ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу 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

Оценка актуальности темы диссертационной работы

В современном мире одной из важнейших форм собираемой, хранимой и обрабатываемой информации становится форма неструктурированных текстовых данных, зачастую содержащих тексты на разных языках, сленг или специальную лексику, а также просто тексты низкого качества с точки зрения грамматики. Такие тексты необходимо обрабатывать, выявлять знания, осуществлять информационный поиск, группировать, рубрицировать и индексировать. Решение подобных задач невозможно без использования математических моделей представления неструктурированных текстовых данных, а также без определения мер сходства текстовых документов и мер релевантности, позволяющих оценить уровень соответствия текста поисковому запросу или набору ключевых слов. Существует много подходов к моделированию представления текстовых данных и определению мер сходства и релевантности. Наиболее часто используется векторная модель коллекции текстов, чуть менее популярна вероятностная языковая модель, использующая, например, аппарат цепей Маркова. В последнее время все большее признание получают модели, полученные с помощью методов на основе латентно-семантического анализа. Разумеется, любая из используемых моделей представления текстов обладает определенными ограничениями. Например, в векторной модели не учитывается порядок слов, многие вероятностные модели делают не всегда обоснованные предположения о форме или параметрах используемых распределений. Модели на основе латентно-семантических методов могут терять важную информацию при преобразовании текстов в тематическое пространство признаков.

Диссертационная работа Е.Л. Черняк посвящена разработке новой модели представления текста и соответствующего набора алгоритмов для ее построения и применения, а также программной реализации предложенных алгоритмов. В отличие от упомянутых выше подходов, в данной работе минимальной единицей считается так называемая строка, состоящая из нескольких последовательных слов. Точное количество слов выбирается эмпирически, в зависимости от постановки конкретной практической задачи.

Такое представление текста позволяет разрешить до некоторой степени проблему потери порядка слов и ввести меру сходства между строкой из нескольких слов и текстом, названную в работе СУВСС (средняя условная вероятность символа в совпадении) и имеющую смысл доли общих подстрок данной строки и текста. В работе исследуется возможность использования для агрегированного представления текста специальной структуры – аннотированного суффиксного дерева, благодаря которому СУВСС вычисляется за линейное время. Разумеется, спектр задач, в которых такая модель может быть использована, ограничен теми задачами, в которых результат не имеет нового текстового преставления: тематической классификацией, рубрикацией или кластеризацией текстов, что ни в коей мере не умаляет актуальность исследуемой задачи.

Содержание работы

Текст диссертации, отражающий результаты исследовательской работы, состоит из введения, шести глав, выводов, заключения и списка литературы, включающего 105 наименований.

Введение посвящено постановке задачи и обоснованию ее актуальности. Выдвигаются теоретические требования к разрабатываемой модели представления текста и формулируются основные ее принципы.

Первая глава является обзорной: перечисляются наиболее популярные модели представления текста и их практическое назначение.

Во второй главе определяется понятие релевантности и устанавливается связь между различными моделями представления текстов и мерами релевантности. Вводится мера релевантности СУВСС. Определяется структура аннотированного суффиксного дерева и способ его использования для вычисления СУВСС.

Третья, четвертая и пятая главы посвящены практическим задачам. В третьей главе СУВСС используется для рубрикации научных публикаций без учителя. В четвертой главе предложен метод пополнения таксономии за счет дерева категорий Википедии. В пятой главе представлен фильтр обсценной лексики, использующий СУВСС в качестве критерия.

В шестой главе описаны разработанные комплексы программ, используемые для сбора данных, предварительной подготовки текстов, построения суффиксных деревьев, вычисления СУВСС.

Текст диссертации написан грамотным научным языком, логично выстроен. Автореферат диссертации в полной мере отражает ее основное содержание, структуру, положения и выводы.

Научная новизна

В диссертационной работе получен ряд результатов, содержащих бесспорную научную новизну и обладающих практической значимостью. В совокупности полученных автором результатов необходимо выделить наиболее весомые:

1. Предложена новая модель представления текстовых данных на основе аннотированных суффиксных деревьев, генерируемая по совокупности коротких последовательностей слов исходных текстовых документов, полученных с использованием скользящего окна.
2. Разработана новая параметризуемая шкалирующими функциями мера сходства строки и текста, позволяющая одновременно получать более универсальные с практической точки зрения оценки релевантности и избавляющая от необходимости глубокого морфологического и синтаксического анализа;
3. Применение предложенных новых алгоритмических решений и подходов успешно апробировано на трех важных прикладных задачах. В частности, предложен метод рубрикации научных статей без учителя, предполагающий вычисление оценок сходства рубрик и текстов. Показано преимущество СУВСС по сравнению с другими мерами сходства по точности ранжирования результатов. Предложен метод пополнения таксономии данными из Википедии. Эффективность метода продемонстрирована экспертным оцениванием одной из построенных таксономий по теории вероятностей и математической статистике. Предложен метод фильтрации обсценной лексики, предполагающий вычисление оценок стоп-слов и текстов. Показано преимущество СУВСС по сравнению с другими мерами сходства по полноте результатов.
4. Разработаны два комплекса программ, реализующих все вышеперечисленные методы, программный код которых находится в свободном доступе, что позволяет убедиться в его функциональности и работоспособности.

