увеличением длины сочетания растет число составляемых сочетаний слов. Но сочетания большой длины редко повторяются и не могут служить источником данных о статистической сочетаемости слов. Например, среди сочетаний из 5 слов в текстах тестового информационного фонда повторяющиеся составляют менее 6%. Поэтому устанавливается ограничение на максимальную длину сочетания, равное 4 (K=4).

При определении количества повторений U сочетаний слов в текстах, достаточного для признания сочетания устойчивым, исследовано количество найденных контекстно-зависимых синонимов и качество их определения при числе повторов U=1, 2, 3, 4, 5. Увеличение количества повторов при U>1 незначительно увеличивает качество выделения синонимов (рис. 7). При этом существенно сокращается количество найденных синонимов. Таким образом, устойчивыми признаются сочетания, повторяющиеся больше 1 раза (U=2).

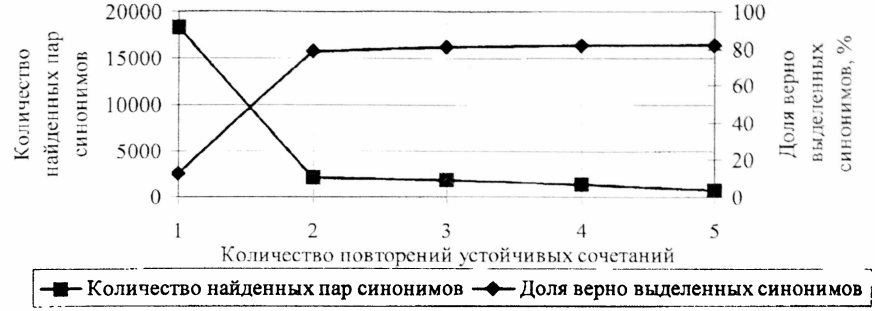

Рис. 7. Определение достаточного числа повторений сочетаний слов

Для выбора значения параметра I_ε, влияющего на количество создаваемых групп и насыщенность этих групп текстами, проведены эксперименты с различными значениями $I_\varepsilon \in [0; 1]$. Из результатов экспериментов следует, что уменьшение значения I_ε повышает количество текстов в каждой группе (рис. 8). Это сокращает время поиска соответствующих запросу групп текстов (первый этап обработки поискового запроса). Увеличение значения I_ε снижает количество текстов в каждой группе, что сокращает время, необходимое для вычисления значений функции релевантности для текстов найденных групп (второй этап).

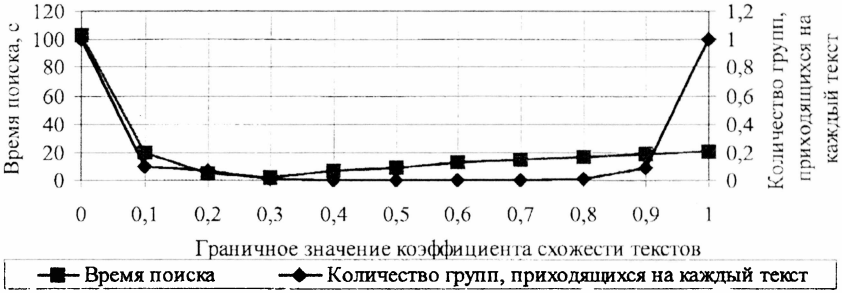

Рис. 8. Определение граничного значения для группирования текстов

При $0{,}2<I_\varepsilon<0{,}3$ время поиска минимально при наибольшем количестве групп текстов, поэтому в данной работе I_ε принято равным 0,25 (I_ε=0,25).

Экспериментальные оценки требуемых ресурсов

Группирование схожих по написанию форм родственных слов, позволило существенно снизить количество слов, используемых для составления сочетаний, так как вместо всех форм слова используется одно заглавное слово. Для рассмотренного тестового набора документов количество обрабатываемых слов сократилось в 2,05 раза. Во столько же раз снизилось время составления сочетаний и дисковое пространство, требуемое для их хранения.

При последовательном добавлении документов в хранилище системы было обнаружено снижение темпов роста количества слов в словаре с каждым новым текстом (рис. 9).

