

На правах рукописи

КРИЖАНОВСКИЙ
Андрей Анатольевич

МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ
ПОСТРОЕНИЯ СПИСКОВ СЕМАНТИЧЕСКИ БЛИЗКИХ СЛОВ
НА ОСНОВЕ РЕЙТИНГА ВИКИ-ТЕКСТОВ

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург
2008

Работа выполнена в Учреждении Российской академии наук Санкт-Петербургском институте информатики и автоматизации РАН.

Научные руководители:
доктор технических наук,
профессор

Смирнов Александр Викторович

Официальные оппоненты:
доктор технических наук,
профессор

Гаврилова Татьяна Альбертовна

доктор технических наук,
профессор

Городецкий Владимир Иванович

Ведущая организация:

Вычислительный центр им. А. А. Дородницына РАН

Защита состоится «16» сентября 2008 г. в 12:30 на заседании диссертационного совета Д.002.199.01 при Санкт-Петербургском институте информатики и автоматизации РАН по адресу: 199178, Санкт-Петербург, В.О., 14 линия, 39.

С диссертацией можно ознакомиться в библиотеке Санкт-Петербургского института информатики и автоматизации РАН

Автореферат разослан «29» июля 2008 г.

Ученый секретарь
диссертационного совета Д.002.199.01

 **Ронжин Андрей Леонидович**

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертации

Увеличение числа и изменение качества электронных документов на локальных компьютерах и в сети Интернет, а также развитие новых информационных технологий требуют разработки математического и программного обеспечения для более точного и быстрого текстового поиска.

Одной из актуальных задач данного направления является поиск похожих объектов, который включает такие (на первый взгляд разные, но общие по способам решения) задачи, как поиск похожих текстовых документов, поиск семантически близких слов, вычисление меры сходства между вершинами графа. Анализ работ в области вычислительной лингвистики показал большое разнообразие алгоритмов, предлагающих решение этих задач: Hypertext Induced Topic Selection (HITS), PageRank, ArcRank, алгоритм извлечения синонимов из толкового словаря, алгоритм извлечения контекстно связанных слов и др.

Поиск семантически близких слов является подзадачей таких актуальных задач информационного поиска, как: (1) расширение / переформулировка запросов с помощью тезаурусов (в поисковых системах), (2) распознавание запроса в запросно-ответных системах, (3) определение значения многозначного слова и (4) автоматическое создание тезаурусов.

Объектом исследования диссертационной работы является синонимия и семантическая близость слов. Поиск семантически близких слов основан на рейтинге вики-текстов в проблемно-ориентированном корпусе с гиперссылками и категориями. Два текста связаны гиперссылкой, если один из них упоминает (ссылается на) другой текст. Тематическая направленность каждого текста определена экспертом с помощью категорий¹. Эксперт выбирает категории (из заданного набора) и присваивает их тексту.

Под семантически близкими словами подразумеваются слова с близким значением, встречающиеся в одном контексте. Более строго семантически близкие слова определяются в работе через понятия корневого набора (релевантные документы), авторитетных и хаб-документов, вводимые в работах Клейнберга.

Современные алгоритмы поиска синонимов (например, алгоритм SimRank, алгоритм Similarity Flooding) не учитывают такую информацию корпусов проблемно-ориентированных документов, как: (1) ключевые слова

1 Связь, осуществляемая гиперссылкой, не имеет семантики, т. е. не описывает смысла этой связи. Однако категории представляют *бинарную* (связаны два объекта) *семантическую* сеть с иерархическими отношениями (родо-видовые и часть – целое).

и (2) категории, классифицирующие документы по их тематической принадлежности. Это актуальная проблема, поскольку большое количество новых документов представлено в современном формате гипертекстовой среды – *вики*, содержащем и ключевые слова, и категории. Текстовые вики ресурсы были выбраны из-за наличия (1) общего стандарта документов (единообразные метаданные: заголовок документа, категории), (2) классификации документов.

Цель работы состоит в решении задачи автоматизированного построения упорядоченного списка семантически близких слов в проблемно-ориентированных корпусах с гиперссылками и категориями (на примере корпуса текстов открытой энциклопедии Википедия) с возможностью оценки результатов поиска. Для достижения поставленной цели необходимо:

1. Проанализировать методы поиска семантически близких слов, обосновать выбор текстовых ресурсов, алгоритма (с возможной адаптацией) и программных систем для автоматической обработки текстов на естественном языке.

2. Разработать подход к поиску семантически близких слов (в корпусе текстовых документов с гиперссылками и категориями).

3. Разработать алгоритмы поиска семантически близких слов в корпусе текстовых документов с гиперссылками и категориями.

4. Спроектировать и реализовать программный комплекс поиска семантически близких слов; разработать способы численной оценки наборов синонимов.

Методы исследования включают: методы кластерного анализа, теории графов, элементы теории сложности алгоритмов.

Положения, выносимые на защиту:

1. Подход к поиску семантически близких слов на основе метаинформации в проблемно-ориентированном корпусе, содержащем два типа текстовых документов (статья и категория) и два типа отношений: иерархические отношения (родо-видовые и часть – целое) и гиперссылки.