Достоверность и практическая ценность полученных научных результатов

Все основные положения, выносимые на защиту, являются обоснованными. Это подтверждается подробными сравнениями предложенного подхода и получаемых решений с известными. Многочисленные примеры и проведенные численные эксперименты

подтверждают эффективность авторского подхода и убедительно иллюстрируют основные результаты и выводы диссертационной работы.

Практическая ценность подтверждается результатами решения прикладных задач и их успешным применением в реальных проектах компании «ФОРС-Центр разработки» и «ЕС-Лизинг».

Замечания по диссертации

1. В диссертационной работе приведен обзор популярных моделей представления текстовых данных, а также подходов и методов к определению меры сходства между текстами и к оценке релевантности строки тексту. В то же время в данном обзоре не отражены такие два важных современных направления, как использование модели представления типа word2vec, а также подходов к определению меры сходства или релевантности текстовых данных на основе строковых ядерных функций (string kernels). Последнее особенно актуально, так как существуют методы построения ядерных функций для строк символов, позволяющие рассчитывать степень сходства на основе совпадения всех подстрок, что идейно близко к подходу, предложенному автором.
2. В работе не до конца освещены необходимость и полезность использования шкалирующих функций при задании меры релевантности, не обсуждается, как выбор функции шкалирования может повлиять на результат, и никак не ограничивается класс этих функций.
3. При представлении «строка-текст», предложенном автором, набор строк по сути генерируется скользящим окном из нескольких слов. Упоминается, что размер окна и, судя по всему, степень перекрытия должны подбираться эмпирически. При этом вопрос выбора этих параметров в экспериментах практически не обсуждается, хотя, скорее всего, влияние этих параметров существенно.
4. В работе упоминается что «латентно-семантический анализ не является интерактивным методом, то есть, при появлении нового текста в коллекции необходимо заново формировать матрицу терм – текст и заново вычислять сингулярные матрицы и матрицу сингулярных значений». Это верно только отчасти. Некоторые методы LSA, такие как методы на основе неотрицательной матричной факторизации действительно не позволяют представить новый документ в пространстве признаков на основе тематик коллекции без пересчета самой тематической модели. Но большинство методов LSA, в том числе на основе главных компонент или латентного размещения Дирихле, позволяют представить новый

документ в пространстве тематических признаков коллекции без пересчета самой тематической модели.

5. Также есть замечания по структуре работы. Главы с описанием практических задач не достаточно сбалансированы по уровню детализации, а также невелики по объему. Не приведена аргументация, почему именно эти прикладные задачи были выбраны в качестве демонстрационных. Заключительная глава с описанием программной реализации очень короткая и собственно детали реализации, которые позволили бы оценить уровень трудозатрат автора, практически не освещены, а выводы по этой главе отсутствуют. Имело бы смысл объединить главы с третьей по пятую в одну общую главу с обоснованием выбора прикладных задач, а описание программной реализации вынести в приложение, добавив больше деталей и выводы.

6. В работе содержатся опечатки, не препятствующие общему пониманию работы, но иногда вводящие в заблуждение, например:

- На стр. 6 полная релевантность строки тексту определяется как сумма вероятностей максимальных совпадений, а на стр. 41 тот же термин определяется как усредненная сумма.
- На стр. 27 в тексте для описания рисунка 1.1. используются обозначения вершин, не совпадающие с обозначениями на самом рисунке (не u и v, а u и w).
- На стр. 43 упоминается три типа шкалирующих функций, в том время как реально в работе используется семь разных шкалирующих функций.

7. Также присутствует ряд технических опечаток. Например: на стр. 16 – тире в начале строчки «за отсеивание слов»; стр. 21 «скрытое переменной, тема состоит набора»; стр. 32 «Предварительна подготовка»; стр. 39 – «текстовый текст»; стр. 45 – отсутствует точка в последней строчке и другие.

Высказанные замечания не снижают общей положительной оценки диссертационного исследования, а создают поле для научной дискуссии, неизбежной при обсуждении результатов.

Заключительная оценка

Диссертационная работа Екатерины Леонидовны Черняк «Разработка вычислительных методов анализа неструктурированных текстов с использованием аннотированных суффиксных деревьев» является законченной научно-исследовательской квалификационной работой. Полученные результаты диссертационной работы достоверны, практические выводы обоснованы. Основные результаты диссертации с достаточной полнотой изложены в научных

публикациях автора, среди которых 3 статьи в журналах, рекомендованных ВАК при Минобрнауки России.

По своему содержанию, актуальности и научной новизне, объему проведенного исследования, теоретической и практической ценности полученных результатов диссертация Черняк Екатерины Леонидовны соответствует паспорту специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ», а ее автор заслуживает присуждения ученой степени кандидата технических наук.

Доцент кафедры Автоматизации систем вычислительных комплексов
факультета Вычислительной математики и кибернетики
Московского государственного университета имени М. В. Ломоносова,
кандидат физико-математических наук

Петровский Михаил Игоревич

Почтовый адрес: 119991 ГСП-1 Москва, Ленинские горы,
МГУ имени М.В. Ломоносова, 2-й учебный корпус, факультет ВМК
Тел. +7 495 9391789, e-mail: michael@cs.msu.su