

ОТЗЫВ

официального оппонента о диссертационной работе Черняк Екатерины Леонидовны «Разработка вычислительных методов анализа неструктурированных текстов с использованием аннотированных суффиксных деревьев», представленной на соискание ученой степени кандидата технических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ

1. Оценка актуальности темы диссертационной работы

В большинстве практических задач работы с текстами возникает необходимость сравнения двух текстов по их сходству в некотором смысле. Текст всегда рассматривается как конечная последовательность слов. Показатель степени сходства текстов принято называть их релевантностью. Как правило, при измерении релевантности двух текстов, рассматриваемых как две последовательности слов, сравниваются не сами тексты, а их представления на некотором специальном математическом языке, делающем сравнение наиболее удобным. Проблема выбора способа описания, наиболее удобного для сравнения текстов в некотором определенном смысле, называется проблемой представления текстов. В силу огромной сложности и неопределенности этой проблемы методология представления текстов непрерывно развивается.

Основной идеей диссертационной работы Е.Л. Черняк является разработка нового класса моделей представления текстов, названных теоретико-множественными моделями. Необходимость создания нового класса моделей продиктована следующими обстоятельствами.

1. Неотъемлемой составляющей любого способа сравнения текстов является сравнение входящих в них слов. В существующих моделях представления текстов любые два слова рассматриваются либо как полностью совпадающие, либо как совершенно разные. В то же время практика требует более тонкого сравнения слов в тексте, что, в свою очередь подсказывает рассматривать отдельное слово не как единое целое, а как последовательность символов (букв), вообще говоря, произвольной длины.

2. Соответственно, в дополнение к общепринятым функциям релевантности, порождаемыми существующими моделями представления текстов и названными в диссертации четкими, существует настоятельная необходимость создания более сложных функций релевантности, рассматривающих как «почти идентичные» слова, различающиеся несколькими буквами.

3. Существующие модели представления текста основаны на молчаливом предположении, что анализируется текст на конкретном языке (русском, английском и пр.), и при анализе текста необходимы как морфологический, так и синтаксический анализ. В

результате традиционные модели приводят к необходимости огромных вычислительных затрат на предварительную обработку текста. В то же время, в большинстве случаев целью массовой обработки текстов является получение весьма простой информации, сводящейся, в основном, к тематической классификации, что позволяет ограничиться более простыми моделями текстов.

4. В информатике вообще и биоинформатике в частности развит и широко используется существенно более простой аппарат анализа строковых структур, одним из элементов которого является суффиксного дерева. Текстовые данные, по сути, являются строками или последовательностями строк, но методология суффиксных деревьев редко используется для представления текстовых данных.

Именно последнее обстоятельство является отправной точкой рецензируемой диссертации. Ее центральная задача сформулирована как совершенствование методологии тематической классификации текстов на основе использования аппарата аннотированных суффиксных деревьев как инструмента, промежуточного по своей сложности между синтаксическим анализом и простейшей концепцией «мешка слов». Такую постановку задачи следует признать в высшей степени актуальной.

2. Содержание работы

Во введении обоснована актуальность построения новой теоретико-множественной модели и выдвинуты требования к ассоциированной мере релевантности, а также сформулированы цели и задачи проводимых исследований, положения, выносимые на защиту, приведены сведения о структуре диссертации, ее апробации и практическом использовании полученных результатов.

Первая глава посвящена обзору существующих моделей представления текстов, их сфер использования, достоинств и ограничений. Показаны существенные недостатки рассмотренных моделей, а именно, невозможность их использования для представления текстов в виде последовательностей строк произвольной длины и учета нечетких совпадений.

Во второй главе формулируется общая схема решения задач автоматического анализа текстов, использующая понятие релевантности строки тексту. Показано, что основным построением является так называемая таблица релевантности строка – текст и приведен обзор существующих мер релевантности, ассоциированных с моделями, перечисленными выше. Особое внимание уделяется ограничениям, связанным с использованием этих моделей релевантности: необходимость представления текста в виде набора слов или буквенных последовательностей фиксированной длины и невозможность учета нечетких совпадений. Показано, что большая часть рассматриваемых мер релевантности зависит от длины текста, что тоже является существенным недостатком. В этой же главе представлен метод аннотированного суффиксного и дерева и продемонстрировано его использование для построения таблиц релевантности, начи-

ная от алгоритма построения аннотированного суффиксного дерева, заканчивая вычислением нормированных оценок релевантности, независящих от длины текста.

В третьей главе проводится сравнение результатов использования косинусной и вероятностной меры релевантности с предлагаемой мерой релевантности, основанной на аннотированных суффиксных деревьях. Показано превосходство последней в почти полтора раза по сравнению с косинусной и вероятностной мерами релевантности.