2. Адаптированный HITS алгоритм поиска семантически близких слов в корпусе текстовых документов с гиперссылками и категориями. Модификация алгоритма включает: (1) новый способ построения корневого набора (релевантных документов), позволяющий отказаться от предварительного поиска документов, а также (2) использование механизма иерархической кластеризации для объединения слов в смысловые группы.

3. Клиент-серверная архитектура программного комплекса, предназначенного для решения задачи поиска семантически близких слов с возможностью оценки (с помощью удалённого доступа к тезаурусам и на

основе модификации коэффициента Спирмена) семантической близости построенных списков слов.

4. Программный комплекс поиска семантически близких слов в проблемно-ориентированном корпусе текстов с динамической визуализацией результатов поиска.

5. Архитектура системы индексирования вики-текстов и её программная реализация.

Научная новизна

1. Новизна предложенного подхода к поиску семантически близких слов в проблемно-ориентированном корпусе заключается в том, что кроме гиперссылок дополнительно учитывается метаинформация документов (ключевые слова, категории).

2. Новизна адаптированного NITS алгоритма состоит в том, что при поиске наиболее похожих документов в корпусе учитываются не только гиперссылки, но и категории, что позволяет применить механизм иерархической кластеризации, объединяющий семантически близкие слова в смысловые группы.

3. Новый способ построения корневого набора документов в адаптированном NITS алгоритме заключается в выборе документов, связанных гиперссылками с исходным документом (заданным пользователем), что позволяет отказаться от шага «предварительный веб-поиск документов».

4. Коэффициент Спирмена модифицирован для численного сравнения списков семантически близких слов; отличие заключается в возможности сравнивать списки разной длины.

5. Впервые предложен показатель степени синонимичности набора слов, заключающийся в сравнении этого набора с эталонным списком синонимов (например, из тезауруса).

6. Впервые спроектирована распределённая архитектура программного комплекса, позволяющего выполнять поиск семантически близких слов и оценивать результаты поиска на основе удалённого доступа к тезаурусам.

7. Эксперименты подтвердили выполнение закона Ципфа для текстов Русской Википедии и Википедии на английском упрощённом языке на основе построенных индексных баз данных.

Обоснованность и достоверность научных положений, основных выводов и результатов диссертации обеспечивается за счёт тщательного анализа состояния результатов исследований в области вычислительной лингвистики, подтверждается экспериментами на основе трёх корпусов текстов Русской Википедии, Английской Википедии и Simple Wikipedia (Википедия на английском упрощённом языке).

Практическая ценность работы заключается в том, что реализованный программный комплекс *Synarcher*² (на языке Java), включающий алгоритм поиска, позволяет выполнять поиск семантически близких слов в английской и русской версии энциклопедии Википедия с динамической визуализацией результатов поиска.

Поиск семантически близких слов в Википедии позволит пользователям (i) находить энциклопедические статьи, близкие по тематике к заданным, для более углублённого изучения некоторого понятия; (ii) устанавливать недостающие ссылки между связанными по смыслу статьями.

Спроектирован и реализован программный комплекс *Russian POS Tagger (RuPOSTagger)*³, позволяющий интегрировать среду GATE и модуль морфологической обработки русского языка Lemmatizer (компании Диалинг). Комплекс *RuPOSTagger* предоставляет доступ к функциям модуля Lemmatizer на основе XML-RPC протокола из системы GATE или из отдельного Java приложения.

Реализация результатов работы. Исследования, отражённые в диссертации, были поддержаны грантами РФФИ (проект № 02-01-00284 «Методологические и математические основы построения компьютерных систем быстрой интеграции знаний из распределённых источников» 2002-2004 гг., № 06-07-89242 «Методология и модели интеллектуального управления конфигурациями распределённых информационных систем с динамически изменяющимися структурами», 2006-2008 гг.; № 05-01-00151 «Методологические и математические основы построения контекстно-управляемых систем интеллектуальной поддержки принятия решений в открытой информационной среде», 2005-2007 гг.), грантами Президиума РАН (проект № 2.44 «Многоагентный подход к построению компьютерной среды для быстрой интеграции знаний из распределённых источников» 2001-2003 гг. и проект № 2.35 «Контекстно-управляемая методология построения распределённых систем интеллектуальной поддержки принятия решений в открытой информационной среде» 2003-2008 гг.), а также грантом ОИТВС РАН (проект № 1.9 «Разработка теоретических основ и многоагентной технологии управления контекстом в распределённой информационной среде» 2003-2005 гг.).

Часть результатов была использована при выполнении контракта «Интеллектуальный доступ к каталогам и документам» на создание системы поддержки клиентов, реализованной для немецкой промышленной компании Фесто, 2003–2004 гг. Разработана архитектура программной системы поиска

2 Программная реализация: <http://synarcher.sourceforge.net>

3 Программная реализация: <http://rupostagger.sourceforge.net>

семантически близких слов в исследовательском проекте CRDF № RUM2-1554-ST-05 «Онтолого-управляемая интеграция информации из разнородных источников для принятия решений», 2005-2006 гг.

Апробация результатов работы: Основные положения и результаты диссертационной работы представлялись на международном семинаре «Автономные интеллектуальные системы: агенты и извлечение данных» (Санкт-Петербург 2005), международных конференциях: «Диалог» (Бекасово 2006), «Речь и Компьютер» (Санкт-Петербург 2006), «Корпусная лингвистика – 2006» (Санкт-Петербург) и первой конференции в России «Вики-конференции 2007» (Санкт-Петербург 2007).