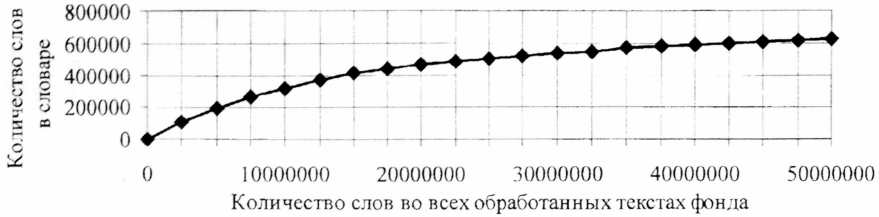

Рис. 9. Зависимость количества слов в словаре от объема обработанного текста

Количество слов (уникальных комбинаций символов), занесенных в словарь в результате анализа тестового набора текстов, составило чуть более 1% количества слов в обработанном тексте.

Разработанный в рамках выполненной работы алгоритм получения всех сочетаний слов требует меньшего числа операций перестановок слов, чем алгоритм последовательного составления сочетаний каждой длины, и выполняет задачу составления сочетаний в 1,5-4 раза быстрее алгоритма последовательного составления сочетаний каждой длины.

При пополнении информационного фонда документов интенсивный рост количества сочетаний продолжается намного дольше, чем рост объема словаря, но, как показано на рис. 10, также постепенно снижается.

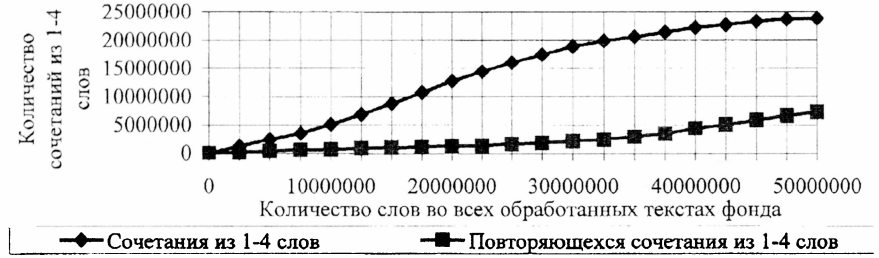

Рис. 10. Зависимость количества сочетаний от объема обработанного текста

Сохраняя тенденцию к снижению темпов роста количества сочетаний, при достижении объема текста в 35 000 000 слов скорость роста количества повторяющихся сочетаний увеличивается. После обработки 35 000 000 слов с каждым новым предложением создается все меньше новых сочетаний и собирается все больше информации о контексте употребления уже сохраненных сочетаний слов. Таким образом, с увеличением объема фонда документов не происходит пропорционального роста объема потребляемых ресурсов, а интенсивность его увеличения постоянно снижается с пополнением фонда новыми документами.

Оценка качественных показателей разработанного метода поиска

Проведенная проверка качества группирования 631 873 слов в 307 193 группы родственных слов показала, что предложенный вариант поиска родственных слов допускает малое количество ошибок: 8 на 1 000 слов. Это позволяет автоматически находить различные формы слов при обработке больших объемов текста.

Автоматически составленный по разработанному алгоритму список незначимых слов был ограничен 2000 слов, что составляет почти 2% от количества слов в словаре. В созданный список незначимых попало 38 слов, которые могут нести самостоятельный смысл. Таким образом, погрешность процедуры выделения значимых слов не превысила 2% от количества слов, признанных незначимыми.

В результате обработки тестового набора текстов были получены 40903 пары слов, устойчиво встречающиеся в одинаковом контексте. Из них 2034 пары слов были признаны синонимичными в соответствии с разработанным алгоритмом выделения синонимов. Экспертная проверка процедуры автоматического поиска контекстно-зависимых синонимов подтвердила успешность для 79% и отвергла 21% результатов.

Для оценки полученного поискового программного продукта проведены эксперименты, в которых использован набор всех запросов, которые были адресованы поисковой системе библиотеки в период с июня 2004 по февраль 2006. В этот период системой было обработано 76615 запросов. Число запросов для полнотекстового поиска равно 40478. Среди них выделено 22369 различающихся запросов, из которых случайным образом было отобрано 500 запросов. Для оценки соответствия результатов поиска запросам введена функция качества поиска $Q_{\%}$.

$$Q^2_{\%} = \frac{k}{1+k}\left[\left(\frac{|F \cap E|}{|F|}\right)^2 + \frac{1}{k}\left(\frac{|F \cap E|}{|E|}\right)^2\right] * 100\%, \qquad (7)$$

где k – коэффициент, определяющий приоритет точности над полнотой поиска; F - множество найденных документов; E - множество искомых документов.