В четвертой главе продемонстрировано использование меры релевантности в задаче пополнения таксономии темами Википедии. Мера релевантности используется для очистки дерева категорий Википедии от шума, а также для определения релевантных тем Википедии основе таксономии. За неимением золотого стандарта проводится экспертное оценивание качества таксономии, результаты которого подтверждают целесообразность и эффективность предложенного метода.

В пятой главе сравниваются различные меры релевантности в очистке текста от обценной лексики. Предложенная мера релевантности, основанная на аннотированных суффиксных деревьях, демонстрирует лучшие результаты как по полноте, так и по временной сложности.

В шестой главе приводится описание комплексов программ, использованных для всех экспериментов и их организация по принципу пайплайна.

Общие теоретические и практические результаты сформулированы в заключении.

Диссертация написана хорошим научным языком и хорошо читается. Большое количество таблиц убеждает в достоверности полученных результатов, а иллюстрации удачно дополняют текстовую часть. Основное содержание диссертации опубликовано в печатных работах автора и достаточно полно отражено в автореферате.

3. Научная новизна

В диссертации получен ряд новых научных результатов.

Использование понятия аннотированного суффиксного дерева позволило разработать принципиально новый метод численного измерения степени вхождения заданной текстовой строки (слова или последовательность слов) в анализируемый текст, что на языке анализа текстов и называется релевантностью строки тексту. В основу метода, позволяющего значительно увеличить гибкость измерения, положен разработанный в диссертации новый принцип вычисления оценок релевантности строки тексту, названный средней условной вероятностью символа в совпадении (СУВСС).

Безусловная новизна этого принципа определяет и новизну целого ряда методов классификационного анализа текстов, разработанных в диссертации.

Во-первых, предложен метод рубрикации научных статей с использованием критерия релевантности СУВСС, который, как показывает проведенное в диссертации исследование, является более точным, чем мировые аналоги.

Во-вторых, разработан оригинальный метод пополнения таксономии с использованием Википедии в качестве внешнего источника данных.

В-третьих, предложен метод фильтрации обценной лексики, более полной и быстрой, чем в стандартных фильтрах.

Эти методы реализованы в виде двух комплексов программ, не имеющих аналогов в мировой практике.

Первый комплекс, осуществляющий извлечение данных из Википедии, работает в интерактивном режиме, позволяя пользователю находить в Википедии интересующие его категории.

Второй комплекс осуществляет предобработку текстов, строит аннотированное суффиксное дерево, находит оценки релевантности для входной коллекции строк, и выдает итоговую таблицу релевантности «строка-тест». Комплекс построен по иерархической методологии пайплайн.

4. 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 подтверждается их успешным применением в реальных проектах компании «ФОРС-Центр разработки».

5. Замечания по диссертации

1. Не освящен вопрос оптимальной организации вычислений, например, параллелизации при очевидной низкой вычислительной эффективности метода аннотированного суффиксного дерева.

2. В диссертации приведены результаты трех групп экспериментов, иллюстрирующих три разные задачи анализа текстов – рубрикация научных публикаций, пополнение таксономии научного знания и фильтрация обценной лексики. Очевидно, что язык текста (русский, английский, немецкий) оказывает существенное влияние на характер результата анализа. Однако в диссертации эксперименты с каждым из видов анализа текста проведены только на одном языке, русском либо английском, и фактически нет экспериментов по влиянию языка на результат анализа текста.

3. К сожалению диссертация не свободна от опечаток, например:

стр. 15: «в керне *не верно* с точки зрения...»,

стр. 54: «англоязычные статьи *в* из области... ».

стр. 72: «подходы к построению или *пополнения*...».

4. Заключительная оценка

Указанные недостатки не снижают научной и практической значимости работы. Тема диссертационной работы Черняк Е.Л. «Разработка вычислительных методов анализа неструктурированных текстов с использованием аннотированных суффиксных деревьев», безусловно, актуальна. Работа написана убедительно, основные результаты диссертации опубликованы в ведущих научных журналах. Автореферат достаточно полно отражает основное содержание диссертации.

Диссертационная работа является законченной научно-квалификационной работой, в которой на основании выполненных автором исследований разработаны новые математические методы моделирования объектов, с применением современных вычислительных технологий разработаны и исследованы эффективные численные методы и алгоритмы, которые реализованы в виде проблемно-ориентированного программного комплекса. В работе содержатся новые научные результаты, представляющие существенный вклад в науку и имеющие важное практическое значение. В целом результаты диссертации можно квалифицировать как значимое научное достижение в области автоматической обработки текстов.

Считаю, что диссертационная работа полностью отвечает требованиям, предъявляемым ВАК к кандидатским диссертациям, соответствует специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ», а ее автор Черняк Е.Л. заслуживает присуждения ей ученой степени кандидата технических наук.

Официальный оппонент

доктор технических наук, профессор,

ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН (Вычислительный центр РАН)

Почтовый адрес: Москва, 119333, ВЦ РАН, ул. Вавилова, д.40.

Тел. (903) 659 0812, e-mail: vmottl@ccas.ru

В.В. Моттль

