На правах рукописи

Соченков Илья Владимирович

РЕЛЯЦИОННО-СИТУАЦИОННЫЕ СТРУКТУРЫ ДАННЫХ, МЕТОДЫ И АЛГОРИТМЫ РЕШЕНИЯ ПОИСКОВО-АНАЛИТИЧЕСКИХ ЗАДАЧ

Специальность: 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата физико-математических наук

Москва, 2014

Работа выполнена в лаборатории динамических интеллектуальных систем федерального государственного бюджетного учреждения науки Института системного анализа Российской академии наук.

Научный руководитель: доктор физико-математических наук, профессор
Осипов Геннадий Семенович

Официальные оппоненты: ***Аншаков Олег Михайлович***
доктор физико-математических наук, профессор, профессор кафедры математики, логики и интеллектуальных систем в гуманитарной сфере отделения интеллектуальных систем в гуманитарной сфере федерального государственного бюджетного учреждения высшего профессионального образования «Российский государственный гуманитарный университет»

Чеповский Андрей Михайлович
кандидат технических наук, доцент,
доцент кафедры анализа данных и искусственного интеллекта федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский университет "Высшая школа экономики"»

Ведущая организация: федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Национальный исследовательский университет "МЭИ"»

Защита состоится «16» октября 2014 г. в 16 часов 00 минут на заседании диссертационного совета Д 002.017.02 в федеральном государственном бюджетном учреждении науки Вычислительном центре им. А.А. Дородницына Российской академии наук по адресу: 119333, Москва, ул. Вавилова, д. 40 (конференц-зал).

С диссертацией можно ознакомиться в библиотеке ВЦ РАН и на официальном сайте ВЦ РАН: http://www.ccas.ru/

Автореферат разослан « 28 » августа 2014 г.

Ученый секретарь
диссертационного совета Д 002.017.02,
доктор физико-математических наук  Рязанов В.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Развитие Интернета привело к росту объёмов доступной информации, которая может быть использована при решении важных задач в ходе научно-исследовательской и экспертной деятельности, для поддержки принятия решений в научно-технической, социальной и других сферах. Анализ этой информации и её использование при принятии стратегических решений даёт преимущество в развитии экономики, науки и технологий. Поисково-аналитическая обработка информации в условиях динамично растущего Интернета не может быть выполнена без автоматизированных информационно-аналитических систем (ИАС).

Современные ИАС включают в себя инструменты решения следующих задач:

- полнотекстовый поиск документов, релевантных информационной потребности пользователя, сформулированной в виде набора ключевых слов, осмысленной фразы, предложения или вопроса на естественном языке (ЕЯ) с дополнительными ограничениями на метаданные, которые могут задаваться как в текстовом, так и нетекстовом виде;
- реферирование результатов полнотекстового поиска, а также отдельных документов;
- поиск текстовых заимствований в коллекциях документов;
- поиск содержательно и тематически близких документов, включая тематическую кластеризацию и классификацию;
- извлечение из текстов ЕЯ структурированных данных и фактов, установление зависимостей и связей между ними, например, выявление отзывов о товарах и услугах и анализ тональности высказанных мнений.

В центре внимания пользователей ИАС находится именно текстовая информация и её содержание. Многие факторы, успешно используемые при ранжировании результатов поиска в поисковых машинах Интернета, не применимы в ИАС. Поэтому современные исследования в области информационно-аналитической обработки текстов направлены на развитие методов, основанных на анализе лингвистической информации. При этом текст и составляющие его элементы характеризуется лексико-морфологическими, синтаксическими и семантическими признаками.

В настоящее время созданы методы лингвистического анализа текстов и разработаны программные системы, позволяющие автоматически выполнять морфологический, синтаксический и семантический анализ предложений текста: AOT[1], Solarix[2], NLTK[3], FreeLing[4] и другие. Вычислительная

[1] Автоматическая Обработка Текста (АОТ). / [Электронный ресурс] URL: http://www.aot.ru (дата обращения 23.01.2014)

эффективность этих систем и уровень качества лингвистического анализа позволяют применять их для обработки больших коллекций текстов. Однако в существующих ИАС, как правило, не применяются наукоёмкие методы лингвистического анализа текстов, поскольку в настоящее время отсутствуют эффективные методы хранения и обработки лингвистической информации, необходимой для решении поисково-аналитических задач, например, морфологических признаков, синтаксических связей, категориально-семантических значений (ролей) и семантических отношений на структурах текстов. Реляционные базы данных и стандартные инвертированные индексы[5] не позволяют эффективно хранить и совместно использовать эту лингвистическую информацию, в то время как использование этой информации обеспечивает повышение качества решения задач аналитической обработки больших коллекций текстовых документов. В настоящей диссертационной работе предложен метод оценки сходства текстов с использованием лексико-морфологической, синтаксической и семантической информации, а также структуры данных и алгоритмы информационного поиска на основе этого метода, обладающие большей эффективностью и обеспечивающие более высокое качество результатов, нежели существующие методы решения поисково-аналитических задач в ИАС, что свидетельствует об **актуальности темы исследования**.

Предмет исследования – методы оценки сходства текстов с использованием лексико-морфологической, синтаксической и семантической информации; метод поиска информации на основе оценки сходства текстов, а также структуры данных и алгоритмы, реализующие указанные методы.

Целью исследования является повышение качества (полноты и точности) и вычислительной эффективности решения поисково-аналитических задач за счёт разработки и применения метода оценки сходства текстов, учитывающего лексико-морфологическую, синтаксическую и семантическую информацию, и создания структур данных и алгоритмов информационного поиска, реализующих этот метод.

---

[2] Solarix: Компьютерная лингвистика. / [Электронный ресурс] URL: http://www.solarix.ru/ (дата обращения 05.03.2014)

[3] Natural Language Toolkit. / [Электронный ресурс] URL: http://nltk.org/ (дата обращения 05.03.2014).
Bird S. Natural Language Processing with Python. / O'Reilly Media Inc, 2009

[4] Freeling: An Open Source Suite Of Language Analyzers. / [Электронный ресурс] URL: http://nlp.lsi.upc.edu/freeling/ (дата обращения 05.03.2014)

[5] Justin Zobel, Alistair Moffat. «Inverted files for text search engines». /ACM Computing Surveys (CSUR). Vol. 38, №2, 2006. Article 6. doi:10.1145/1132956.1132959

Задачи исследования:

1. Разработка лексико-морфологических, синтаксических и семантических критериев оценки сходства текстов, а также метода оценки сходства текстов на основе этих критериев.
2. Разработка модели данных, предназначенной для исследования свойств структур данных и алгоритмов поисково-аналитической обработки текстовой информации.
3. Разработка и программная реализация структур данных, необходимых для решения задачи оценки сходства текстов и предназначенных для представления, хранения и обработки лексико-морфологической, синтаксической и семантической информации, являющейся результатом компьютерного лингвистического анализа текстов.
4. Разработка и программная реализация алгоритмов информационного поиска на основе многокритериальной оценки сходства текстов.
5. Экспериментальные исследования разработанных структур данных, алгоритмов формирования инвертированных индексов и многокритериальной оценки сходства текстов.

Для решения поставленных задач применены следующие **методы исследования**:

1. Методы теории множеств и алгебры логики.
2. Методы объектно-ориентированного проектирования программного обеспечения.
3. Методы исследования качества результатов информационного поиска.

В ходе решения поставленных задач получены следующие **новые научные результаты**:

1. Разработаны лексико-морфологические, синтаксические и семантические критерии оценки сходства текстов, а также метод многокритериальной оценки сходства текстов.
2. Предложена и исследована модель данных, предназначенная для анализа свойств структур данных и алгоритмов поисково-аналитической обработки текстовой информации.
3. Разработаны структуры данных инвертированного поискового индекса, предназначенные для хранения и обработки лексико-морфологической, синтаксической и семантической информации, являющейся результатом компьютерного лингвистического анализа текстов документов. Эти структуры данных применяются для эффективного решения задач информационного поиска, в частности, многокритериальной оценки сходства текстов.
4. Разработаны и исследованы следующие алгоритмы поисково-аналитической обработки текстовой информации:

- алгоритм построения инвертированного поискового индекса коллекций документов,
- алгоритм поиска информации по запросу, основанный на разработанном методе многокритериальной оценки сходства текстов, реализующий следующие типы поиска с учётом метаданных документов: поиск по ключевым словам, фразовый поиск, семантический и вопросно-ответный поиск.