Средний показатель качества поиска созданным программным комплексом по 500 тестовым запросам среди документов тестового фонда составил 88%. Для сравнения рассмотрена поисковая методика, реализованная в локальной версии поисковой машины Яндекс (www.yandex.ru), используемой в некоторых библиотеках для поиска по фондам документов. Для нее средний показатель качества поиска, оцененный по такой же методике и при таких же условиях, составил 65%.

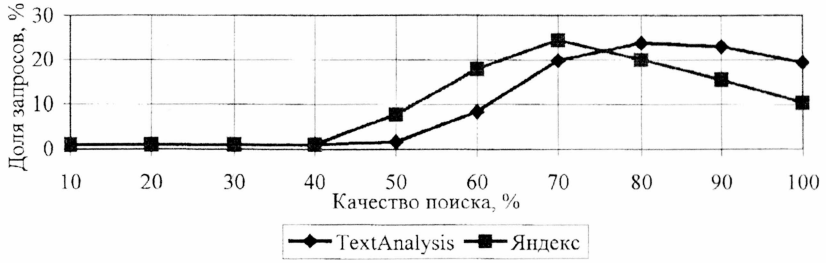

Рис. 11. Сравнение качества поиска исследуемыми поисковыми системами

Сравнение распределения значений функции качества поиска показало, что качество поиска созданной поисковой системой выше (рис. 11) благодаря учету контекстных связей слов, а также использованию словаря контекстно-зависимых синонимов. При этом на выполнение запросов требовалось от 2 до 15 секунд.

Созданная система использовалась для анализа текстов византийских источ-

ников на древнегреческом языке, что подтвердило ее способность обрабатывать тексты на различных языках. Качество обработки текстов положительно оценено сотрудниками Института Всеобщей Истории Российской Академии Наук.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Создан метод поиска информации в фондах электронных документов, базирующийся на статистической обработке контекста слова и сопоставлении найденных устойчивых сочетаний слов в текстах документа и запроса.
2. Разработаны новые алгоритмы автоматического выполнения морфологической, синтаксической и семантической обработки текстов, обеспечивающие возможность интеллектуального поиска информации на основе введенной функции оценки значимости слова с использованием словаря синонимов, составленного без участия человека.
3. Предложен метод автоматического обучения поисковой системы, позволяющий проводить поиск вне зависимости от предметной и тематической направленности документов, по фондам, содержащим тексты на различных языках.
4. Создан программный комплекс, реализующий разработанный метод и позволяющий пополнять и обрабатывать информационные фонды, выполнять по ним поиск документов.
5. Проведено исследование практической применимости предложенного метода поиска. Разработан критерий качества поиска, с использованием которого подтверждено увеличение полноты и степени релевантности найденных документов по сравнению известными поисковыми системами.

РАБОТЫ ПО ТЕМЕ ДИССЕРТАЦИИ

1. *Григорьев А.С.* Организация хранения, обработки и доступа к полнотекстовым документам в современных АБИС // Новые информационные технологии: Матер. шестого всерос. научн.-практ. сем. - М., 2003. - С. 128-138.
2. *Григорьев А.С.* Принципы создания современной библиотечной поисковой системы // Информатика и системы управления в XXI веке: Труды молодых ученых, аспирантов и студентов. Сб. тр. - 2003. - №1. - С. 330-333.
3. *Григорьев А.С.* Машинное понимание естественного языка при составлении запросов к поисковой системе библиотеки // Новые информационные технологии: Матер. седьмого всерос. научн.-практ. сем. - М., 2004. - С. 70-76.
4. *Григорьев А.С.* Новая система обработки естественно-языковых текстов в исследовании понятий и терминов византийских источников // Межкультурное взаимодействие и его интерпретации: Материалы всероссийской научной конференции. - М., 2004. - С. 197-200.
5. *Григорьев А.С.* Исследование проблемы проектирования систем обработки естественно-языковых текстов и организации поиска по ним // Информатика и системы управления в XXI веке: Труды молодых ученых, аспирантов и студентов. Сб. тр. - 2004. - №2. - С. 184-188.
6. *Григорьев А.С.* Автоматическое получение ключевых словосочетаний текста электронного документа на произвольном языке // Новые информационные технологии: Матер. восьмого всерос. научн.-практ. сем.- М., 2005.-С.110-118.

Подписано к печати 26.04.2006 г. Заказ № 200. Объем 1.0 п.л. Тир. 100 экз.
Типография МГТУ им. Н.Э. Баумана.