

Федеральный исследовательский центр «Информатика и управление» Российской
Академии Наук (ФИЦ ИУ РАН)

Россия, 119333, г. Москва, ул. Вавилова, д.44, корп. 2
Тел.8 (499) 135-62-60, факс 8 (495) 930-45-05
E-mail: ipiran@ipiran.ru <http://www.frccsc.ru/>

ОТЗЫВ

на автореферат диссертации Черняк Е.Л. «Разработка вычислительных методов анализа неструктурированных текстов с использованием аннотированных суффиксных деревьев» на соискание ученой степени кандидата технических наук по специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ»

Диссертация Черняк Екатерины Леонидовны посвящена актуальной проблеме создания эффективных методов, алгоритмов и программных комплексов, предназначенных для решения ряда важных задач анализа естественно-языковых текстовых документов, включая задачу информационного поиска, на уровне последовательностей символов. Автор диссертационного исследования предлагает и обосновывает теоретико-множественную модель совокупности «строка – текст», при этом в качестве адекватной модели данных для вычисления параметров нечеткой оценки релевантности предлагается аннотированное суффиксное дерево - структура данных, которая позволяет вычислять все частоты всех подстрок. Для реализации предложенного подхода автор диссертации разработал новый лингвистический ресурс, алгоритмы, программы.

Выполненная диссертационная работа обладает существенной новизной, которая заключается в создании нового метода вычисления оценок релевантности строки тексту; разработке метода использования справочных материалов интернета с учетом наличия в них шумовой компоненты для пополнения предметных таксономий, апробированного в задачах пополнения таксономий чистой и прикладной математики с использованием русскоязычной Википедии; применении критерия релевантности средней условной вероятности символа в совпадении в классе задач поиска по однословному ключу, в котором полнота важнее, чем точность.

В диссертации Черняк Е.Л. разработан понятийный аппарат для решения поставленных задач и проведено систематическое изложение основных методов и путей их решения. Автор диссертации вводит понятие меры средней условной вероятности символа в совпадении (СУВСС) и проводит эксперимент, в котором рассматривает наиболее популярные в международной литературе меры релевантности, производит сравнение результатов их использования с результатами использования СУВСС и наглядно демонстрирует достоинства предложенного подхода. Эксперименты в диссертационном исследовании выполнены на материале представительной выборки текстов: автором были использованы 5079 аннотаций научных статей из журналов ACM, представленных в свободном доступе в электронной библиотеке ACM Digital Library; таксономия ACM CCS 2012, состоящая из 2074 тем 6 уровней; авторские темы, приписанные аннотациям научных статей их авторами – 2-3 темы низших уровней и все темы, лежащие на пути от корня до них в дереве таксономии ACM CCS.

Предлагаемый автором метод имеет большое практическое значение: он позволяет создавать системы, ориентированные на работу с неструктурированными текстами. Результаты диссертационного исследования могут существенно усилить такие практические инструменты как системы извлечения знаний из текстов, аналитического мониторинга заданной предметной области, а также могут эффективно применяться в учебных и исследовательских целях в области информационного поиска, компьютерной лингвистики, информатики.

Автореферат диссертации состоит из следующих разделов: общая характеристика диссертации, основное содержание диссертации, основные результаты работы, список публикаций автора.

В разделе "Общая характеристика диссертации" обосновывается актуальность темы диссертации, определяется цель работы, описываются методы исследования, формулируются основные результаты, обсуждаются теоретическая и практическая значимость работы.

В разделе "Структура диссертации" приводится краткое изложение введения, шести глав и заключения работы. В первой главе приводится обзор основных видов формализации коллекций текстовых документов: векторная модель, языковая модель, модель скрытых тем, модель суффиксного дерева. Вводится теоретико-множественная модель представления коллекции текстовых документов, вводится понятие нормированного аннотированного суффиксного дерева. Вторая глава посвящена задаче вычисления релевантности «строка – текст», которая, как отмечает автор диссертации, является базовым этапом любой практической задачи обработки и анализа коллекций текстовых документов. В третьей, четвертой и пятой главах рассматриваются примеры использования теоретико-множественной модели коллекции текстовых методов и ассоциированного метода вычисления релевантности с использованием нормированного аннотированного суффиксного дерева для решения следующих задач: рубрикации аннотаций научных публикаций темами таксономии, пополнения научной таксономии, фильтрации обценной лексики. Для решения задачи фильтрации обценной лексики автор диссертации создал оригинальный лингвистический ресурс - аннотированную коллекцию текстов, общий размер которой составляет 294916 словоупотреблений и 60868 словоформ. В шестой главе приводится описание программных комплексов, в которых реализованы разработанные в диссертационном исследовании модели и методы, а также представлены решения целого ряда вспомогательных задач сбора и обработки данных. В заключении диссертации приведены основные результаты, выводы и практические итоги работы.

Автореферат диссертации написан грамотным научным языком по установленной форме при соблюдении необходимой меры подробности в представлении содержания работы.

В автореферате отчетливо обоснованы актуальность темы диссертации, научная новизна, практическая и теоретическая значимость работы, сформулированы основные результаты. Автореферат дает хорошее представление о диссертационном исследовании и позволяет сделать вывод о том, что работа соответствует требованиям ВАК к диссертациям на соискание степени кандидата наук, а ее автор заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ».

Зав. Лабораторией Компьютерной лингвистики
и когнитивных технологий обработки текстов ИПИРАН
кандидат филологических наук



Е.Б. Козеренко

Подпись Е.Б. Козеренко заверяю
Ученый секретарь ИПИРАН
доктор технических наук



В.Н. Захаров

14.10.2016

Козеренко Елена Борисовна
E-mail: kozerenko@mail.ru

Тел. +7 (499) 135-62-60 (ОДО), +7 916 631 58 83 (моб.)

Адрес: 119333, г. Москва, ул. Вавилова, 44, корп. 2, ФИЦ ИУ РАН