

«УТВЕРЖДАЮ»

Проректор по научной работе
НИУ «МЭИ»

 Драгунов В.К.
« 1 » 10 2014 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

Федерального государственного бюджетного образовательного учреждения
высшего профессионального образования
«Национальный исследовательский университет "МЭИ"»
на диссертацию Соченкова Ильи Владимировича
«РЕЛЯЦИОННО-СИТУАЦИОННЫЕ СТРУКТУРЫ ДАННЫХ, МЕТОДЫ И
АЛГОРИТМЫ РЕШЕНИЯ ПОИСКОВО-АНАЛИТИЧЕСКИХ ЗАДАЧ»,
представленную на соискание ученой степени кандидата физико-
математических наук по специальности 05.13.17 — «Теоретические основы
информатики»

В большинстве поисковых и аналитических информационных систем для индексирования текстовой информации применяется векторная модель, не учитывающая лингвистическую природу текста и, по этой причине, не позволяющая решать многие задачи информационного поиска с достаточной полнотой и точностью. Речь идет, в частности, о вопросно-ответном поиске, определении наиболее значимых фрагментов текстов, установлении тематического и семантического сходства. В работах российских и зарубежных исследователей показано, что применение для решения этих задач лингвистических методов анализа текстовой информации способствует повышению полноты и точности информационного поиска.

В современных информационно-аналитических системах при решении задач поиска информации применяются инвертированные индексы, структура которых обеспечивает эффективное хранение и выборку информации, описывающей текстовые документы в векторной модели. Реляционные и графовые базы данных не обеспечивают эффективного хранения и быстрых операций выборки указанных данных, в первую очередь, из-за количества хранимых словоупотреблений и связей между ними, исчисляемых десятками миллиардов на миллион документов. Реализация информационно-аналитических функций на подобных базах данных требует многократного увеличения вычислительных мощностей. В то же время в случае классической векторной модели инвертированные индексы успешно решают стоящие перед ними задачи для миллиона документов на одном

вычислительном сервере средней конфигурации. В то же время применение для индексирования тестовой информации результатов синтаксического и семантического анализа требует создания и исследования таких способов представления информации, структура которых отличается от векторов числовых признаков. Остаётся также открытым вопрос об организации информационных хранилищ для компактного хранения и быстрой выборки данных, представляющих собой семантические сети и синтаксические деревья предложений текста.

Работа И.В. Соченкова посвящена решению именно этих задач, что свидетельствует о её актуальности.

Целью диссертационного исследования является повышение качества (полноты и точности) и вычислительной эффективности решения поисково-аналитических задач в информационно-аналитических системах.

Диссертация И.В. Соченкова 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.

Во введении обоснована актуальность темы, определен предмет исследования, сформулированы цель и задачи исследования, научная новизна, практическая значимость полученных результатов.

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 информации в системах информационного поиска. Далее рассматриваются структуры данных (инвертированные поисковые индексы), применяемые для представления текстовой информации в задачах информационного поиска, и основные методы ранжирования результатов информационного поиска.

Вторая глава посвящена методу оценки структурного сходства текстов на основе сопоставления результатов лексико-морфологического, синтаксического и семантического анализа (Semantic Role Labeling). Приведено формальное описание неоднородной семантической сети, представляющей значения лексико-морфологических признаков словоупотреблений в текстах, а также синтаксические и семантические связи. На её основе определены критерии сравнения текстов по предложениям и доказаны утверждения о свойствах введённой оценки сходства текстов.

Третья глава посвящена разработке и исследованию модели данных, структурам данных и алгоритмам для решения поисково-аналитических

задач. Показано, как результаты лингвистического анализа в форме неоднородной семантической сети, описанной во второй главе, представляются в виде предложенного И.В. Соченковым реляционно-ситуационного инвертированного поискового индекса. Описаны разработанные автором алгоритмы индексации текстовой информации, а также алгоритмы, применяемые в ходе оценки сходства текстов с помощью предложенного поискового индекса. Доказаны утверждения о свойствах операций над данными, а также утверждения, содержащие оценки вычислительной сложности алгоритмов, предложенных автором.

Четвёртая глава посвящена реализации и экспериментальному исследованию метода оценки сходства текстов, структур данных и алгоритмов информационного поиска.

