

ОТЗЫВ

на автореферат диссертации

Екатерины Леонидовны Черняк

«Разработка вычислительных методов анализа неструктурированных текстов

с использованием аннотированных суффиксных деревьев»

по специальности 05.13.18 «Математическое моделирование, численные методы и

комплексы программ»

Актуальность темы, освещаемой в диссертации, обоснована в автореферате потребностью в разработке методов, не требующих предварительной разметки текстов.

Научная новизна диссертации подтверждена рядом новых результатов: теоретико-множественной моделью совокупности «строка-текст» с новым эффективным методом оценки соответствия строк тексту; методом рубрикации научных статей на основе предложенного автором критерия релевантности; методом пополнения предметных таксономий на основе справочных материалов интернета; комплексами программ, реализующими предложенные методы и модели.

Практическая значимость подтверждена успешным применением методов и алгоритмов, разработанных в рамках диссертации, в реальных проектах, а также в преподавательской деятельности автора.

Работа логично структурирована и аккуратно оформлена. Содержание диссертации полно отражено в публикациях, в том числе в изданиях из списка ВАК.

В качестве недостатков работы можно указать следующее:

- На наш взгляд, *модель релевантности* – не очень удачный термин, в зависимости от приложения лучше использовать *модель информационного поиска* (*information retrieval model*) или *мера близости текста* (*text similarity measure*). *Релевантность* – это соответствие документа информационной потребности пользователя (а не тексту запроса).
- Автор обосновывает актуальность предложенного метода недостатками классических моделей информационного поиска (BM25, языкового моделирования). Было бы логично и последовательно провести сравнение оценки предлагаемого метода и сравнение с другими моделями на задаче поиска документа по запросу, благо есть достаточное количество размеченных коллекций.
- Текст автореферата не содержит важных деталей оценки предложенного метода в задаче рубрикации аннотаций научных статей. Например, непонятно, почему были выбраны метрики MAP и nDCG, предназначенные для оценки ранжированного поиска, или, почему не были использованы стандартные коллекции для оценки методов классификации текстов (например,

RCV1), которые позволили бы сравнить эффективность разработанного метода с более широким спектром аналогов.

- Наряду с определением аннотированного суффиксного дерева (АСД), в автореферате дается 2 свойства АСД. При этом нет каких-либо пояснений о важности этих свойств и их применении в работе.
- Не указана оценка сложности наивного алгоритма построения АСД и, соответственно, не приводится сравнения по сложности модификации алгоритма Укконена и наивного алгоритма.
- Из описания четвертой главы неясно, почему согласованность разных частей исследования измеряется по-разному. Логично было бы в обоих случаях приводить и процент согласия, и каппу Козна. Значение каппы Козна 0,319 говорит о низкой согласованности экспертов.
- Результаты сравнения методов в описании пятой главы было бы логично оформить в виде таблицы. Не очень понятно, что являлось единицей оценки – отдельное слово или весь документ?

Указанные недостатки, однако, не снижают научную ценность работы и не влияют на ее общую положительную оценку.

На основании знакомства с авторефератом можно заключить, что работа соответствует требованиям Положения о порядке присуждения ученых степеней и паспорту научной специальности, а ее автор достоин присуждения степени кандидата технических наук.

м.н.с. Института математики и компьютерных наук Уральского федерального университета

А.В. Белобородов

с.н.с. Института математики и компьютерных наук Уральского федерального университета,
канд. техн. наук

П.И. Браславский



Подпись

А.В. Белобородов, П.И. Браславский

Заверяю

Институт математики и компьютерных наук
Уральский федеральный университет
Имени первого Президента России
Б. Н. Ельцина
Ул. Тургенева, 4, Екатеринбург,
Свердловская Область, 620075