5. Теоретически исследованы свойства разработанных структур данных и алгоритмов поисково-аналитической обработки текстовой информации, в том числе получены оценки вычислительной сложности и доказаны утверждения, обосновывающие корректность указанных алгоритмов.

Теоретическая значимость. Разработанные метод многокритериальной оценки сходства текстов, представление текстовой информации, алгоритмы и структуры данных информационного поиска служат основой решения ряда поисково-аналитических задач. Методы семантического аннотирования текстов и поиска потенциально некорректных заимствований (рассмотрение которых выходит за рамки диссертационной работы) опираются на разработанный метод многокритериальной оценки сходства текстов и используют алгоритмы и структуры данных, предложенные и исследованные в настоящей работе.

Практическая значимость. Разработанный метод многокритериальной оценки сходства текстов, в основе которого лежит сопоставление лексико-морфологической, синтаксической, семантической информации, а также алгоритмы и структуры данных для решения поисково-аналитических задач нашли применение в ИАС. Программная реализация указанных алгоритмов и структур данных ориентирована на обработку больших коллекций текстовых документов в ИАС для информационной поддержки аналитической деятельности в научно-технической сфере.

Результаты исследований по теме диссертационной работы использованы при выполнении научно-исследовательских работ по следующим проектам Минобрнауки РФ, программам ОНИТ РАН и грантам РФФИ:

1. «Создание методов и программных средств выявления перспективных направлений научных исследований в России и за рубежом по данным из открытых источников на основе потребностей реального сектора экономики и обеспечения конкурентных позиций отечественных производителей на перспективных рынках инновационных товаров и услуг и созданных научно-технических заделов» (в рамках ФЦП «Научные и научно-педагогические кадры инновационной России», ГК № *16.740.11.0753*, 2011–2013 г.г.).
2. «Создание программного комплекса информационно-аналитической поддержки научно-технической деятельности на основе вычислительного

семантического поиска и анализа неструктурированной текстовой информации» (в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007—2013 годы», ГК № *07.551.11.4003*, 2011–2013 г.г.).

3. «Разработка вычислительных методов объективной оценки качества научно-технических документов на естественных языках» (в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007—2013 годы», ГК № *14.514.11.4018*, 2011–2013 г.г.).
4. «Исследование и разработка методов и алгоритмов анализа связанности сложно-структурированных данных в научно-технической сфере» (в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007—2013 годы», ГК № *14.514.11.4024*, 2011–2013 г.г.).
5. «Исследование и разработка программного обеспечения понимания неструктурированной текстовой информации на русском и английском языках на базе создания методов компьютерного полного лингвистического анализа» (в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007—2013 годы», ГК № *07.514.11.4134*, 2011–2013 г.г.).
6. «Развитие методов анализа полуструктурированной информации и моделирования целенаправленного поведения» (в рамках ФЦП «Научные и научно-педагогические кадры инновационной России», ГК № *П952*, 2009–2011 г.г.).
7. Проект «Система высокоточного интеллектуального поиска, индексации и анализа информации для поддержки принятия решений» (проект ПР4 «Исследование и разработка параллельных алгоритмов анализа больших объемов текстовой информации из глобальной сети и алгоритмов принятия решений на основе когнитивных методов» программы «ТРИАДА», 2006–2007 г.г.).
8. «Развитие методов и программных средств многоязычного семантического поиска» (в рамках проекта *2.6* ОНИТ РАН 2009–2011 г.г.).
9. «Развитие методов и технологии семантического поиска и анализа научных публикаций Exactus Expert» (в рамках проекта *2.9* ОНИТ РАН 2012–2013 г.г.).

10. «Разработка и исследование структур данных и алгоритмов поисково-аналитической обработки текстовой информации» (в рамках проекта *14-07-31149\14 мол_а* РФФИ 2014–2015 г.г.)

Созданное программное обеспечение (ПО), включающее программную реализацию структур данных и алгоритмов поисково-аналитической обработки текстовой информации, внедрено в следующих системах:

- электронная библиотека международных клинических руководств[6] в Медицинском центре Банка России;
- портал «Руконт» – национальный цифровой ресурс[7] в виде информационно-поисковых сервисов портала;
- информационно-аналитическая система Exactus Expert[8] и поисковая машина Exactus[9].

Достоверность результатов подтверждена строгой математической формализацией основных положений диссертационного исследования и доказательствами теоретических утверждений, а также результатами экспериментальных исследований разработанных программных средств, реализующих предложенные методы, структуры данных и алгоритмы.

Апробация результатов исследования. Основные положения диссертации докладывались и обсуждались на следующих конференциях и семинарах:

- XIII национальная конференция по искусственному интеллекту с международным участием (КИИ: Россия, Белгород, Белгородский государственный технологический университет, 2012 г.);
- European Intelligence and Security Informatics Conference (IEEE EISIC: 2011 (Greece, Athens);

---

[6] Назаренко Г.И., Плотникова В.А., Смирнов И.В., Соченков И.В., Тихомиров И.А. Программные средства создания и наполнения полнотекстовых электронных библиотек / «Электронные библиотеки: перспективные методы и технологии, электронные коллекции: XII Всероссийская научная конференция RCDL’ 2010, Казань, Россия - 2010. – С38-42.

[7] Национальный цифровой ресурс Руконт - межотраслевая электронная библиотека на базе технологии Контекстум / [Электронный ресурс] URL: http://www.rucont.ru/ (дата обращения 16.02.2014)

[8] Osipov, G.; Smirnov, I.; Tikhomirov, I. and Shelmanov, A. Relational-Situational Method for Intelligent Search and Analysis of Scientific Publications. / In Proceedings of the Workshop on Integrating IR technologies for Professional Search Moscow, Russian Federation, March 24, 2013, p.57-64

[9] Завьялова О.С., Киселёв А.А., Осипов Г.С., Смирнов И.В., Тихомиров И.А., Соченков И.В. Система интеллектуального поиска и анализа информации Exactus на РОМИП-2010 / Труды российского семинара по оценке методов информационного поиска РОМИП'2010. - Казань: Казан. ун-т, 2010. С49-69.

- Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем – всероссийская конференция с международным участием в 2010, 2011 г.г. (Россия, Москва, Российский университет дружбы народов)
- XII Всероссийская научная конференция RCDL’ 2010: «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Россия, Казань, 2010 г.)
- Российский семинар по оценке методов информационного поиска в 2008–2010 г.г.;
- Семнадцатая международная Конференция "Крым 2010" (Россия, Геленджик, 2010 г.).

Информационные системы, содержащие в своём составе программную реализацию разработанных структур данных и алгоритмов поисково-аналитической обработки текстовой информации, представлены на российских и международных выставках программного обеспечения и информационных технологий SofTool (в 2010–2013 г.г.) и CeBIT (в 2008–2014 г.г.).

Публикации. Всего по теме исследования опубликовано 14 работ: 6 из них в рецензируемых журналах из списка ВАК РФ, 8 в материалах российских и международных конференций. Получен патент Российской Федерации на изобретение и 3 свидетельства о государственной регистрации программ для ЭВМ.

Структура и объем работы. Диссертация состоит из введения, четырёх глав, заключения, списка сокращений и условных обозначений, а также списка использованной литературы. Диссертация содержит 148 страниц, 25 рисунков, 2 таблицы, 152 источника в списке используемой литературы.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы, определен предмет исследования, сформулированы цель и задачи исследования, научная новизна, практическая значимость полученных результатов, а также приведены данные о структуре и объеме диссертации.

В **первой главе** приведен обзор методов компьютерного анализа и информационного поиска текстовой информации. Обзор включает в себя рассмотрение различных моделей, предназначенных для представления и анализа текстовой ЕЯ информации в системах информационного поиска. Рассматриваются существующие методы компьютерного анализа текстов: токенизация, морфологический анализ и лемматизация (определение леммы[10] по словоформе некоторой лексемы[11]), синтаксический анализ и семантический анализ (включающий в себя установление семантических значений («ролей») слов и понятий в текстах (semantic role labeling, или SLR) и разрешение кореференции). Описаны также программные системы, реализующие рассмотренные методы.

Далее рассматриваются структуры данных (инвертированные поисковые индексы (ИПИ)), применяемые для представления текстовой информации в задачах информационного поиска, и методы ранжирования результатов информационного поиска: булев поиск (Boolean search), ранжирование на основе TF-IDF и их модификации. В заключительной части главы обоснована необходимость разработки и исследования метода оценки сходства текстов, использующего результаты лексико-морфологического, синтаксического и семантического анализа, поставлена цель и сформулированы задачи исследования.