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

В качестве сильных сторон работы отметим, что теоретические положения работы хорошо отражают практический аспект применения результатов лингвистического анализа в приложениях к реальным задачам. В частности, в предложенном автором представлении текстовой информации учитывается потенциально не снятая многозначность (омонимия), математически описан и алгоритмически реализован фразовый и вопросно-ответный поиск. Кроме того, определена операция пополнения неоднородной семантической сети для оценки сходства текстов с учётом разрешения анафоры, а также тезаурусных отношений (например, синонимии) для реализации семантического поиска. О достижении целей работы свидетельствуют данные экспериментального исследования, выполненного диссертантом.

Результаты исследований по теме диссертационной работы использованы при выполнении ряда научно-исследовательских работ по проектам Минобрнауки РФ, программам ОНИТ РАН и грантам РФФИ. Таким образом, полученные в работе результаты обладают не только теоретической значимостью, но и практической ценностью, и могут быть использованы в ИСА РАН, ВЦ РАН, ИСП РАН, ИППИ РАН, ВМиК МГУ и в других ведущих научных учреждениях Российской Федерации и коллективах, где проводятся исследования в области информационного поиска.

По диссертации И.В. Соченкова имеются следующие замечания.

1. В обзоре литературы не упоминаются базы данных на основе графовых моделей в качестве одного из возможных способов хранения и обработки структурированной информации в виде сетей и деревьев.
2. В диссертации не обоснован выбор именно линейной комбинации значений критериев в качестве результирующей оценки сходства текстов.
3. В сериях экспериментов при оценке метрик качества результатов поиска использованы только русскоязычные коллекции текстов. Поэтому вопрос повышения полноты и точности поиска в случае коллекций текстов на иностранном языке остаётся открытым.
4. При записи ряда формул, в частности (2.6), (2.8) и др., в качестве индексов операции суммирования используются громоздкие и трудноинтерпретируемые конструкции, содержащие кортежи, кванторы, логические операции.
5. Замечание по автореферату. Отмечено, что разработаны и исследованы алгоритмы поисково-аналитической обработки текстовой информации и получены оценки их вычислительной сложности, однако сами оценки, подтверждающие практическую применимость этих алгоритмов для обработки больших объемов текстовой информации, в автореферате отсутствуют (в диссертации эти оценки приведены).

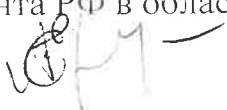
Приведенные замечания не снижают общей положительной оценки диссертационной работы.

Научные результаты диссертации, выносимые на защиту, являются новыми и получены диссертантом лично. Диссертация И.В. Соченкова является законченной научно-исследовательской работой, выполнена на высоком научном уровне. Сформулированные утверждения подкрепляются строгими математическими доказательствами. Автореферат достаточно полно отражает содержание диссертации.

Результаты диссертационной работы обоснованы и докладывались на 6 российских и международных конференциях. По теме исследования И.В. Соченковым опубликовано 14 работ: 6 из них в рецензируемых журналах из списка ВАК, 8 в материалах российских и международных конференций. Получен патент Российской Федерации на изобретение и 3 свидетельства о государственной регистрации программ для ЭВМ.

Диссертация соответствует критериям, установленным Положением о присуждении ученых степеней: в ней содержится решение задачи, имеющей существенное значение для теоретических основ информатики, она написана автором самостоятельно, обладает внутренним единством и содержит новые научные результаты, полученные автором лично. В диссертации приведены рекомендации по использованию полученных результатов. Работа полностью соответствует требованиям ВАК, предъявляемым к диссертациям на соискание ученой степени кандидата физико-математических наук, а ее автор И.В. Соченков заслуживает присуждения ему ученой степени кандидата физико-математических наук по специальности 05.13.17 - «Теоретические основы информатики».

Отзыв обсужден и одобрен на заседании кафедры прикладной математики ФГБОУ ВПО «Национальный исследовательский университет "МЭИ"» 1 октября 2014 г., протокол № 2.

Зав. кафедрой прикладной математики НИУ «МЭИ»
лауреат Премии Президента РФ в области образования,
д.т.н., проф.  Еремеев А.П.