О Т З Ы В

на автореферат диссертационной работы Исхаковой Анастасии Олеговны «Метод и программное средство определения искусственно созданных текстов», представленной на соискание учёной степени кандидата технических наук по специальности 05.13.17 «Теоретические основы информатики»

Алгоритмы автоматизированной обработки текста широко используются в современных информационных системах для проведения семантического анализа структуры текста, в DLP-системах (системах предотвращения утечки информации), для выявления плагиата и для других целей. Поэтому любая работа, направленная на дальнейшее развитие и совершенствование подобных алгоритмов, в частности, и работа, представленная в данном автореферате и посвященная разработке нового, более эффективного метода выявления в потоке информации текстов, сгенерированных автоматически, актуальна и представляет несомненный научный, методический и практический интерес.

В своей работе Исхакова А.О. для определения происхождения текстов предлагает использовать меру принадлежности анализируемого текста к известным классам, основанную на расчете геометрического расстояния между точками в n-мерном пространстве характерных признаков анализируемого текста. Одна из сопоставляемых точек соответствует входному тексту, а остальные - инвариантам некоторых известных классов. Созданное на основе предложенного подхода к классификации текста методическое обеспечение и его программная реализация были применены для идентификации происхождения искусственного текста, полученного в результате генерации методом синонимизации и методом цепей Маркова, и показали высокую достоверность идентификации разработанного метода. Важно отметить, что этот метод достаточно универсален и легко может быть адаптирован к другим видам текста и для решения других задач, связанных с обработкой текстовой информации. Практическую значимость выполненного исследования подтверждает и его внедрение в две организации, деятельность которых связана с обработкой текста, а также наличие зарегистрированного в установленном порядке программного обеспечения, предназначенного для расчета характеристик текста, необходимых для формирования его инвариантов. Среди других положительных особенностей работы нужно отметить и большую самостоятельность соискателя, что вытекает из большого числа работ, выполненных им без соавторства.

Замечания по автореферату

1. В качестве основной области применения разработанного метода идентификации в автореферате указывается выявление спама. Однако спам редко генерируется автоматически, часто представляется в графическом виде и его текстовая составляющая редко превышает 200 - 500 символов, тогда как для получения репрезентативных статистических оценок текста этот объем должен быть в разы больше (на стр.15 указывается 1200 символов). В то же время другие, упомянутые на стр.15 возможные области применения метода (системы поиска и выявления плагиата) в работе не исследовались.

2. На стр. 15 отмечается, что выходными данными метода являются заключение о происхождении текста и точность этого заключения, но никак не поясняется, в какой форме представлено это заключение и что автор понимает под точностью заключения? Естественно предположить, что это не расстояние между точками в многомерном пространстве, а вероятность ошибок первого и второго рода. Но тогда неясно, как они находятся для конкретной реализации анализируемого текста?

3. Из автореферата неясно, почему для идентификации использовался расчет расстояния между точками в многомерном пространстве, а не более эффективные методы, например, опорных векторов, главных компонент или скрытых цепей Маркова? Не раскрыты в нем и такие теоретические вопросы, как влияние вида предметной области на значения компонент инвариантов текстов (в частности, как автоматически вычисляется зависящая от предметной области частотность уникальных и популярных слов), наличие среди них компонент, прини-

мающих интервальные или нечисловые значения, влияние степени рассеяния компонент на достоверность идентификации.

4. Крайне скупо в автореферате представлена информация о практическом применении метода и его программной реализации. В частности, неясно, каким образом разработанный и представленный на рисунке 2 программный продукт интегрировался в аналитические системы, какие функции он там выполнял и к каким видам текста применялся?

5. Имеются отдельные недостатки в содержательной части автореферата. Так, на стр. 16 в третьей строке сверху написано "… значение максимальное значение …", а на рисунке 2 между базой данных и аналитическим модулем стрелка должна быть двунаправленной (поскольку этот же программный продукт используется и для формирования записываемых в базу параметров инвариантов).

Тем не менее, перечисленные замечания в большей части носят не принципиальный, а дискуссионный характер, работа имеет несомненное научное и практическое значение для специалистов в области автоматизированной обработки текстовой информации. Содержание автореферата позволяет утверждать, что в целом работа выполнена на высоком научном уровне, соответствует требованиям к кандидатским диссертациям, изложенным в п.9. Положения о присуждении ученых степеней в редакции постановления Правительства РФ от от 21.04.2016 № 335, а ее автор достоин присуждения ему учёной степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики».

Зав. кафедрой информатики, вычислительной
техники и информационной безопасности,
зав. лабораторией информационно-измерительных
систем ФГБОУ ВО «Алтайский государственный
технический университет им. И.И. Ползунова»,
д.т.н., профессор
656038, Барнаул, пр. Ленина, 46
тел. +7(3852) 290-786,
E-mail: yakunin@agtu.secna.ru

Алексей Григорьевич Якунин

Подпись заверяю: