На правах рукописи

ИСХАКОВА АНАСТАСИЯ ОЛЕГОВНА

МЕТОД И ПРОГРАММНОЕ СРЕДСТВО ОПРЕДЕЛЕНИЯ ИСКУССТВЕННО СОЗДАННЫХ ТЕКСТОВ

05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Томск – 2016

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Томский государственный университет систем управления и радиоэлектроники» (ТУСУР)

Научный руководитель – доктор технических наук, профессор
Мещеряков Роман Валерьевич

Официальные оппоненты: Тузовский Анатолий Федорович,
доктор технических наук, профессор,
профессор кафедры оптимизации систем управления Института кибернетики Национального исследовательского Томского политехнического университета»

Крючкова Елена Николаевна,
кандидат физико-математических наук, доцент,
профессор, заместитель заведующего кафедрой прикладной математики Алтайского государственного технического университета им. И.И. Ползунова,
г. Барнаул

Ведущая организация – Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский институт информатики и автоматизации» Российской академии наук

Защита состоится 22 декабря 2016 г. в 15 часов 15 мин. на заседании диссертационного совета Д 212.268.05 в ТУСУРе по адресу: 634050, г. Томск, пр. Ленина, 40, ауд. 201.

С диссертацией можно ознакомиться в библиотеке ТУСУРа по адресу: 634045, г. Томск, ул. Красноармейская, 146 и на сайте ТУСУРа по адресу: https://storage.tusur.ru/files/51408/dissertation.pdf

Автореферат разослан « ____ » _____ 2016 г.

Ученый секретарь
диссертационного совета  Костюченко Евгений Юрьевич

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Использование текстовой формы представления данных для хранения и передачи сведений используется во всех сферах деятельности. Создание текста долгое время было связано исключительно с умственной деятельностью человека – его автора. На сегодняшний день тексты выполняют не только функции хранения, накопления и передачи информации. Благодаря развитию способов моментального обмена и возможности распространения данных посредством сети Интернет, авторы имеют широкие возможности донести информацию до большого количества читателей.

Объемы создаваемой текстовой информации за последние десятилетия постоянно возрастают, об этом свидетельствует развитие дата-центров, интернет-ресурсов различного назначения, электронного документооборота и т.д. Вместе с тем создание самих текстов уже не является уникальной прерогативой человека. Специальные алгоритмы и программные средства позволяют генерировать тексты автоматически на основе некоторых исходных данных. **В диссертационной работе тексты, созданные автоматически с помощью специальных алгоритмов или программных генераторов, называются искусственно созданными или искусственными.**

Методы создания искусственных текстов позволяют генерировать множество уникальных экземпляров на основе некоторого авторского произведения или на основе модели формирования текста-результата, которая может быть представлена в виде шаблонов на базе словарей, сформированных конструкций предложений и др. Такие методы широко применяются для создания информационного контента, так как размещаемая в сети Интернет информация должна обладать достаточным уровнем уникальности, чтобы сайт был проиндексирован поисковыми системами и пользователи смогли его найти.

Интернет и различные инфокоммуникационные технологии играют значительную роль в общественных отношениях и взаимодействиях на всех уровнях. В этой связи информация, распространяемая в сети, зачастую воспринимается человеком как современный аналог энциклопедии или справочника, а также телевидения и газет, которым люди привыкли верить. Однако есть основания полагать, что эта информация не всегда отражает действительность и может быть использована для введения пользователей Интернета в заблуждение, распространения заведомо ложной или подстрекающей информации.

Массовое автоматическое порождение текстов определенной направленности может быть нацелено на пропаганду различных идей, в том числе социального, политического, а также преступного характера. Кроме того, оно может быть использовано для манипуляции населением или парализации работы электронных ресурсов. Использование искусственных текстов в виде информационного контента популярных или специально разработанных для этого веб-ресурсов позволяет распространять уникальные публикации любого содержания в неограниченном количестве. Таким образом, учитывая неоспоримую значимость

Интернет-технологий в жизни человека, злоумышленники могут использовать их в собственных неправомерных целях.

Задача определения искусственных текстов напрямую связана с текстовой атрибуцией – исследованием текстов для получения сведений об условиях их создания. Данное направление получило свое развитие с возникновением исследований по определению авторства, однако на сегодняшний день текстовая атрибуция включает в себя более широкий спектр задач. Проблема анализа и выявления искусственных текстов пока не получила достаточной освещенности. В настоящий момент в литературе не представлено описание методов для выявления такого класса текстов. В то же время скорость распространения искусственно созданной текстовой информации постоянно возрастает.

