

Утверждаю
ВРИО директора федерального
государственного учреждения
«Федеральный исследовательский центр
Институт прикладной математики им.
М. В. Келдыша РАН, д.ф.-м.н., проф.



А. И. Аптекарев

3 октября 2016

Отзыв ведущей организации на диссертацию
Черняк Екатерины Леонидовны

«Разработка вычислительных методов анализа неструктурированных текстов с использованием аннотированных суффиксных деревьев», представленную на соискание ученой степени кандидата технических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ

Актуальность темы диссертационной работы.

Актуальность темы диссертации Е. Л. Черняк определяется важностью задачи компьютерного (автоматического) анализа неструктурированных текстов; эта задача возникает в различных прикладных областях, в частности, при анализе Интернет-данных. В настоящее время известен ряд подходов к представлению текстов, которые используются для анализа текстов различной природы. Исследование целесообразности применения тех или иных подходов в различных областях и выделение подходов, адекватных конкретной прикладной области (в случае работы Е.Л.Черняк – это прежде всего тексты Википедии), безусловно, является актуальной задачей.

Новизна и значимость основных научных результатов.

Работа носит практический характер, проведена на основе анализа результатов экспериментов по обработке текстов. Основные научные результаты, полученные автором, состоят в следующем.

1. Предложена мера релевантности «строка – текст», для вычисления которой используется представления текста в виде набора строк, организуемых в аннотированное суффиксное дерево. Предложенная мера применима для анализа текстов различных длин и позволяет учитывать нечеткие совпадения строки и текста.
2. Предложенная мера релевантности использована
 - a. В задаче рубрикации текстов. Показано, что при ее использовании, точность рубрикации возрастает в полтора–два раза по сравнению с мировыми аналогами;
 - b. В задаче пополнения таксономии для сравнения тем заданной основы таксономии с категориями Википедии. На выходе метода пополнения таксономии получается таксономия высокого качества (качество оценено экспертами предметной области);
 - c. В задаче фильтрации обценной лексики. Показано, что при ее использовании, полнота и скорость фильтрации превосходит мировые аналоги.

3. Разработаны два комплекса программ, используемые для сбора и последующей обработки текстов. Заметим, что комплекс программ, посвященный непосредственно обработке текстов может быть с легкостью использован и для обработки любых символьных последовательностей.

Содержание работы.

Диссертационная работа состоит из введения, 6 глав, заключения и списка литературы, насчитывающего 105 наименований.

Введение посвящено, прежде всего, неформальной постановке задачи, обоснованию ее актуальности. В нем также сформулированы цели исследования, описаны основные шаги на пути к достижению этих целей, а также введено понятие «модели представления текстов».

Первая глава носит обзорный характер: в ней представлены векторная и вероятностная модели представления текста, языковая модель, модели представления текста, использующие суффиксные деревья.

Понятие релевантности строки тексту вводится во второй главе. В этой же главе перечислены меры релевантности, связанные с моделями представления текста, описанными в предыдущей главе и, в частности, мера релевантности в модели аннотированного суффиксного дерева. В последующих трех главах рассмотрены различные примеры использования этой меры в практических задачах.

Третья глава посвящена задаче рубрикации текстов. Рассматривается простейший подход к решению этой задачи без использования методов машинного обучения: выбираются те рубрики, которые получают максимальные оценки по той или иной мере релевантности. В качестве тестовой коллекции используется стандартный для этой задачи набор публикаций, опубликованных в электронной библиотеке Ассоциации Вычислительной Техники (DL ACM). Показано, что лучшие результаты достигаются при использовании меры релевантности, использующей аннотированные суффиксные деревья.

В четвертой главе рассмотрена задача пополнения таксономии. Предполагается, что из внешнего источника заимствуется основа таксономии, т.е. два - три уровня, задающих достаточно общие понятия. Эту основу таксономии нужно автоматических пополнить темами менее общими. В качестве источника таких тем автор выбирает Википедию, а именно, дерево категорий и названия статей из предметной области, соответствующей основе таксономии. Рассматриваются две предметные области: а) теория вероятностей и математическая статистика, б) численные методы. Оценки одной из пополненных, полученные при помощи экспертного оценивания, таксономий демонстрируют высокое качество построенной таксономии.

В пятой главе проводится сравнение различных строковых мер сходства для поиска слов по заданному списку, в данном случае, obscene слов по заданному стоп-списку. Утверждается, в задаче фильтрации obscene лексики полнота результатов важнее точности, поэтому речь идет именно о вычислении оценок сходства, а не релевантности, поскольку необходимо найти все вхождения стоп-слов и их производных. Показано, что при использовании меры сходства, основанной на аннотированных суффиксных деревьях, достигаются лучшие показатели не только полноты, но и скорости работы метода.

В заключении подводятся основные итоги работы и еще раз подчеркиваются все обсуждаемые выше достоинства меры, основанной на аннотированных суффиксных деревьях, ее ограничения и возможные направления работы.

Рекомендации по использованию результатов и выводов диссертации.

1. Необходимо продолжить работу по развитию метода пополнения таксономии: добавить поддержку синонимов и протестировать метод на других языках и предметных областях.
2. Необходимо протестировать метод рубрикации на других тестовых коллекциях, возможно, на других языках.
3. Необходимо опубликовать в свободном доступе размеченный корпус обценной лексики и расширить метод на случай многословных обценных выражений.

Общие замечания.

Текст диссертации не свободен от недостатков. Литературный обзор, в целом очень подробный, в некоторых местах написан тяжело для восприятия. Во второй главе, посвященной различным мерам релевантности, не хватает иллюстративных примеров для вычисления косинусной меры релевантности, меры релевантности BM25 и др.

Заключение.

Диссертация представляет собой завершенную научно-исследовательскую работу на актуальную тему. Новые научные результаты, полученные диссертантом, имеют существенное значение для области автоматической обработки текстов и могут быть использованы специалистами по компьютерной лингвистике и информационному поиску, а также в прикладных задачах тематической классификации текстов или классификации текстов по тональности. Выводы и рекомендации достаточно обоснованы. Работа отвечает требованиям п. 9 «Положения о порядке присуждения ученых степеней», предъявляемым к кандидатским диссертациям, а её автор заслуживает присуждения ему ученой степени кандидата биологических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ.

Отзыв на диссертацию и автореферат обсужден на заседании семинара лаборатории прикладной математики филиала ФГУ ФИЦ ИПМ им. М.В.Келдыша РАН 16 сентября 2016 г.

Адрес: 125047, Москва, Миусская пл.,
д.4, ИПМ им. М.В.Келдыша РАН, тел.: +7 499 978-13-14, e-mail:
office@keldysh.ru

Отзыв подготовил зав. лабораторией прикладной математики филиала ФГУ ФИЦ
ИПМ им. М.В.Келдыша РАН, д.ф.-м.н.

М.А.Ройтберг

М.А.Ройтберг
30.09.2016

Подпись М.А. Ройтберга
Начальник Отдела кадров
ИПМ им. М.В. Келдыша РАН



А.В. Федосин