Публикации: Основные результаты по материалам диссертационной работы опубликованы в 8 печатных работах, в том числе в 2 журналах из списка ВАК.

Структура и объем работы: Диссертация объемом 156 страниц (188 с приложениями) содержит введение, четыре главы и заключение, приложения, список литературы (189 наименований), 35 рисунков, 14 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована важность и актуальность темы диссертации, сформулированы цели диссертационной работы и решаемые задачи, определена научная новизна работы и указана её практическая ценность, кратко изложены основные результаты.

В первой главе диссертации кратко описаны предыдущие исследования в области теории графов и вычислительной лингвистики, цель которых – поиск похожих интернет-страниц, а также вычисление меры сходства между вершинами графа. В настоящее время широкое распространение получил ряд алгоритмов, предлагающих решение этих задач, например, алгоритмы HITS, PageRank, алгоритм распределения рангов ArcRank, алгоритм извлечения синонимов из толкового словаря, метод извлечения контекстно связанных слов и др. Такое разнообразие алгоритмов неслучайно.

Сложность организации поиска семантически близких слов (и оценки качества поиска) определяется рядом причин. Во-первых, понятие семантической близости определено не для слов, а для значений слов, то есть неразрывно связано с контекстом. Во-вторых, язык – это открытая система. Слова могут устаревать или получать новые значения. Особенно активное словообразование и присвоение новых значений словам наблюдается в науке, в её молодых, активно развивающихся направлениях. В-третьих, нет однозначного общепринятого способа вычисления близости значений слов. Это обуславливает сложность численной оценки работы алгоритмов поиска

семантически близких слов.

Поиск синонимов и семантически близких слов является одной из задач автоматической обработки текста. Проведённый анализ проблемы автоматизированного построения списков семантически близких слов показал, что здесь можно выделить следующие подзадачи.

1. *Разработка (выбор) алгоритмов.* Предложена классификация алгоритмов, выполняющих поиск похожих документов и близких по значению слов: (1) поиск на основе анализа ссылок, когда ссылки заданы явно гиперссылками (HITS, PageRank, ArcRank, Green, WLVM и др.), и когда ссылки нужно построить (Similarity Flooding, алгоритм извлечения синонимов из толкового словаря); (2) поиск на основе анализа текста: (2.1) статистические алгоритмы, а именно: ESA, сходство коротких текстов, извлечение контекстно связанных слов на основе частотности словосочетаний и (2.2) автоматическое понимание текстов; и (3) поиск на основе анализа и ссылок, и текста (алгоритмы учёных Bharat и Maguitman). Требованием к алгоритму поиска семантически близких слов является учёт дополнительных возможностей, предоставляемых рассматриваемым корпусом документов, а именно: (1) наличие категорий, классифицирующих документы по их тематической принадлежности, (2) наличие метаинформации в виде ключевых слов (например, заголовков документа).

2. *Оценка результатов работы алгоритма.* Анализ работ в данной области показал, что необходима разработка оригинальных показателей степени синонимичности полученных списков семантически близких слов. Задача *выбора текстового ресурса* определяет способ оценки результатов поиска. Среди множества проблем создания, использования корпусов можно выделить общую проблему отсутствия единого стандарта. В диссертации в качестве корпуса вики-текстов предлагается использовать коллективную онлайн энциклопедию Википедия, содержащую 2.4 млн документов.

3. *Программные ресурсы для обработки текста.* При выборе программных инструментальных средств разработки и проектирования архитектуры программного комплекса автор придерживался следующих требований: открытость исходного кода, кроссплатформенность, модульность архитектуры.

4. *Визуализация результатов поиска.* Анализ поисковых систем показал, что некоторые из них обеспечивают визуализацию результатов поиска. Автором выделены системы, представляющие результаты поиска в виде статического и динамического изображений.

Во второй главе разработан подход к поиску семантически близких слов (СБС) в корпусе текстов с гиперссылками и категориями на основе указанных выше подзадач поставленной проблемы (рис. 1). Входными

данными для поиска семантически близких слов являются исходное слово, корпус документов и список слов, уточнённый пользователем. Последовательность взаимодействия частей системы указаны на рис. 1 числами (1-7). Входными данными для поискового алгоритма являются слово, заданное пользователем (1-2), и данные корпуса (2). Алгоритм строит упорядоченный список СБС (3), а пользователь получает возможность работать с ним благодаря визуализации (4-5). В ходе работы пользователь уточняет список СБС и может запустить алгоритм повторно (6-7). Достоинствами подхода являются (1) визуализация результатов поиска, (2) возможность уточнения запроса в ходе работы пользователя.