Вторая глава посвящена методу многокритериальной оценки сходства текстов на основе сопоставления лексико-морфологической, синтаксической и семантической информации. В первом параграфе представлено разработанное представление текстовой информации для решения задачи многокритериальной оценки сходства текстов.

Рассматриваются универсальное множество D лемм – словарь нормальных форм всех лексем ЕЯ, Ф – множество всевозможных форм (словоформ) всех лексем ЕЯ, множество текстов $E = \{\tau\}$ и произвольный текст $\tau \in E$. Этот текст содержит конечное множество словоупотреблений $W^{\tau} = \{w_i\}$. Далее определены:

---

[10] Лемма – каноническая (словарная) форма слова

[11] Лексема – совокупность парадигматических форм одного слова, как элемента лексики естественного языка. Одной лексеме, как правило, соответствует одна статья в словаре

– Бинарное отношение δ^τ на множестве словоупотреблений W^τ текста τ и множестве D нормальных форм лексем: $\delta^\tau \subseteq W^\tau \times D$, сопоставляющее каждому словоупотреблению его нормальную форму: $\forall w \in W^\tau \exists d \in D :< w,d >\in \delta^\tau$. При этом, вообще говоря, у каких-либо словоупотреблений имеется более одного варианта нормализации (для некоторых $w \in W^\tau \ \exists d' \in D, d'' \in D\, d' \neq d'' \& < w,d' >\in \delta^\tau \& < w,d'' >\in \delta^\tau$).
– Бинарное отношение ψ^τ на множестве словоупотреблений W^τ текста τ и множестве Ф всевозможных форм различных лексем: $\psi^\tau \subseteq W^\tau \times \Phi$. Это отношение определяет форму каждого словоупотребления в тексте: $\forall w \in W^\tau \exists g \in \Phi :< w,g >\in \psi \ \& \left(\exists g' \in \Phi :< w,g' >\in \psi \rightarrow g = g'\right)$.
– Конечное множество меток (тегов) $T=\{t_i\}$, и бинарное отношение $\theta^\tau \subseteq W^\tau \times T$, сопоставляющее каждому словоупотреблению некоторую метку (тег гипертекстовой разметки или разметки по метаданным, например, для словоупотреблений в заголовке текста, в списке авторов и т.п.): $\forall w \in W^\tau \exists t \in T :< w,t >\in \theta^\tau \ \& \left(\exists t' \in T :< w,t' >\in \theta^\tau \rightarrow t = t'\right)$.
– Числовая функция, задающая вес («значимость») меток: $\omega(t)$.
– Числовая функция, задающая вес словоупотребления в тексте: $\nu(w^\tau)$.

Разбиение множества словоупотреблений на предложения (возможно, сложные) задано в виде множества $S^\tau = \left\{ s_i \right\}$, где $\mathrm{s}_k = \left\{ w_{i_k},...,w_{j_k} \mid \forall w_x \in W^\tau \right\}$, где $i_k,...,j_k$ – последовательность индексов словоупотреблений в тексте, входящих в k-е предложение.

Для учёта синтаксических зависимостей между словоупотреблениями в тексте определены следующие объекты:

– Конечное множество типов синтаксических связей: SR^{12}.
– Бинарное отношение на множестве словоупотреблений: $\Sigma^\tau \subseteq W^\tau \times W^\tau$, представляющее всевозможные синтаксические связи между словоупотреблениями. Синтаксические структуры рассматриваются в виде деревьев, и считается, что в паре элементов $< w_i, w_j >$ первый элемент – $w_i \in W^\tau$ – соответствует главному словоупотреблению (ГС), а второй элемент – $w_j \in W^\tau$ – зависимому (подчинённому) словоупотреблению.

[12] А.Сокирко. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) / Дисс канд.т.н. // [Электронный ресурс] URL: http://www.aot.ru/docs/sokirko/sokirko-candid-1.html (дата обращения 23.01.2014)

– Семейство отношений $Syntax^{\tau} = \left\{ \Sigma_z^{\tau} \subseteq \Sigma^{\tau} \mid z \in SR^{\tau}, \bigcup_{z \in SR} \Sigma_z^{\tau} = \Sigma^{\tau} \right\}$ – разбиение множества Σ^{τ}, где каждое из отношений Σ_z^{τ} определяет тип синтаксической связи $z \in SR$ для пар словоупотреблений $< w_i, w_j > \in \Sigma_z^{\tau}$.

Для представления семантической информации в текстах определены:

– Конечное множество категориально-семантических значений синтаксем[13] *Roles*, называемых для краткости "семантическими значениями".
– Бинарное отношение $SemRoles^{\tau} \subseteq W^{\tau} \times Roles$, сопоставляющее словоупотреблениям текста семантические значения[14] соответствующих синтаксем[15]. Каждое словоупотребление может иметь 0 и более семантических значений в тексте.
– Множество типов отношений *R* на множестве значений синтаксем[16].
– Бинарное отношение $\Omega^{\tau} \subseteq W^{\tau} \times W^{\tau}$, определяющее семантически связанные словоупотребления в тексте[17].
– Семейство отношений $SemRels^{\tau} = \left\{ \Omega_x^{\tau} \subseteq \Omega^{\tau} \mid x \in R, \bigcup_{x \in R} \Omega_x^{\tau} = \Omega^{\tau} \right\}$ – разбиение множества Ω^{τ}, где каждое отношение Ω_x^{τ} определяет тип семантической связи $x \in R$ для пар словоупотреблений $< w_i, w_j > \in \Omega_x^{\tau}$, где $w_i \in W^{\tau}$ и $w_j \in W^{\tau}$.

Разработанное представление текстовой информации является модификацией реляционно-ситуационной модели (РСМ) текстовой информации, предложенной Г.С. Осиповым и Г.А. Золотовой, для решения информационно-аналитических задач.

Во втором параграфе **второй главы** введены критерии оценки сходства текстов и описан метод многокритериальной оценки сходства текстов.

---

[13] Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка. – М. 2004. – 544 с.

[14] G. Osipov. Methods for extracting semantic types of natural language statements from texts. // In 10th IEEE International Symposium on Intelligent Control, Monterey, California, USA, 1995

[15] В общем случае в тексте синтаксема является некоторой синтаксической конструкцией. Например, синтаксема может быть представлена синтаксической группой (предложной или именной). При решении задачи информационного поиска будем связывать семантическое значение не с синтаксемой в целом, а с главным словом синтаксической конструкции, представляющей синтаксему.

[16] Осипов Г.С. Приобретение знаний интеллектуальными системами. // М.: Наука. Физматлит, 1997

[17] Осипов Г.С. Методы искусственного интеллекта. – М.: ФИЗМАТЛИТ, 2011. – 296 с.

Сходство эталонного текста $\varepsilon \in E$ и сопоставляемого с ним текста $\tau \in E$ оценивается по множествам предложений этих текстов: S^{ε} и S^{τ} соответственно. Для произвольных предложений $\mathrm{s}^{\varepsilon} \in S^{\varepsilon}$ и $\mathrm{s}^{\tau} \in S^{\tau}$ определено множество $N\left(\mathrm{s}^{\varepsilon}, \mathrm{s}^{\tau}\right) = \left\{ < w^{\varepsilon}, w^{\tau} > \in W^{\varepsilon} \times W^{\tau} \mid w^{\varepsilon} \in \mathrm{s}^{\varepsilon} \exists w^{\tau} \in \mathrm{s}^{\tau} \exists d \in D : < w^{\varepsilon}, d > \in \delta^{\varepsilon} \,\&\, < w^{\tau}, d > \in \delta^{\tau} \right\}$ – множество пар совпадающих по нормальной форме словоупотреблений, называемых *соответственными*.

Далее определены критерии оценки сходства предложений s^{ε} и s^{τ}.

1. ***Покрытие предложения-эталона*** s^{ε} ***предложением*** s^{τ} ***сопоставляемого текста*** τ:

$$I_1\left(\mathrm{s}^{\varepsilon}, \mathrm{s}^{\tau}\right) = \sum_{< w^{\varepsilon}, w^{\tau} > \in N\left(\mathrm{s}^{\varepsilon}, \mathrm{s}^{\tau}\right)} v(w^{\varepsilon}). \tag{1}$$

   В качестве функции для определения весов $v(w^{\varepsilon})$ может применяться классическая оценка информационной значимости на основе inverse document Frequency[18] (IDF) или характеристика тематической значимости[19] (ХТЗ), если для эталонного текста s^{ε} известна его тематическая принадлежность.

