

ОТЗЫВ

официального оппонента на диссертацию А.О. Исхаковой «Метод и программное средство определения искусственно созданных текстов», представленную на соискание ученой степени кандидата технических наук по специальности 05.13.17 – Теоретические основы информатики

1. Актуальность темы.

В настоящее время в сети Интернет формируется и хранится огромное количество информации в виде не структурированного текста. При этом, текстовая информация может создаваться не только людьми, но и с помощью специальных программных средств автоматической генерации текстов, которые позволяют формировать большое количество различных вариантов текста на основе некоторых данных или оригинального авторского текста. Такие искусственно созданные тексты могут быть направлены на достижение противоправных или непристойных целей. В связи с этим часто важным является принятие решения о том, создан ли некоторый текст человеком или он сформирован искусственно с помощью программ. На основе этого можно утверждать, что тема диссертационной работ Исхаковой А.О., посвященная разработке метода и программных средств определения искусственно созданных текстов, несомненно, является актуальной.

2. Содержание диссертации:

Диссертация Исхаковой А.О. состоит из введения, трех глав и заключения. В них содержится описание следующих выполненных исследований.

Рассмотрены методы текстовой атрибуции – определения набора показателей, которые могут быть численно рассчитаны для конкретного текста. Базовым предположением является то, что у разных источников текста имеется устойчивый (характерный) набор таких показателей (профиль, стиль, инвариант автора). Проанализированы два метода формирования параметров текста (атрибуции): статистический и с использованием искусственного интеллекта.

Рассмотрены способы формирования искусственных текстов: на основе цепей Маркова и на основе SIMP-таблиц, с использованием словарей и на основе синонимизации.

На основе обзора существующих подходов определения искусственно созданных текстов автором сделана формализованная постановка задачи определения искусственных текстов на основе инвариантов и разработан алгоритм формирования инвариантов классов текстов. По утверждению автора данный алгоритм является универсальным (стр. 49) и может быть применен при решении любой задачи связанной с классификацией текстов.

С использованием разработанного алгоритма на основе трех выборок размером в 3210 текстов были рассчитаны инварианты (наборы показателей) естественного и искусственных текстов (сформированных методом цепей Маркова и синонимизации).

Разработан метод отнесения произвольного текста к одной из групп: естественный текст, искусственный текст, созданный с использованием метода цепей Маркова и искусственный текст, созданный с использованием метода синонимизации

Разработана программная система выделения (фильтрации) искусственно созданных текстов в виде web-приложения с использованием языка программирования PHP.

Материал диссертации изложен достаточно ясно, диссертация хорошо оформлена.

3. Степень обоснованности и достоверности научных положений, изложенных в диссертации достаточно высокая и обеспечивается корректностью использования теории множеств, математической статистики и методов функционального и математического моделирования. Достоверность полученных результатов также подтверждается путем согласования полученных результатов с исследованиями других авторов и результатами выполненных вычислительных экспериментов.

4. Полнота опубликования результатов диссертации, соответствие автореферата содержанию диссертации.

Основные результаты диссертационной работы изложены в 11 публикациях, в числе которых 5 статей в рецензируемых журналах, рекомендованных ВАК для опубликования научных результатов диссертаций, а также докладывались на 9 международных и российских конференциях. Таким образом, полнота опубликования результатов диссертации полностью соответствует требованиям ВАК. Автореферат содержит все основные положений диссертационной работы.

5. Теоретическое значение диссертации состоит в следующем:

- Создан новый метод определения искусственно созданных текстов, отличающийся использованием меры принадлежности входного текста к известным классам текстов и позволяющий определить его происхождение.
- Разработан алгоритм формирования инвариантов классов текстов, отличающийся от существующих использованием качественных различий между классами для создания наборов характеристик.
- Определены новые инварианты для текстов, созданных с помощью методов цепей Маркова и синонимизации, с использованием разработанного алгоритма.

6. Практическое значение диссертации состоит в том, что полученные результаты диссертационного исследования могут использоваться при разработке систем поиска и обработки информации в сети Интернет. Кроме того, на основе полученных результатов разработаны программы определения искусственных текстов. Получено свидетельство об официальной регистрации программы расчета статистических значений текстовых характеристик. Результаты работы использовались при выполнении гранта РФФИ № 16-47-700350.

7. Замечания и недостатки диссертации

Работа в целом выполнена на высоком научном уровне, однако, при чтении диссертации возникает ряд вопросов и замечаний:

1. Нет описания практического применения результатов работы в организациях, которые предоставили акты о внедрении (ООО «Агентство медиа-решений», ООО «Лингвистические и информационные технологии»). Есть только результаты тестирования метода определения искусственных текстов.
2. В тексте диссертации в качестве классов искусственных текстов, которые анализируются, выбраны только тексты, сформированные двумя методами (цепи Маркова и синонимизации). Не понятно, как будет обработан текст, созданный с помощью использования SIMP-таблиц и словарей. К какому классу он будет отнесен?
3. В тексте диссертации используется понятие «семантическое ядро», но нет его точного определения.
4. В тексте Главы 2 диссертации нет явного описания алгоритма расчета таких количественных показателей, как «Количество слов в семантическом ядре» и «Наличие единства тематики в разных частях текста» (хотя сами числовые значения приведены).
5. Журнал «Ползуновский вестник» есть в списке изданий, рекомендуемых ВАК, но у него не указана специальность 05.13.00, по которой могут публиковать статьи соискатели степени к.т.н.

8. Заключение

Не смотря на сделанные замечания считаю, что диссертация Исхаковой А.О. представляет собой законченную научно-исследовательскую работу, выполненную на актуальную тему. Новые научные результаты, полученные диссертантом, имеют существенное значение для разработки математического и программного обеспечения для решения актуальных задач разработки и исследования методов и алгоритмов анализа текста, извлечения данных из текстов на

естественном языке, а также разработки новых интернет-технологий. Исходя из этого, данная диссертация соответствует 2, 5, 6 и 9 пунктам паспорта специальности 05.13.17.

В целом считаю, что диссертационная работа *Исхаковой Анастасии Олеговны* удовлетворяет требованиям п. 9 «Положения о порядке присуждения ученых степеней» утвержденного Постановлением правительства РФ от 24.09.2013 №842, а ее автор заслуживает присуждения ему ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики».

Профессор кафедры Программная инженерия
Института кибернетики Национального исследовательского
Томского политехнического университета,
д.т.н., профессор по специальности 05.13.11

А.Ф. Тузовский

2 декабря 2016 г.

Подпись д.т.н., профессора Тузовского А.Ф. заверяю

Ученый секретарь ученого Совета
Института кибернетики Национального исследовательского
Томского политехнического университета



О.А. Ананьева

Тузовский Анатолий Федорович
профессор кафедры «Программной инженерии»
Института кибернетики Национального исследовательского
Томского политехнического университета
адрес: 634034, г. Томск, ул. Советская, 84/3, ком. 316
тел.: (3822) 606132
e-mail: tuzovskyaf@tpu.ru