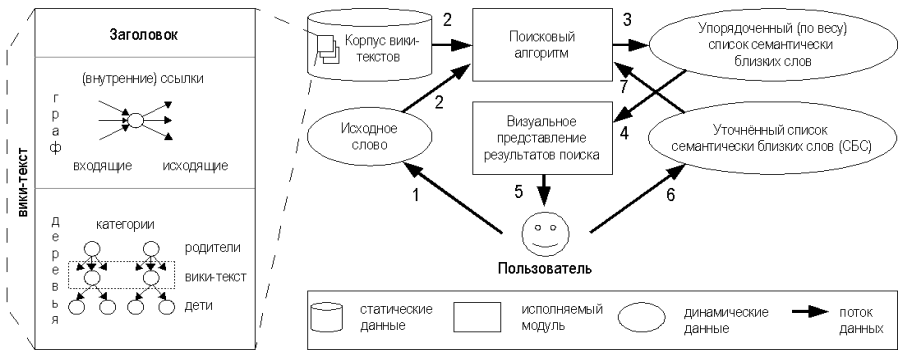


Рис. 1. Подход к поиску семантически близких слов

Из двух классических алгоритмов HITS и PageRank, удовлетворяющих вышеприведённым требованиям (учёт ключевых слов, гиперссылка), был выбран алгоритм HITS по следующим причинам: (1) формулы вычисления в PageRank требуют экспериментального подбора коэффициента, (2) значения весов (рассчитанные с помощью PageRank) не могут быть использованы напрямую для поиска похожих страниц, то есть нужен дополнительный алгоритм, который будет искать похожие страницы на основе весов PageRank.

В HITS алгоритме для поиска интернет-страниц, соответствующих запросу пользователя, используются такие понятия, как авторитетный документ и хаб-документ. *Авторитетный документ* – это документ, соответствующий запросу пользователя, имеющий больший удельный вес среди документов данной тематики, то есть большее число документов ссылаются на данный документ. *Хаб-документ* – это документ, содержащий много ссылок на авторитетные документы.

Каждому документу в HITS алгоритме сопоставляются веса a (*authority*) и

h (*hub*), которые показывают, соответственно, насколько документ является авторитетным и насколько он является хорошим хаб-документом. Формулы алгоритма HITS для итеративного вычисления весов таковы:

$$h_j = \sum_{i:(j,i) \in E} a_i ; \quad a_j = \sum_{i:(i,j) \in E} h_i \quad (1)$$

где значение веса h_j показывает насколько документ j является хорошим указателем на релевантные документы (j -ый документ рассматривается как хаб-документ) и a_j — насколько документ j является авторитетным документом.

$$\begin{aligned} A \subset V, |A| = N, \sum_{v \in A} a_v \rightarrow \max \\ H \subset V, |H| = M, \sum_{v \in H} h_v \rightarrow \max \\ A \subset V, H \subset V, k \in [0, 1]: \\ k \cdot \sum_{v \in A} a_v + (1 - k) \cdot \sum_{v \in H} h_v \rightarrow \max \\ A \subset V, H \subset V, \forall a \in A \exists h \in H: \\ \Gamma^+(h) \ni s, a \end{aligned}$$

Предложена формализация понятия «похожие вершины» графа, а также понятий, неформально введённых Клейнбергом (см. выше): авторитетные (A) и хаб (H) документы (здесь вершины). В формулах: s обозначает исходную вершину, k — весовой параметр, определяющий приоритет авторитетных либо хаб-документов.

Адаптированный HITS алгоритм (АНИТС), разработанный автором, учитывает метаинформацию проблемно-ориентированного корпуса документов: (1) ключевые слова, (2) категории, классифицирующие документы по их тематической принадлежности и (3) гиперссылки. В HITS алгоритме граф содержит взвешенные вершины одного типа. Предлагается модификация алгоритма для учёта трёх типов вершин (авторитетный документ, хаб-документ и новый тип вершин для HITS алгоритма — категория) и трёх типов дуг (документ-документ и новые для для HITS: документ-категория и категория-категория), определяемых проблемно-ориентированным корпусом текстов.

Таким образом, исходными данными для алгоритма являются (рис. 1): сеть документов-энциклопедических статей (вершины — документы, дуги — гиперссылки) и дерево категорий (вершины — категории, дуги связывают категории-родителей с категориями-детьми). Причём элемент сети (статья) связан с одним или несколькими элементами дерева (категории). Для каждого документа определены: (i) список документов, ссылающихся на данный документ, (ii) список документов, на которые ссылается данный документ, (iii) список категорий, определяющих его тематическую принадлежность.

Шаги АНИТС алгоритма представлены на рис. 2. Шаги, предложенные автором, выделены пунктиром. Тематическая классификация документов и