2. ***Общая оценка информационной значимости слов предложения-эталона*** s^{ε} ***в предложении*** s^{τ} ***сопоставляемого текста*** τ:

$$I_2\left(\mathrm{s}^{\varepsilon}, \mathrm{s}^{\tau}\right) = \sum_{< w^{\varepsilon}, w^{\tau} > \in N\left(\mathrm{s}^{\varepsilon}, \mathrm{s}^{\tau}\right)} f(w^{\varepsilon}, w^{\tau}) v(w^{\varepsilon}) v'(w^{\tau}). \tag{2}$$

   В этой формуле функционал $f(w^{\varepsilon}, w^{\tau})$ соответствует «штрафу» за несовпадение форм словоупотреблений w^{ε}, w^{τ}:

$$f(w^{\varepsilon}, w^{\tau}) = \begin{cases} 1, & \exists \mathrm{g} \in \Phi :< w^{\varepsilon}, g > \in \psi^{\varepsilon} \,\&\, < w^{\tau}, g > \in \psi^{\tau}, \\ f_0, & \text{в противном случае,} \end{cases} \tag{3}$$

   где $0 \le f_0 \le 1$ – параметр метода.

   В качестве весов $v(w^{\varepsilon})$ могут использоваться IDF или ХТЗ, а в качестве функции для определения весов $v'(w^{\tau})$ – классическая оценка term frequency[20] (TF) или её модификации[21].

---

[18] S.E. Robertson, C.J. van Rijsbergen and P.W. Williams. «Probabilistic models of indexing and searching». / In R.N. Oddy Information Retrieval Research, pp. 35-56, London, 1981. Butterworths.

[19] Р.Е. Суворов, И.В. Соченков. Определение связанности научно-технических документов на основе характеристики тематической значимости. // Искусственный интеллект и принятие решений. М.: ИСА РАН, №1, 2013. С.33-40

[20] Joachims, T. «A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization». / DTIC Document, 1996.

3. ***Оценка сходства предложения-эталона*** s^{ε} ***и предложения*** s^{τ} ***сопоставляемого текста*** τ ***на основе совпадения синтаксических структур***:

$$I_3\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)=\frac{\sum\limits_{<w^{\varepsilon},w^{\tau}>\in N_{Syn}\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)} v(w^{\varepsilon})}{\sum\limits_{w^{\varepsilon}\in\left\{w\in W^{\varepsilon}\mid\exists w'\in W^{\varepsilon}:<w,w'>\in\Sigma^{\varepsilon}\right\}} v(w^{\varepsilon})}. \quad (4)$$

В этой формуле $N_{Syn}\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)=\left\{<w^{\varepsilon},w^{\tau}>\in N\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)\mid\exists\tilde{w}^{\varepsilon}\in W^{\varepsilon},\exists\tilde{w}^{\tau}\in W,\exists z\in SR:\right.$ $\left.<\tilde{w}^{\varepsilon},\tilde{w}^{\tau}>\in N\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)\&<w^{\varepsilon},\tilde{w}^{\varepsilon}>\in\Sigma_z^{\varepsilon}\&<w^{\tau},\tilde{w}^{\tau}>\in\Sigma_z^{\tau}\right\}$ представляет собой множество пар соответственных словоупотреблений в эталонном предложении s^{ε} и сопоставляемом предложении s^{τ}, для которых совпадают (по нормальным формам лексем) главные $<w^{\varepsilon},w^{\tau}>\in N\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)$ и зависимые $<w^{\varepsilon},w^{\tau}>\in N\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)$ слова, а сами словоупотребления связаны в контексте эталонного и сопоставляемого предложения однотипными синтаксическими связями: $<w^{\varepsilon},\tilde{w}^{\varepsilon}>\in\Sigma_z^{\varepsilon}\&<w^{\tau},\tilde{w}^{\tau}>\in\Sigma_z^{\tau}$.

4. ***Оценка сходства предложения-эталона*** s^{ε} ***и предложения*** s^{τ} ***сопоставляемого текста*** τ ***на основе совпадения семантических значений***:

$$I_4\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)=\frac{\left|\rho\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)\right|}{\left|SemRoles^{\varepsilon}\right|}. \quad (5)$$

В этой формуле $\rho\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)=\left\{<w^{\varepsilon},a>\in SemRoles^{\varepsilon}\mid w^{\varepsilon}\in\mathrm{s}^{\varepsilon}\ \&\ a\in Roles\ \&\right.$ $\left.\exists w^{\tau}\in\mathrm{s}^{\tau}<w^{\varepsilon},w^{\tau}>\in N\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)\ \&<w^{\tau},a>\in SemRoles^{\tau}\right\}$ – множество таких словоупотреблений с семантическими значениями в предложении эталонного текста, что для них в сопоставляемом тексте имеются соответственные словоупотребления, которым приписаны такие же семантические значения.

5. ***Оценка сходства предложения-эталона*** s^{ε} ***и предложения*** s^{τ} ***сопоставляемого текста*** τ ***на основе семантических связей***:

$$I_5\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)=\frac{\sum\limits_{<w^{\varepsilon},w^{\tau}>\in N\left(\mathrm{s}^{\varepsilon},\mathrm{s}^{\tau}\right)}\left|SemR^{\varepsilon}_{w^{\varepsilon}}\cap SemR^{\tau}_{w^{\tau}}\right|}{\left|SemRoles^{\varepsilon}\right|}. \quad (6)$$

---

[21] Joaquin Perez-Iglesias. «Integrating the Probabilistic Model BM25: BM25F into Lucene», 2009 / [Электронный ресурс] URL: http://nlp.uned.es/~jperezi/Lucene-BM25/ (дата обращения 23.03.2014)

Где $SemR_w^\tau = \{a \in Roles \mid \exists w' \in W^\tau, \exists x \in R: < w, w' > \in \Omega_x^\tau \;\&\; < w', a > \in SemRoles^\tau\}$ – множество значений синтаксем, приписанных в тексте τ тем словоупотреблениям *w'*, которые связаны в этом тексте со словоупотреблением *w* всевозможными семантическими связями.

Общая оценка сходства предложения s^ε текста-эталона и предложения s^τ сопоставляемого текста определяется взвешенной суммой:

$$Sim(s^\varepsilon, s^\tau) = \sum_{n=1}^{5} \lambda_n I_n(s^\varepsilon, s^\tau). \quad (7)$$

В формуле (7) набор $L=\{\lambda_i | i=1..5\}$ ($\sum_{n=1}^{5} \lambda_n = 1$) задаёт набор параметров метода, определяющих вклад каждого из критериев оценки сходства в итоговую величину. Из всех возможных предложений s^τ в сопоставляемом тексте τ выбираются те, которые наилучшим образом (в смысле максимизации оценки (7)) соответствуют предложению эталона s^ε:

$$Y(s^\varepsilon, \tau) = \max_{s^\tau \in S^\tau} \{Sim(s^\varepsilon, s^\tau)\}. \quad (8)$$

Общая оценка сходства текста-эталона ε по отношению к тексту τ выражается следующей формулой:

$$X(\varepsilon, \tau) = \sum_{s^\varepsilon \in S^\varepsilon} Y(s^\varepsilon, \tau). \quad (9)$$

В завершение второго параграфа **второй главы** сформулированы и доказаны следующие утверждения, представляющие интерес с точки зрения практического применения величины, заданной формулой (9), для оценки сходства текстов в ИАС и ранжирования результатов поиска.

Утверждение 2.1. $0 \le Sim(s^\varepsilon, s^\tau) \le 1$.

Утверждение 2.2. $0 \le X(\varepsilon, \tau) \le 1$.

В третьем параграфе **второй главы** рассмотрено применение разработанного метода оценки сходства текстов для решения задач полнотекстового поиска по ключевым словам и словосочетаниям, а также фразового, семантического и вопросно-ответного поиска. Под фразовым поиском подразумевается такой вид поиска, когда фрагмент предложения (фраза) запроса и фрагменты предложений в текстах найденных документов совпадают по синтаксическим структурам (но при этом они могут отличаться порядком и формой словоупотреблений). При семантическом и вопросно-ответном поиске предложения запроса и найденных текстов должны совпадать по семантическим значениям и связям словоупотреблений. В этом же параграфе описана процедура расширения поискового запроса (в общем случае – произвольного эталонного текста) с помощью концептов из тезауруса,

связанных некоторыми отношениями с фразами запроса. Показано, как в рамках предложенного представления учитывается кореферентность.

