

ОТЗЫВ

официального оппонента о диссертации **Ильвовского Дмитрия Алексеевича**
**«Методы и алгоритмы обработки текстовых данных на основе графовых
дискурсивных моделей»,**

представленной на соискание ученой степени кандидата технических наук
по специальности 05.13.18 «Математическое моделирование, численные методы и
комплексы программ»

Актуальность темы. Обработка текстов на естественном языке буквально пронизывает современные информационные технологии: поиск информации, классификация документов и выдача рекомендаций. В последние несколько десятилетий в этой области достигнут значительный прогресс. Опираясь на достижения современной лингвистики, специалисты по информатике смогли перейти от использования простейших методов морфологической нормализации (лемматизации) к сложным моделям синтаксического (и даже поверхностно-семантического) анализа предложений. Следующим естественным шагом в усложнении используемых моделей является «дискурсивный» уровень, когда строится модель не одного предложения, а целого абзаца. Хотя теоретиками-лингвистами разработаны (главным образом, для английского языка) различные модели дискурса (теория риторических структур, теория речевых актов), на практике еще следует создать специальные модели, алгоритмы и программы для использования этих лингвистических знаний при автоматической обработке текста. Рецензируемая диссертационная работа посвящена этой актуальной теме.

Оценка содержания диссертации

Диссертация написана ясным и четким языком и состоит из введения, пяти глав, заключения, списка литературы и приложений. Объем работы – 250 листов. Библиография включает 139 названий.

Введение

Во введении приведена общая характеристика работы, обоснована актуальность темы диссертации, ее научная новизна и практическая значимость, сформулированы предмет, цель и задачи исследования.

Глава 1.

В главе 1 даются теоретические основы используемых моделей и методов (анализ формальных понятий, проекции узорных структур, синтаксические и дискурсивные модели текста, определяется ключевое понятие работы – чаша разбора – обогащенное риторическими и коммуникативными связями множество деревьев разбора, ядра свертки для машины опорных векторов, онтологии).

Глава 2.

Эта глава описывает графовую модель абзацев текста, на базе которой развивается метод построения решетки замкнутых описаний с использованием проекций узорных структур для ускорения вычислительно-сложного алгоритма поиска сходств на множествах графов.

Глава 3.

В этой главе диссертант на основе модели «чаша разбора» предложил и исследовал специальные ядра свертки для текстов, которые позволяют применять к текстам на естественном языке широко-известный метод опорных векторов (SVM) В.Н. Вапника. Экспериментальное сравнение с ранее используемыми ядрами свертки, основанными на модели «мешок слов», показало преимущества предложенного Д.А. Ильвовским подхода.

Глава 4.

В этой главе исследуется задача поиска тождественных денотатов для формальных описаний фрагментов текста. Эта задача очень полезна при применении к онтологиям. Здесь снова применяется методы анализа формальных понятий, но здесь это – процедура фильтрации элементов решетки замкнутых описаний относительно индексов устойчивости.

Глава 5.

В этой главе описывается построенный в рамках диссертационного исследования программный комплекс, предназначенный для работы с чашами разбора, которые автоматически порождаются из текстов с как использованием программ с исходными текстами, так и созданных Д.А. Ильвовским на основании лингвистической теории, а также модули интеллектуального анализа данных, разрабатываемых в НИУ-ВШЭ командой программистов, одним из активных участников которой является диссертант.

Заключение

В заключении приведены общие выводы из работы и сформулированы перспективные направления дальнейших исследований.

Приложения.

В приложениях (около 80 страниц) приводятся основные фрагменты кода программной системы.

Оценка автореферата диссертации

Автореферат диссертация написан по стандартной схеме ясным и четким языком. Содержание диссертации отражено точно и полно.

Достоверность и новизна полученных результатов. Для достижения поставленной в работе цели диссертантом использовался широкий круг методов анализа формальных понятий, метода ядер сверток, онтологий и лингвистических моделей для дискурса. Графовая модель фрагмента текста — чаша разбора — получила свою алгоритмическую реализацию. Все полученные Д.А. Ильвовским результаты являются новыми и оригинальными.

Ценность для науки и практики. Полученные результаты представляются ценными для практического использования. Можно надеяться, что разработанные соискателем методы приведут к дальнейшему развитию информационного поиска, рекомендательных систем и онтологий.

Подтверждение апробации работы и публикации результатов. Основные результаты диссертации своевременно опубликованы в научной печати. Следует отметить активное участие соискателя в национальных и международных конференциях по искусственному интеллекту и информационным технологиям с докладами и сообщениями.

Замечания к тексту диссертации и рекомендации по его улучшению.

1. Во введении при обзоре предыдущих подходов к проблеме не упомянуты работы отечественных исследователей.
2. Рекомендуется описание математических деталей частных подходов (ядра и онтологии) перенести из первой главы в соответствующие главы (3 и 4, соответственно). Тогда первая глава станет компактнее, а читателю будет удобно читать о применении ядер для коротких текстов и поиске

тождественных сущностей в онтологиях, ознакомившись с относящимися к делу определениями в первых параграфах соответствующих глав.

3. Рецензент находит странным порядок перечисления использованных источников в списке литературы.
4. В списке литературы некоторые отечественные работы процитированы на английском языке (В.К. Финн, С.А. Евтушенко, одна и та же работа которого цитируется 2 раза в [11] по-русски, а в [103] – по-английски).

Работа содержит несколько опечаток. Наиболее важные:

1. Страница 22: в определении 1.9 следует опустить индексы у знаков частичного порядка.
2. Страница 24: операция может быть алгоритмически (все оценки обычно асимптотические) сложной.
3. Страница 27: в формуле для $\text{idf}(t,D)$ содержится опечатка.

Соответствие уровня диссертации.

Указанные недостатки не снижают научной и практической значимости работы. Представленные Ильвовским Д.А. исследования отвечают цели работы и демонстрируют высокий научный уровень. Работа имеет важное практическое значение.

Диссертационная работа Д.А. Ильвовского является законченной научной работой, в которой на основании проведенных автором исследований разработаны новые математические методы моделирования текстов, разработаны и исследованы эффективные вычислительные методы и алгоритмы, которые реализованы в виде программного комплекса. В работе имеются новые научные результаты.

Считаю, что диссертационная работа полностью соответствует требованиям, предъявляемым ВАК к кандидатским диссертациям на соискание степени кандидата технических наук, а ее автор Ильвовский Дмитрий Алексеевич заслуживает присуждения ему искомой степени по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ».

Официальный оппонент

С. н. с. ИПИ РАН ФИЦ ИУ РАН, к. ф. -м. н.

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

e-mail: krrguest@yandex.ru

Тел.: 8(499) 135-01-40

Д.В. Виноградов