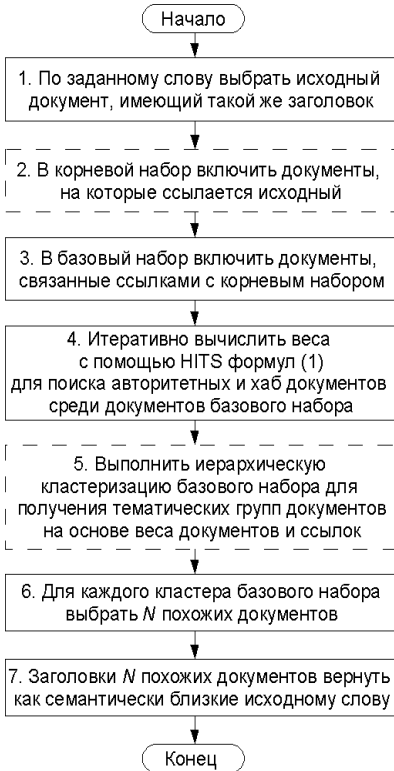


Рис. 2. Адаптированный HITS алгоритм

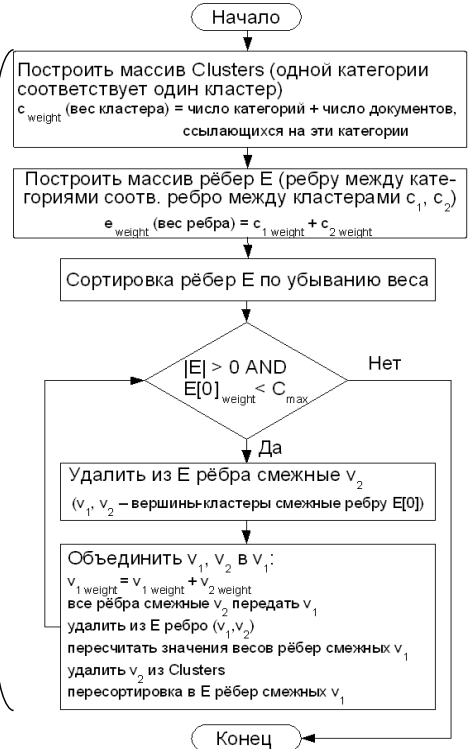


Рис. 3. Иерархическая кластеризация

жёсткая связь заголовка и гиперссылок документа (две особенности вики) позволяют применить разработанный алгоритм иерархической кластеризации (рис. 3), группирующий документы по тематической принадлежности, что даёт кластеры (группы) СБС, соответствующие подзначениям искомого слова. Одним из параметров алгоритма, задаваемых пользователем, является максимально допустимый вес кластера (C_{max}), определяющий максимальное число документов и категорий в одном кластере.

Временная сложность вычисления весов на шаге 4 (рис. 2) составляет $O(I \cdot N^2)$, где I – число итераций, N – число документов в базовом наборе. Сложность кластеризации (рис. 3) определяется как $O(C^3 + N \cdot C^2)$, где C – число категорий, N – число документов в базовом наборе. Итак, временная сложность АНITS алгоритма есть $O(C^3 + N \cdot C^2 + I \cdot N^2)$, она выросла по

сравнению со сложностью HITS алгоритма $O(I \cdot N^2)$, но при этом выросла точность поиска (см. далее результаты экспериментов).

Предложены показатели численной оценки построенного упорядоченного списка семантически близких слов, разработанные автором. *SER (Strong Error Rate)* – оценка грубой ошибки, а именно: число слов (в процентах от общего числа найденных слов), которые не имеют отношения к исходному слову (определяется либо вручную экспертом либо автоматически – по наличию / отсутствию в тезаурусе). Обозначим $Words_{AHITS}^w$ – множество слов, найденных с помощью AHITS алгоритма для слова w , $Words_{WordNet}^w$ – список синонимов из тезауруса WordNet для слова w , $Words_{Moby}^w$ – список семантически близких слов из Moby, знак «\» – вычитание множеств.

$$SER(w) = \frac{|Words_{AHITS}^w \setminus (Words_{WordNet}^w \cup Words_{Moby}^w)|}{|Words_{AHITS}^w \cup Words_{WordNet}^w \cup Words_{Moby}^w|} \cdot 100 \quad (2)$$

Для численной оценки степени сходства эталонного списка (WordNet, Moby) и автоматически построенного списка СБС адаптирован коэффициент Спирмена. Адаптация позволяет сравнивать ранжирование элементов в списках разной длины. Итак, для исходного слова даны: эталонный список A , построенный экспертом, и список B , построенный автоматически. В конец списка B добавляются отсутствующие в нём элементы A . Каждому элементу списка назначается ранг (порядковый номер) от 1 до N . Далее применяется формула (3), где сравниваются положения в списках общих элементов, то есть вычисляется сумма модулей расстояний между i -ми элементами набора, S – число общих элементов.

$$F^S(s_1, s_2) = \sum_{i=1}^S |s_1(i) - s_2(i)| \quad (3)$$

Коэффициент Спирмена позволяет сравнивать с эталонным списком ранжирование одного и того же набора слов AHITS алгоритмом при разных входных параметрах (размер корневого набора, максимальный вес кластера).

В третьей главе представлена архитектура программного комплекса *Synarcher*, включающего такие модули, как: модуль *kleinberg*, предоставляющий доступ к данным Википедии и реализующий алгоритм AHITS, модуль визуализации *TGWikiBrowser*, модуль вычисления степени синонимичности списков слов на основе удалённого доступа к тезаурусам Moby и WordNet. Данные Википедии (тексты, ссылки) хранятся в базе данных MySQL, размещённой локально или удалённо. Параметры поиска и слова, помеченные пользователем как синонимы, хранятся на компьютере пользователя.

Модуль визуализации написан на основе кода программы визуализации вики-страниц – *TouchGraph WikiBrowser*. Для более удобной навигации код был существенно модифицирован, в контекстное меню были добавлены команды: спрятать все вершины (Hide all except node), пометить вершину как синоним (Rate synonym), показать категории (Expand Categories).