В заключение **второй главы** на основе многокритериальной оценки сходства текстов предложен метод ранжирования результатов поиска и сделан вывод о необходимости разработки новых алгоритмов и структур данных информационного поиска, поскольку известные структуры данных и алгоритмы не позволяют эффективно работать с предложенным представлением текстовой информации.

Третья глава посвящена разработке и исследованию модели данных, структурам данных и алгоритмам для решения поисково-аналитических задач. Представлены разработанные структуры данных, необходимые для представления информации о текстовых документах в ИАС, структуры инвертированных индексов, а также алгоритмы их построения на этапе индексирования документов.

Индекс текста документа (ИТД) представляет собой последовательность (массив) элементов данных (ЭД). Каждый ЭД имеет фиксированный размер в памяти и содержит набор полей данных, представляющих лексико-морфологическую, синтаксическую и семантическую информацию словоупотреблений текста документа. В ИТД присутствуют ЭД нескольких видов, различающиеся (в зависимости от вида ЭД) конкретным набором полей данных. Каждому словоупотреблению текста соответствует один или более ЭД различных видов в зависимости от информации, сопоставленной этому словоупотреблению. В реляционно-ситуационной модели семантические значения и связи назначаются не всем синтаксическим структурам, а только именным синтаксемам. Поэтому ЭД, содержащие поля данных с семантическими значениями и связями, присутствуют в ИТД только для таких словоупотреблений, которые входят в состав именных синтаксем, имеющих семантические значения и участвующих в семантических связях. Такой подход позволяет уменьшить количество памяти, необходимой для хранения информации в индексных структурах данных.

Для эффективной реализации процедуры поиска разработан ИПИ, ориентированный на хранение информации о текстах масштабных коллекций текстовых документов с возможностью поиска с учётом синтаксической и семантической информации. ИПИ реализуется в виде хеш-таблицы. В качестве ключа ИПИ выступает идентификатор нормальной формы лексемы (ИНФЛ). Каждому ключу сопоставлена последовательность σ-ЭД – пост-лист. Схема σ-ЭД ИПИ представлена на рисунке 1. Каждый σ-ЭД занимает в памяти ЭВМ 8 байт, что позволяет эффективно реализовать представление σ-последовательностей в памяти ЭВМ в виде массивов.

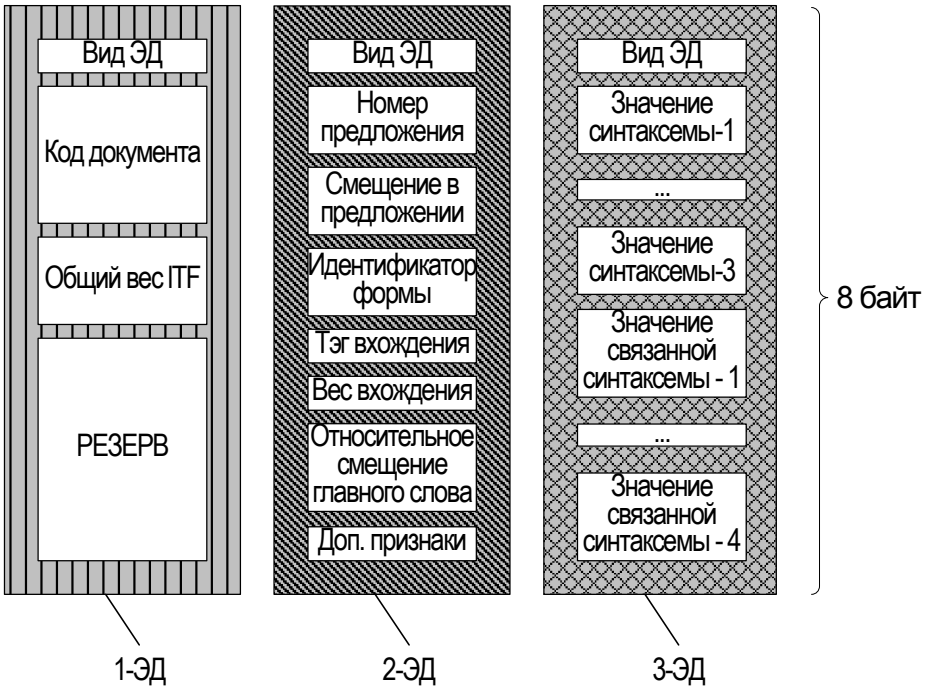

Рисунок 1 – Схема элементов данных разработанного инвертированного поискового индекса

Для описания модели данных ИПИ в диссертации введены следующие определения.

Определение 3.1. Последовательность σ-ЭД, относящихся к употреблениям некоторой лексемы в тексте отдельно взятого документа назовём *σ-цепочкой*; σ-цепочка строится по следующим правилам:

1. В начале цепочки размещается 1-ЭД.
2. За ним следует 2-ЭД, соответствующий первому (по порядку) употреблению лексемы в тексте.
3. За ним опционально следует 3-ЭД, если с рассматриваемым употреблением лексемы связана некоторая семантическая информация.
4. Далее следуют чередующиеся 2-ЭД и 3-ЭД (опционально), соответствующие остальным употреблениям рассматриваемой лексемы в тексте.

Определение 3.2. Для представления информации обо всех употреблениях некоторой лексемы всех документах коллекции применяется последовательность σ-цепочек – *σ-последовательность*. σ-последовательность формируется с помощью конкатенации σ-цепочек по следующему правилу: σ-цепочка σ_1 предшествует σ_2 тогда и только тогда, когда код документа, представляемого σ-цепочкой σ_1 меньше кода документа, представляемого σ-цепочкой σ_2.

Определение 3.3. Объект, указывающий на текущий обрабатываемый ЭД в последовательности элементов, называется итератором.

На классе σ-цепочек индуцируется линейный порядок (*σ-правило-1*) и частичный порядок на классе σ-ЭД внутри σ-цепочки (*σ-правило-2*). В

совокупности *σ-правило-1* и *σ-правило-2* вводят отношение частичного порядка на классе σ-ЭД в σ-последовательностях – рисунок 2.

Рисунок 2 – Пример упорядочения σ-ЭД в σ-последовательностях

Это отношение частичного порядка, однако, не позволяет сравнивать σ-ЭД двух σ-последовательностей, что требуется, например, в алгоритме слияния (merge) двух упорядоченных множеств. Вводимое Для устранения этого недостатка введено следующее *σ-правило-3*: из двух сравниваемых σ-ЭД m_1 и m_2, входящих в некоторую σ-последовательность ($m_1, m_2 \in \sigma$), ЭД m_1 предшествует m_2 ($m_1 < m_2$) тогда и только тогда, когда:

1. m_1 является 3-ЭД, а m_2 – 2-ЭД.
2. m_1 является 3-ЭД, а m_2 – 1-ЭД.
3. m_1 является 2-ЭД, а m_2 – 1-ЭД.
4. m_1 и m_2 являются 1-ЭД, и при этом код документа в m_1 меньше кода документа в m_2: $doc\_id(m_1) < doc\_id(m_2)$;
5. m_1 и m_2 являются 2-ЭД, и при этом (аналогично *σ-правилу-2*)
   a. Номер предложения в тексте документа, в котором содержится словоупотребление, представленное 2-ЭД m_1, меньше номера предложения в тексте, содержащего словоупотребление, представленное 2-ЭД m_2: $sent(m_1) < sent(m_2)$;
   b. Номера вышеуказанных предложений совпадают ($sent(m_1) = sent(m_2)$), а смещение (от начала предложения) словоупотребления, представленного 2-ЭД m_1, меньше смещения словоупотребления, представленного 2- ЭД m_2: $sofs(m_1) < sofs(m_2)$.

Это правило задаёт отношение частичного порядка на ЭД σ-последовательностей: рассмотрены 5 вариантов сравнения σ-ЭД друг с другом из 6 возможных, причём два 3-ЭД не сравнимы между собой.

Далее рассмотрен класс алгоритмов S , таких что всякий алгоритм A класса S обладает рядом свойств. К классу алгоритмов S относится, в частности, алгоритм слияния σ-последовательностей. Схематический пример обработки σ-последовательностей алгоритмом класса S приведён на рисунке 3.

Справедливо следующее утверждение.

