

ОТЗЫВ

Научного руководителя о диссертации Е. Л. Черняк на тему «Разработка вычислительных методов анализа неструктурированных текстов с использованием аннотированных суффиксных деревьев», представляемой на соискание учёной степени кандидата технических наук по специальности 05.13.18 — Математическое моделирование, численные методы и комплексы программ

Перед диссертантом была поставлена задача формирования модельно-алгоритмического аппарата анализа неструктурированных научных текстов, рассматриваемых как последовательности символов языка, прежде всего, русского языка, и проверки того, насколько такой аппарат применим в реальных ситуациях. Дело в том, что популярные методы анализа неструктурированных текстов основаны на использовании знаний о тех или иных свойствах, иногда грамматических, а иногда даже и семантических, языка и тем самым привязаны к определенному естественному языку и связанными с ним правилами. Мне же хотелось иметь аппарат, который можно применять к системе текстов, не очень заботясь об их языковой привязке. Вопросом было, насколько сложные задачи можно решать без использования знаний о конкретной языковой структуре.

Е. Л. Черняк разработала такой аппарат на основе теоретико-множественной модели совпадений в системе строка-текст, привязанный к апробированному вычислительному средству – суффиксным деревьям, но удачно избегающим недостатков, свойственным ранним реализациям подходов с использованием суффиксных деревьев. Здесь прежде всего стоит отметить предложенную Е.Л. Черняк предварительную разбивку текста на короткие – но осмысленные, потому что следуют структуре пробелов между словами – фрагменты, позволяющую резко сократить требования к объему памяти, необходимой для хранения создаваемых деревьев. Дальнейшее улучшение вычислительных свойств удалось получить путем разделения этапов построения и аннотации суффиксных деревьев частотами. Наиболее удачная инновация – предложенная интуитивно осмысленная мера релевантности как сумма средних условных вероятностей символов в максимальных совпадающих фрагментах. Разработанный аппарат был применен к трем конкретным актуальным задачам анализа текстов – как на английском, так и на русском языках, без каких-либо изменений - и доказал свою эффективность путем сравнения с популярными аналогами из международной литературы. И в вычислительных экспериментах по рубрикации научных статей, и по выявлению нежелательных слов, методология, разработанная Е.Л. Черняк, и реализованная ею в программных комплексах, с большим отрывом победила популярные аналоги из международной литературы. Особняком стоит задача пополнения таксономии. При всей её важности, она относится не к наблюдаемому миру, а миру идей, в котором традиционные оценки качества результатов – такие как полнота и точность – не применимы. Здесь возможно только сравнение с результатами человеческой экспертизы. К несчастью, задача категоризации относится к числу наиболее трудных для человека, как мне не раз приходилось убеждаться при выполнении тех или иных проектов, в том числе международных. Поэтому из 6 специалистов, докторов и кандидатов наук, согласившихся провести экспертизу

результатов использования метода Е.Л. Черняк в задаче пополнения таксономии «теории вероятностей и математической статистики», довели работу до конца только двое.

В целом, считаю, что диссертантка Е.Л. Черняк не только успешно справилась с поставленной задачей, но и разработала при этом аппарат, который несомненно найдет дальнейшее применение в задачах анализа и интерпретации неструктурированных текстов.

Представленная диссертация соответствует специальности 05.13.18, например, в части 3. Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий. Описанная в ней работа с большим запасом покрывает требования к кандидатским диссертациям, а ее автор Е.Л. Черняк заслуживает присуждения ей учёной степени кандидата технических наук по этой специальности.

Научный руководитель,
Профессор Департамента Анализа данных и
искусственного интеллекта Факультета компьютерных наук
Национального исследовательского университета
«Высшая школа экономики» доктор технических наук ст. н. с.



Миркин Б.Г.

8 августа 2016 г.

Подпись заверяю