В главе описаны экраны программы: (1) экран Article, позволяющий просмотреть энциклопедическую статью, соответствующую выбранному слову, (2) экран Database, позволяющий подключиться к базе данных и получить статистику по базе данных, (3) экран Synonyms, на котором задаются параметры АНITS алгоритма, выводятся результаты поиска в табличной и текстовой форме, (4) экран с результатами поиска СБС в виде графа (рис. 8).

Далее описана модель, позволяющая интегрировать модуль морфологического анализа *Lemmatizer* в систему GATE (на основе разработанных автором XML-RPC клиента и сервера, поскольку XML-RPC протокол связывает приложения, написанные на разных языках программирования, здесь C++ и Java). Разработана архитектура системы индексирования вики-текстов, включающая программные модули GATE и Lemmatizer (рис. 4). Реализован программный комплекс индексации текстов Википедии на трёх языках: русский, английский, немецкий.

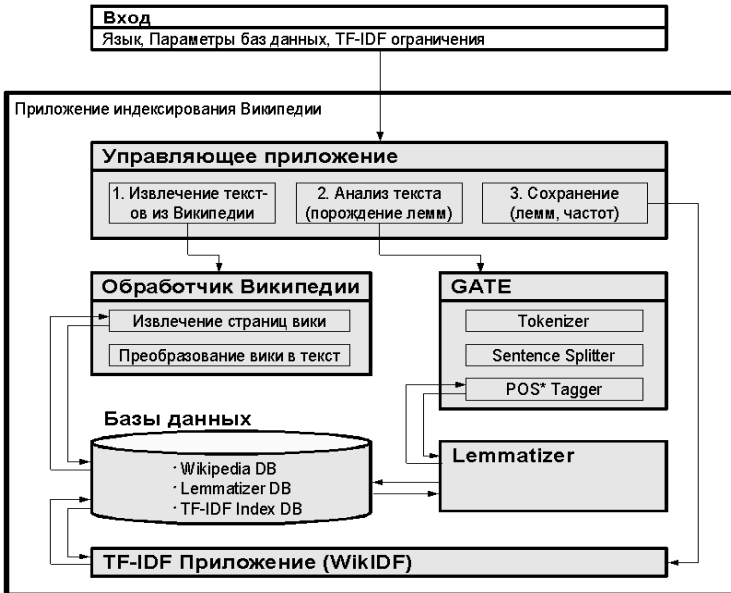


Рис. 4. Архитектура системы индексирования вики-текстов

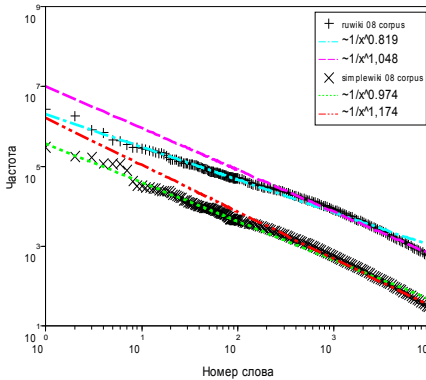


Рис. 5. Экспериментальная оценка выполнения закона Ципфа для корпусов Русской Википедии и Simple Wikipedia

число слов, лексем). Обнаружен более быстрый рост английской: за пять месяцев (сент. 2007 — февр. 2008) скорость роста числа статей была больше на 12% и на 6% быстрее чем в русской пополнялся лексикон Википедии на английском упрощённом языке. Эксперименты подтверждают выполнение закона Ципфа⁴ для текстов Русской Википедии и Википедии на английском упрощённом языке (рис. 5).

Далее описаны эксперименты поиска СБС в английской и русской версии Википедии с помощью АНИТС алгоритма. Описан пример работы в программе *Synarcher*. Эксперименты показали, что разработанный программный комплекс *Synarcher* позволяет найти синонимы и семантически близкие слова в английской Википедии, отсутствующие в современных тезаурусах WordNet, Moby (например, найден синоним *Spationaut* для слова *Astronaut*). Тем не менее, некоторые синонимы, представленные в тезаурусах и Википедии, не были найдены.

На рис. 6 представлен пример оценки времени работы и точности поиска СБС для слов и словосочетаний, имеющих энциклопедические статьи в Русской Википедии (*Аэродром*, *Беспилотный летательный аппарат*, *Движитель*, *Интернационализация*, *Истина*, *Пропеллер*, *Самолёт*, *Сюжет*,

В четвёртой главе описаны данные морфологического анализа, доступные в подсистеме *Russian POS Tagger*. Представлены: (i) пример инициализации подсистемы, (ii) параметры подключения к XML-RPC серверу *LemServer* и (iii) результаты работы подсистемы в составе системы GATE.

С помощью разработанной системы индексирования вики-текстов построены индексные базы Русской Википедии и Википедии на английском упрощённом языке, выполнено сравнение основных показателей индексных баз данных (чис-

4 Эмпирический закон Ципфа говорит о том, что частота употребления слова в корпусе текстов обратно пропорциональна его рангу в списке упорядоченных по частоте слов этого корпуса.