Утверждение 3.1. При обработке σ-последовательностей σ_1 и σ_2 любым алгоритмом *A* класса *S* не происходит сравнения 3-ЭД между собой.

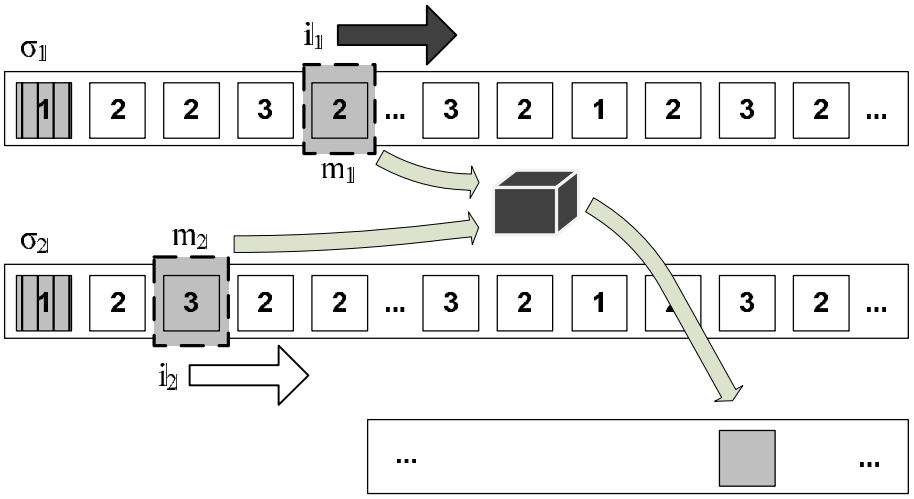

Рисунок 3 – Пример обработки σ-последовательностей алгоритмом класса *S*

Это утверждение означает, что к σ-последовательностям применим алгоритм слияния (и аналогичные алгоритмы), использующие *σ-правило-3* в качестве отношения порядка.

Далее в **третьей главе** описаны алгоритмы формирования поисковых индексов и наполнения базы ИПИ. Описана модель данных буфера для хранения промежуточных ИТД в памяти, разработаны алгоритмы слияния буфера и ИПИ. Модель данных опирается на понятие ξ-последовательности, расширяющей понятие σ-последовательности. Введено отношение частичного порядка на ξ-ЭД, разработаны алгоритмы *фильтрации σ-последовательности*, *помещения ИТД в буфер ИПИ*, *обновления ИПИ* и доказаны утверждения об оценках сложности разработанных алгоритмов (**утверждения 3.2 – 3.4**).

Далее предложено представление поискового запроса в виде последовательности ЭД – индекс текста запроса (ИТЗ). Разработаны алгоритмы оценки релевантности и ранжирования результатов информационного поиска, использующие метод многокритериальной оценки сходства текстов и структурах данных ИПИ: *первичной фильтрации σ-последовательностей*, *фильтрации σ-последовательности по словоупотреблению запроса*. Для этих алгоритмов получены оценки вычислительной сложности (**утверждения 3.5, 3.6**).

Предложен алгоритм *предварительной оценки релевантности* найденных документов, позволяющий провести быструю оценку соответствия найденных словоупотреблений тексту запроса и на первом этапе исключить из дальнейшего рассмотрения нерелевантные результаты поиска. В основе алгоритма предварительной оценки релевантности лежит понятие ρ-последовательности – массив линейно упорядоченных ρ-ЭД. Алгоритм предварительной оценки релевантности использует алгоритм *преобразования σ-последовательности в ρ-последовательность* и алгоритм *оценки сходства двух*

множеств семантических значений, разработанные **в третьей главе**. Доказано **утверждение 3.7**, обосновывающие оценки вычислительной сложности алгоритма.

Для формирования множества документов, соответствующих поисковому запросу, разработаны следующие операции над двумя ρ-последовательностями: *слияния*, *условного слияния*, *одностороннего условного слияния*, *условного исключения*. Установлено свойство операции условного соединения (**утверждение 3.8**), разработаны алгоритмы, реализующие эти операции и получены оценки их вычислительной сложности (**утверждения 3.9**, **3.10**). Разработаны алгоритмы *поиска синтаксических связей в ρ-последовательности* и *фразовой фильтрации ρ-последовательности*. Получены оценки их вычислительной сложности (**утверждения 3.11**, **3.12**).

Далее в работе представлен общий *алгоритм формирования множества потенциально релевантных документов*, опирающийся на ранее разработанные алгоритмы. Показано, что общая оценка вычислительной сложности линейна относительно количества σ-ЭД в σ-последовательностях, извлечённых из ИПИ для всех словоупотреблений, составляющих поисковый запрос. На основе ρ-последовательности ρ^F, являющейся результатом работы алгоритма формирования множества потенциально релевантных документов, вычисляются итоговые оценки соответствия найденных документов запросу пользователя и выполняется итоговое ранжирование результатов поиска.

Четвёртая глава посвящена реализации и экспериментальному исследованию метода оценки сходства текстов, структур данных и алгоритмов информационного поиска. Описана программная реализация метода оценки сходства текстов, алгоритмов и структур данных для решения поисково-аналитических задач. Предложена методика оценки разработанного метода оценки сходства текстов в приложении к задаче поиска информации. В основе методики – стандартные метрики оценки качества результатов информационного поиска[22]: 11-точечный график «полнота-точность»[23] (по методике TREC) и метрика nDCG[24]. В качестве тестовой коллекции для проведения экспериментов использована коллекция нормативных документов[25]

---

[22] Агеев М.С., Кураленок И.Б. Официальные метрики РОМИП. — В кн.: Труды третьего российского семинара РОМИП'2005 (Ярославль, 6 октября 2005 г.) - Санкт-Петербург: НИИ Химии СПбГУ, 2005. -224 с.

[23] The Twelfth Text Retrieval Conference (TREC 2003). Appendix 1. Common Evaluation Measures. / [Электронный ресурс] – URL: http://trec.nist.gov/pubs/trec12/appendices/measures.ps (дата обращения 14.01.2014).

[24] Kalervo Jarvelin, Jaana Kekalainen. Cumulated gain-based evaluation of IR techniques. / ACM Transactions on Information Systems №20(4), 2002. P.P.422–446

[25] Коллекция нормативных документов 2007. / 2007 // [Электронный ресурс] URL: http://romip.ru/ru/collections/legal07.html (дата обращения 24.02.2014)

(300 тыс. документов), предоставленная организаторами Российского семинара по оценке методов информационного поиска (РОМИП).

Было сформировано множество профилей ранжирования (наборы параметров метода многокритериальной оценки сходства), где каждый параметр изменялся по заданной дискретной решётке числовых значений таким образом, чтобы для профиля в целом сохранялось условие нормировки. Значения метрик рассчитаны на основе таблиц релевантности для 95 запросов, сформированных независимыми экспертами (в рамках семинара РОМИП). На рисунке 4 представлено сравнение метрик качества результатов информационного поиска разработанного алгоритма ранжирования ("Лучший профиль" и "Худший профиль") с результатами участников 2008 года: лучший результат и результат алгоритма Exactus-2008.