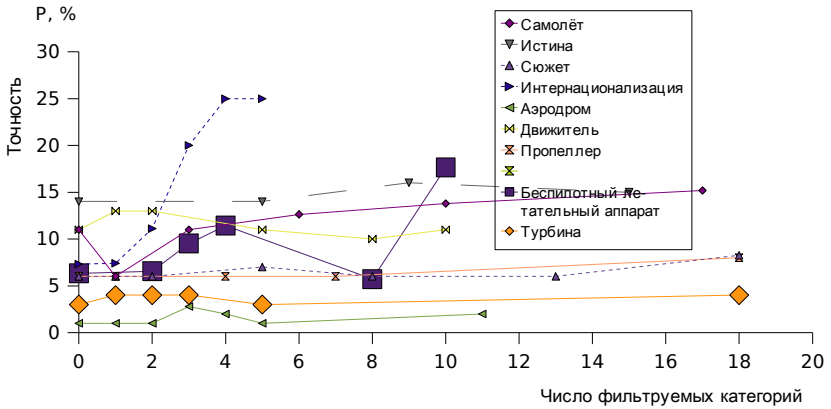


Рис. 6. Пример изменения точности⁵ поиска (P) АНITS алгоритма в зависимости от числа фильтруемых категорий

Турбина) с помощью АНITS алгоритма в зависимости от числа категорий.

Чёрный список категорий (blacklist) составляется экспертом и сужает пространство поиска. Например, фильтрация по категории *XX век* позволяет отсеять множество документов с заголовками: *1900*, *1901* и т. д. В эксперименте для фильтрации выбираются категории с максимальным числом слов, не являющихся семантически близкими заданному слову. Рис. 7 показывает, что при использовании категорий (для слова *Самолёт*) время работы вырастает, но точность поиска также увеличивается.

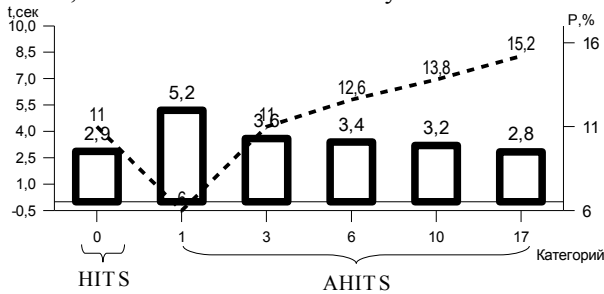


Рис. 7. Изменение времени работы *t* и точности поиска *P* АНITS алгоритма в зависимости от числа фильтруемых категорий для слова *Самолёт*

⁵ Точность — отношение числа семантически близких слов, найденных программой, к общему числу найденных слов. Списки СБС (синонимы, антонимы, гипонимы, гиперонимы, согипонимы, меронимы, холонимы) этих слов представлены в Русском Викисловаре (<http://ru.wiktionary.org>).

Русской Википедии соответствует граф, содержащий 171 тыс. вершин, 3.4 млн дуг (на 11.05.07). При поиске в графе АНITS алгоритм строит базовый набор с числом вершин 200-800, числом дуг 800-12 000 (для слова *Самолёт*). Указан диапазон вершин и дуг, поскольку изменение фильтруемых категорий меняет число вершин, включаемых в базовый набор. Таким образом, рис. 7 обобщает результаты шести опытов с разными размерами базовых наборов, построенных для слова *Самолёт*.

Основные отличия НITS и АНITS алгоритмов – не учёт и учёт категорий соответственно. При числе категорий ноль (первый вертикальный ряд на рис. 7) работа АНITS алгоритма (по скорости и точности поиска) соответствует работе НITS алгоритма. Это позволяет сравнить НITS и АНITS алгоритмы. Сравнение для указанных девяти слов показало, что работа АНITS алгоритма медленнее НITS алгоритма в среднем на 52%, а точность поиска АНITS алгоритма выше на 33% (рис. 6).

Выполнена численная оценка положительного влияния эвристики на качество строящегося автоматически списка семантически близких слов из Русской Википедии (коэффициент Спирмена показал улучшение результата в среднем на 41.9% после её применения). Суть эвристики в том, чтобы не включать в корневой и в базовый набор те энциклопедические статьи, названия которых не содержат пробелы (то есть состоят из одного слова).

Предложенная модификация коэффициента Спирмена позволила оценить

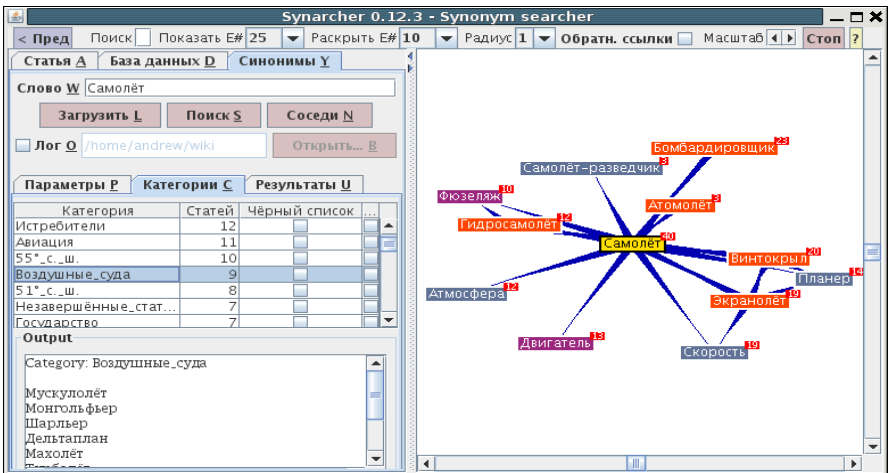


Рис. 8. Результаты поиска семантически близких слов для слова *Самолёт* в виде таблицы (со списком категорий), текста и графа