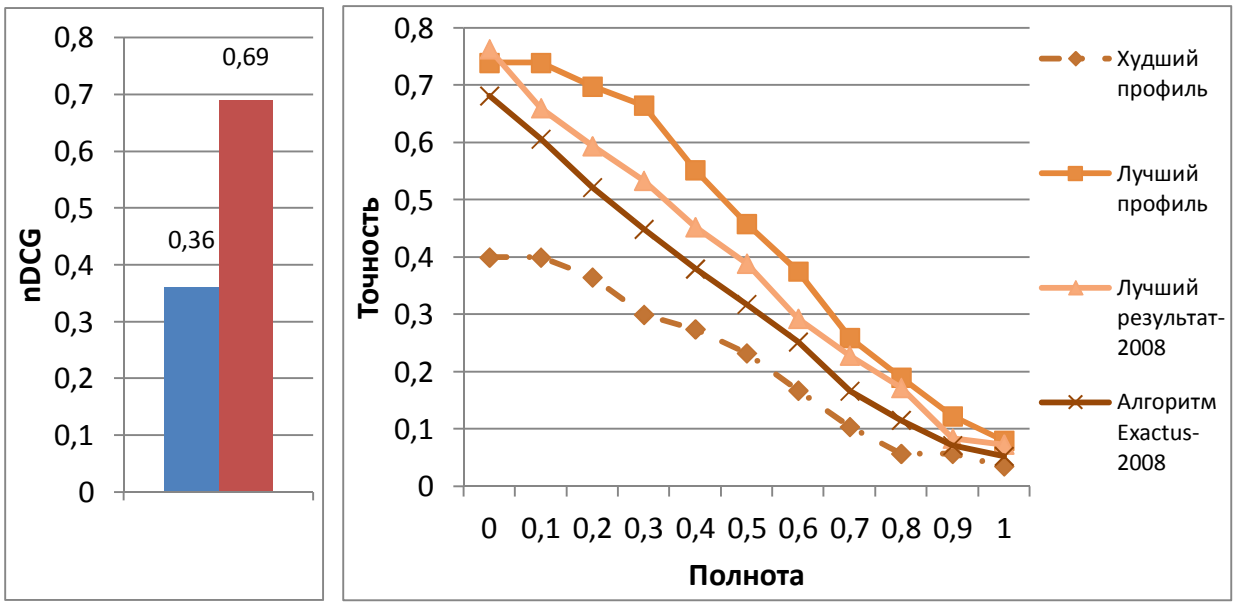

Рисунок 4 – Значения метрик качества результатов информационного поиска

Средние значения метрики nDCG (левая диаграмма) приведены для лучшего и худшего профиля ранжирования. Сравнение оценок качества ранжирования с результатами участников РОМИП 2008 по средней точности приведено на рисунке 5. На верхней диаграмме представлены абсолютные значения, а на нижней – изменение средней точности относительно лучшего результата 2008 г. Из представленных диаграмм видно, что относительное увеличение средней точности по сравнению с лучшим значением участников 2008 года составило 17%. В проведённых экспериментах доля найденных документов, для которых отсутствует оценка в таблице релевантности, по отношению к общему количеству найденных документов составляет 56%.

Вышесказанное означает, что полученные оценочные значения метрик качества информационного поиска для разработанного алгоритма ниже их реального значения, поскольку среди неоценённых документов (которые считались априорно нерелевантными) могут присутствовать релевантные.

Достигнутое повышение качества информационного поиска (на имеющихся таблицах релевантности) по сравнению с лучшими результатами других участников позволяет говорить о значимости результатов работы.

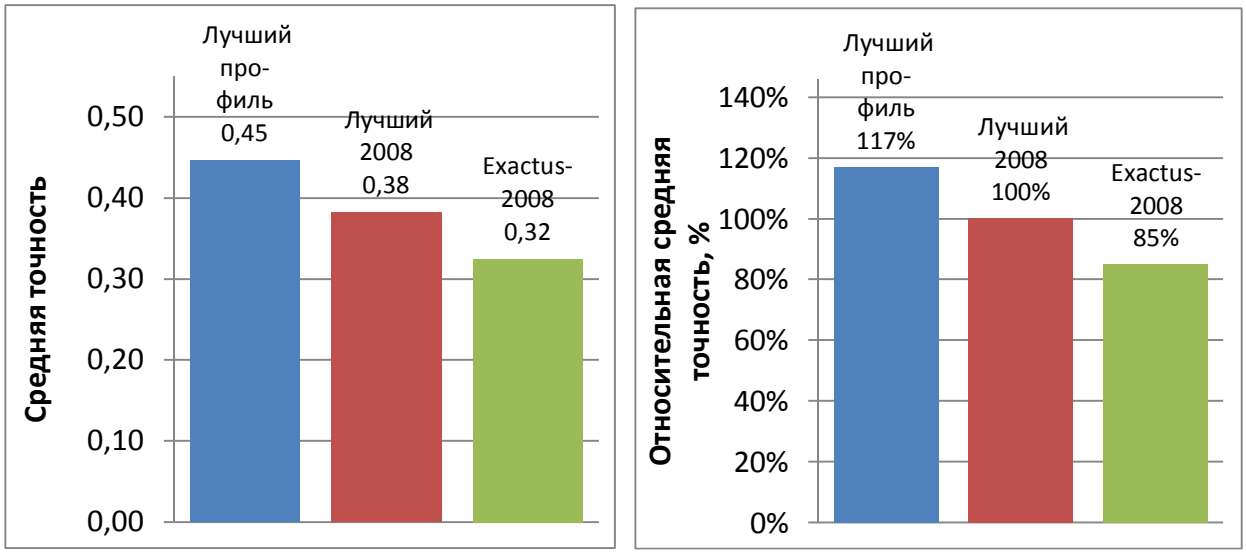

Рисунок 5 – Средняя точность результатов информационного поиска

На рисунке 6 приведены результаты оценки разработанных программных средств по объёму дисковой памяти, необходимой для хранения индексных структур, показавшие их эффективность в сравнении с известными структурами данных для решения задачи поиска информации.

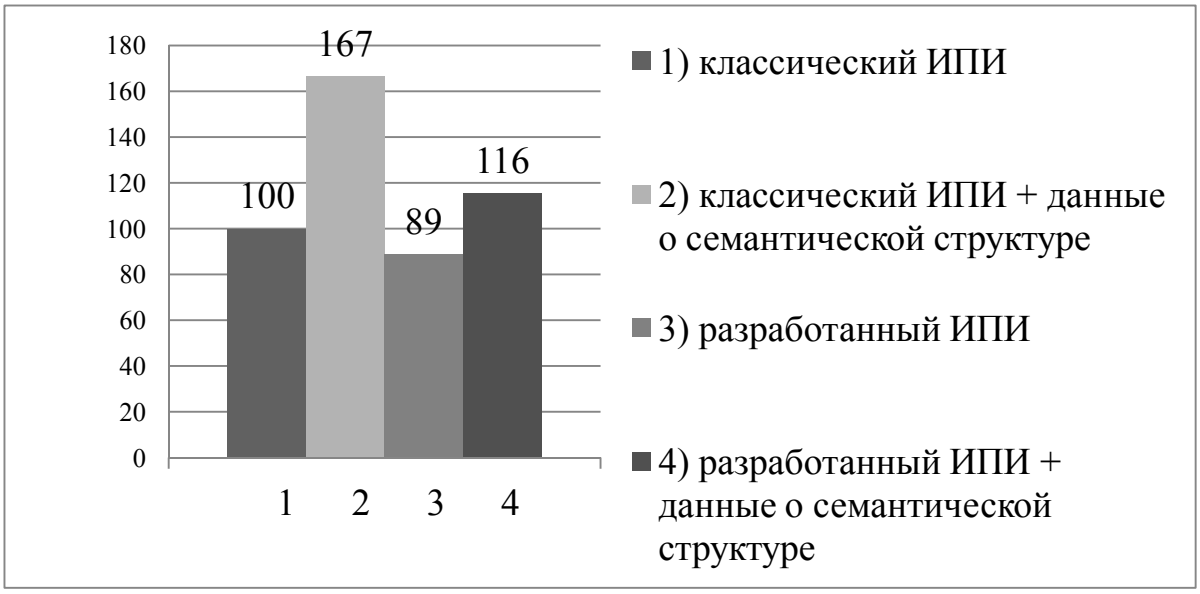

Рисунок 6 – Соотношение размеров инвертированных поисковых индексов при различных схемах представления информации

По сравнению с классической схемой представления информации в ИПИ (левый столбец) учёт семантических связей и значений увеличивает размер памяти, необходимой для хранения ИПИ, на 16% (в среднем). По сравнению с наиболее эффективной схемой организации ИПИ без учёта семантической информации (третий слева столбец) увеличение размера ИПИ составляет 30%. Использование классической схемы организации ИПИ с учётом семантической

информации приводит к увеличению затрат памяти на 67% (второй столбец слева). Преимущество разработанного ИПИ перед таким вариантом реализации ИПИ составляет 51% (в среднем).

Проведённые экспериментальные исследования разработанных алгоритмов и структур данных показали их практическую применимость. Полученные оценки качества информационного поиска свидетельствуют о том, что результаты поиска в значительной степени отвечают информационным потребностям экспертов. Характеристики программной реализации также соответствуют современным требованиям к скорости работы и объёмам занимаемой памяти информационно-поисковых и поисково-аналитических систем.

В **заключении** приведены основные результаты и выводы диссертационной работы, а также рассмотрены направления дальнейших исследований.

На защиту выносятся:

1. Новые синтаксические и семантические критерии оценки сходства текстов.
2. Новый метод многокритериальной оценки сходства текстов.
3. Новая модель данных, предназначенная для описания и исследования свойств структур данных и алгоритмов поисково-аналитической обработки текстовой информации.
4. Структуры данных предложенного в работе инвертированного поискового индекса, обеспечивающие эффективное хранение и обработку лексико-морфологической, синтаксической и семантической информации.
5. Алгоритмы поисково-аналитической обработки текстовой информации:
   – алгоритм построения инвертированного поискового индекса коллекций документов,
   – алгоритм поиска информации на основе метода многокритериальной оценки сходства текстов, реализующий поиск по ключевым словам, фразовый, семантический и вопросно-ответный поиск с учётом метаданных документов.
6. Программная реализация структур данных и алгоритмов, предназначенных для решения поисково-аналитических задач с помощью метода многокритериальной оценки сходства текстов.
7. Результаты экспериментального исследования разработанных структур данных и алгоритмов поисково-аналитической обработки текстовой информации, демонстрирующие, что эти алгоритмы обладают большей эффективностью и обеспечивают более качественное решение задач информационного поиска в ИАС, нежели известные методы.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ИССЛЕДОВАНИЯ

Основные положения диссертации отражены в следующих публикациях.

Публикации в журналах, входящих в перечень ВАК:

1. И.В. Соченков. Метод сравнения текстов для решения поисково-аналитических задач // Искусственный интеллект и принятие решений. М.: ИСА РАН, 2013, №2, с.95-106.
2. И. В. Соченков, Р. Е. Суворов. Сервисы полнотекстового поиска в информационно-аналитической системе (Часть 1) // Информационные технологии и вычислительные системы. М.: ИСА РАН, №2, 2013 , С. 69-78.
3. И. В. Соченков, Р. Е. Суворов. Сервисы полнотекстового поиска в информационно-аналитической системе (Часть 2) // Информационные технологии и вычислительные системы. М.: ИСА РАН, №3, 2013 , С. 71-87.
4. Р.Е. Суворов, И.В. Соченков. Определение связанности научно-технических документов на основе характеристики тематической значимости. // Искусственный интеллект и принятие решений. М.: ИСА РАН, №1, 2013. С.33-40.
5. Д.А. Девяткин, Р.Е. Суворов, И.В. Соченков. Метод тематической кластеризации масштабных коллекций научно-технических документов. // Информационные технологии и вычислительные системы. М.: ИСА РАН, №1, 2013 , С.33-42.
6. Э. Мбайкоджи, А.А. Драль, И.В. Соченков "Метод автоматической классификации коротких текстовых сообщений" // Информационные технологии и вычислительные системы. М.: ИСА РАН №3, 2012. С. 93-102.

Публикации в сборниках докладов российских и международных конференций:

7. Тихомиров И.А., Смирнов И.В., Соченков И.В., Девяткин Д.А., Шелманов А.О., Зубарев Д.В., Швец А.В., Лешкин А.В., Суворов Р.Е. Exactus Expert: Поисково-аналитическая система поддержки научно-технической деятельности / Труды тринадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2012. Б.: БГТУ, 2012. т. 4. - С. 100-108
8. O. Vybornova, I. Smirnov, I. Sochenkov, A. Kiselyov, I. Tikhomirov, N. Chudova, Y. Kuznetsova and G. Osipov Social Tension Detection and Intention Recognition Using Natural Language Semantic Analysis (on the material of Russian-speaking social networks and web forums) // In: Proceedings of the

European Intelligence and Security Informatics Conference (IEEE EISIC 2011), September 12-14; Athens, Greece, 2011

9. Соченков И. В., Мбайкоджи Э. Модель представления текста для решения задач машинного анализа естественно-языковой информации // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: Тезисы докладов Всероссийской конференции с международным участием. 18-22 апреля 2011г. – М.: Изд-во РУДН, 2011. Стр. 138–140.
10. Завьялова О.С., Киселёв А.А., Осипов Г.С., Смирнов И.В., Тихомиров И.А., Соченков И.В. Система интеллектуального поиска и анализа информации Exactus на РОМИП-2010 / Труды российского семинара по оценке методов информационного поиска РОМИП'2010. - Казань: Казан. ун-т, 2010. С49-69.
11. Назаренко Г.И., Плотникова В.А., Смирнов И.В., Соченков И.В., Тихомиров И.А. Программные средства создания и наполнения полнотекстовых электронных библиотек / «Электронные библиотеки: перспективные методы и технологии, электронные коллекции: XII Всероссийская научная конференция RCDL’ 2010, Казань, Россия - 2010. – С38-42.
12. Осипов Г.С., Смирнов И.В., Соченков И.В., Тихомиров И.А. Полнотекстовые электронные библиотеки с сервисами автоматического наполнения и высокоточного поиска / Семнадцатая международная Конференция "Крым 2010".
13. Смирнов И.В., Соченков И.В., Тихомиров И.А. Система интеллектуального поиска и анализа информации «Exactus» на РОМИП-2009 // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2009. Россия, Санкт-Петербург: НУ ЦСИ, 2009. Стр. 41-52.
14. Смирнов И.В., Муравьев В.В., Тихомиров И.А. Соченков И.В. Результаты и перспективы поискового алгоритма Exactus // Труды российского семинара по оценке методов информационного поиска РОМИП'2007-2008. Санкт-Петербург: НУ ЦСИ, 2008, с. 66-76

Прочие публикации:

15. Осипов Г.С., Тихомиров И.А., Соченков И.В., Смирнов И.В. СПОСОБ И СИСТЕМА СЕМАНТИЧЕСКОГО ПОИСКА ЭЛЕКТРОННЫХ ДОКУМЕНТОВ. / Патент РФ на изобретение №2473119, дата отсчета срока действия патента: 05.08.2011
16. Соченков И.В. «Программа определения степени релевантности запросу пользователя». / Свидетельство о государственной регистрации программ для ЭВМ. № 2013613284, 2013 г.

17. Девяткин Д.А., Соченков И.В., Суворов Р.Е. «Программа объективной вычислительной оценки соответствия научно-технического документа заданному набору тематик». / Свидетельство о государственной регистрации программы для ЭВМ. №2013613412, 2013 г.
18. Зубарев Д.В., Соченков И.В. «Программа хранения научно-технических документов на естественном языке». / Свидетельство о государственной регистрации программы для ЭВМ. №2013613410, 2013 г.

Личный вклад соискателя: в работах 1–18 автору принадлежат метод оценки сходства текстов, представление текстовой информации в информационно-аналитических системах, структуры данных и алгоритмы информационного поиска.

Соченков Илья Владимирович (Россия)

РЕЛЯЦИОННО-СИТУАЦИОННЫЕ СТРУКТУРЫ ДАННЫХ, МЕТОДЫ И АЛГОРИТМЫ РЕШЕНИЯ ПОИСКОВО-АНАЛИТИЧЕСКИХ ЗАДАЧ

Разработаны лексико-морфологические, синтаксические и семантические критерии оценки сходства текстов, а также метод многокритериальной оценки сходства текстов.

Предложена и исследована модель данных, предназначенная для формализации и анализа свойств структур данных и алгоритмов поисково-аналитической обработки текстовой информации.

Разработаны структуры данных инвертированного поискового индекса, предназначенные для хранения и обработки лексико-морфологической, синтаксической и семантической информации, являющейся результатом компьютерного лингвистического анализа текстов документов.

Разработанные алгоритмы и структуры данных информационного поиска программно реализованы и экспериментально проверены. Получены оценки параметров качества информационного поиска. Разработанные структуры данных и алгоритмы поисково-аналитической обработки текстовой информации обеспечивают более эффективное и качественное решение задач информационного поиска, нежели известные методы, за счёт применения многокритериальной оценки сходства текстов на основе лингвистической информации.

Sochenkov Ilya (Russia)

RELATIONAL-SITUATIONAL DATA STRUCTURES, ALGORITHMS AND METHODS FOR SEARCH AND ANALYTICAL TASKS SOLVING

Lexico-morphological, syntactic and semantic criteria for text similarity measurement are developed. Multicriteria text similarity measure is presented.

New data model for formalization and analysis of data structures and algorithms of search and analytical text processing is proposed.

Inverted search index for storing linguistic information of text is developed. These data structures and algorithms are meant for storing and processing of lexico-morphological, syntactic and semantic information as results of computational linguistic analysis of text.

Implemented algorithms and data structures for informational search were experimentally-verified. Information retrieval evaluation showed that proposed search method on the base of linguistic information comparison performs better for information search tasks than known methods.