чувствительность результатов АНITS алгоритма к изменению параметров поиска с помощью экспериментов. Для ряда слов из Русской Википедии (например, *Самолёт*) точность поиска было достаточно стабильной (значение стандартного отклонения коэффициента Спирмена равно 4.41), что избавляет пользователя от необходимости тщательно подбирать параметры поиска. Для более часто употребляемого (в данном корпусе текстов) слова *Сюжет* качество результата оказалось в большей степени зависимым от входных параметров алгоритма (значение стандартного отклонения коэффициента Спирмена составило 95.97). На рис. 8 представлены результаты поиска семантически близких слов для слова *Самолёт* в виде списка категорий (в левой части экрана в центре), списка статей для выбранной категории («Воздушные суда») и графа.

ЗАКЛЮЧЕНИЕ

Диссертационное исследование выполнено в соответствии с положениями п.п. 9 и 17 областей исследований паспорта специальности 05.13.11: разработанные алгоритмы решают задачи семантического анализа текстовой информации на основе рейтинга текстов. Разработаны математическое и программное обеспечение для автоматизированного построения упорядоченного списка семантически близких слов. Результаты таковы:

1. Предложен подход к поиску семантически близких слов на основе учёта метаинформации (ключевые слова, категории, гиперссылки) в проблемно-ориентированном корпусе, содержащем два типа текстовых документов (статья и категория) и два типа отношений: иерархические отношения (родо-видовые и часть – целое) и гиперссылки.
2. НITS алгоритм адаптирован к поиску семантически близких слов (в корпусе текстовых документов с гиперссылками и категориями) на основе (1) нового способа построения корневого набора документов и (2) механизма иерархической кластеризации, позволяющего объединять слова в смысловые группы.
3. Спроектирована клиент-серверная архитектура программного комплекса поиска семантически близких слов с возможностью оценки списков слов на основе удалённого доступа к тезаурусам (WordNet, Moby) и модифицированного коэффициента Спирмена.
4. Разработан программный комплекс поиска семантически близких слов в проблемно-ориентированном корпусе текстов с динамической визуализацией результатов поиска.
5. Спроектирована архитектура и реализована система индексирования вики-текстов.

ОСНОВНЫЕ ПОЛОЖЕНИЯ ДИССЕРТАЦИИ ОПУБЛИКОВАНЫ В РАБОТАХ

В рецензируемых журналах из списка ВАК:

1. Крижановский, А.А. Автоматизированный поиск семантически близких слов на примере авиационной терминологии // *Автоматизация в промышленности*. – 2008. – Т. 4. – С. 16–20. (0,47 а.л.).
2. Крижановский, А.А. Формирование контекста задачи для интеллектуальной поддержки принятия решений / А.В. Смирнов, М.П. Пашкин, Н.Г. Шилов, Т.В. Левашова, А.А. Крижановский // *Фундаментальные основы информационных технологий и систем. // Труды Института системного анализа РАН*. – М.: ИСА РАН, 2004. – Т. 9. – С. 125–188. (0,56 а.л.).

В других изданиях:

3. Крижановский, А.А. Оценка результатов поиска семантически близких слов в Википедии // *Труды СПИИРАН*. Вып. 5. — СПб.: Наука, 2007 – С. 113–116. (0,24 а.л.).
4. Крижановский, А.А. Автоматизированное построение списков семантически близких слов на основе рейтинга текстов в корпусе с гиперссылками и категориями // *Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006»*. Бекасово, 2006. – С. 297–302. (0,55 а.л.).
5. Krizhanovsky, A. Synonym search in Wikipedia: Synarcher / A. Krizhanovsky // In: *11-th International Conference "Speech and Computer"*. Russia, St. Petersburg, 2006. – pp. 474–477. (0,64 а.л.).
6. Krizhanovsky, A. Context-sensitive access to e-document corpus / A. Smirnov, T. Levashova, M. Pashkin, N. Shilov, A. Krizhanovsky, A. Kashevnik, A. Komarova // *Труды международной конференции «Корпусная лингвистика–2006»*. – СПб.: Изд-во С.-Петерб. ун-та, 2006. – С. 360–364. (0,18 а.л.).
7. Krizhanovsky, A. Ontology-based users and requests clustering in customer service management system / A. Smirnov, M. Pashkin, N. Chilov, T. Levashova, A. Krizhanovsky, A. Kashevnik // In: *Autonomous Intelligent Systems: Agents and Data Mining*. Springer-Verlag GmbH, Lecture Notes in Computer Science, 2005, Vol. 3505, 231-246. (0,98 а.л.).
8. Krizhanovsky, A. Free text user request processing in the system “KSNet” / A. Smirnov, M. Pashkin, N. Chilov, T. Levashova, A. Krizhanovsky // *Proceedings of the 9th International Conference “Speech and Computer”*. Russia, St.Petersburg, 2004. – pp. 662–665. (0,36 а.л.).