Актуальность угроз общественной безопасности и тот факт, что на сегодняшний день задача автоматического выявления такого контента не решена, обусловливают необходимость создания метода определения искусственно созданных искусственных текстов.

Целью диссертационной работы является повышение точности определения искусственно созданных текстов за счет создания авторского метода и программного средства.

Объектом исследования являются простые текстовые сообщения, сгенерированные автоматически на основе специальных алгоритмов и исходные естественные тексты.

Предметом исследования являются методы и алгоритмы классификации текстов.

Для достижения указанной цели были сформулированы следующие задачи:

1) исследовать методы и алгоритмы создания инвариантов, оценить возможность их применения для создания инвариантов текстов, сгенерированных автоматически;

2) выделить инварианты для классификации естественных и искусственно созданных текстов;

3) разработать метод определения искусственно созданных текстов;

4) провести экспериментальную апробацию метода определения искусственно созданных текстов;

5) разработать программное средство, позволяющее определять искусственно созданные тексты.

Методы исследования. Для решения поставленных задач в диссертационной работе использовались методы функционального и математического моделирования, теории множеств, математической статистики, матричных вычислений.

Научная новизна проведенных исследований и полученных в работе результатов заключается в следующем:

1. Создан оригинальный метод определения искусственно созданных текстов, отличающийся использованием меры принадлежности входного текста к известным классам и позволяющий принять решение о происхождении текста.

2. Разработан новый алгоритм формирования инвариантов классов текстов, отличающийся от существующих использованием качественных и уточняющих их

количественных текстовых характеристик и позволяющий осуществить выбор компонентов инварианта на основе лингвистических особенностей текстов.

3. Предложены новые инварианты для текстов, созданных искусственно с помощью синонимизации и метода Марковских цепей, полученные с использованием авторского алгоритма и позволяющие провести классификацию текстов по способу их создания.

Практическая значимость. Практическая значимость диссертационной работы подтверждается использованием полученных в ней результатов для решения практических задач. Разработанный метод и основанное на нем программное средство позволяют решить задачи автоматизированной фильтрации интернет-контента, входящих сообщений и иных электронных текстов. Использование разработанного автором программного средства позволяет определить нежелательные страницы, содержащие искусственно созданный контент, оценить приходящие онлайн-запросы, а также идентифицировать потенциально опасные, вредоносные текстовые сообщения на электронных ресурсах.

Положения, выносимые на защиту:

1. Метод определения искусственно созданных текстов, основанный на расчете меры принадлежности входного текста к известным классам, позволяет определить, является исследуемый текст естественным или искусственным в 93 % случаев. *Соответствует п. 6 паспорта специальности 05.13.17 – Разработка методов, языков и моделей человекомашинного общения; разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения данных из текстов на естественном языке.*

2. Алгоритм формирования инвариантов классов текстов описывает последовательность действий для определения обладающих различительной способностью качественных, а также уточняющих их количественных характеристик текста, значения которых формируются в инварианты. *Соответствует п. 5 паспорта специальности 05.13.17 – Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.*

3. Инварианты естественных текстов и текстов, созданных искусственно с помощью синонимизации и метода цепей Маркова, позволяют представить различия данных классов текстов формально – в виде набора значений характеристик и использовать их при определении искусственных текстов. *Соответствует п. 2 паспорта специальности 05.13.17 – Исследование информационных структур, разработка и анализ моделей информационных процессов и структур.*

4. Программное средство фильтрации искусственно созданных текстов позволяет производить автоматическое определение способа созданиях входного текста. *Соответствует п. 9 паспорта специальности 05.13.17 – Разработка новых интернет-технологий, включая средства поиска, анализа и фильтрации информации, средства приобретения знаний и создания онтологии, средства интеллектуализации бизнес-процессов.*

Достоверность результатов обеспечивается строгостью применения математических методов, проверкой непротиворечивости и адекватности результатов, полученных как на промежуточных, так и на окончательных этапах работы, а также их согласованностью с результатами проведенных практических экспериментов.

Внедрение результатов работы. Результаты диссертационной работы были внедрены в деятельность ООО «Агентство медиарешений» и ООО «Лингвистические и информационные технологии», а также в учебную деятельность ТУСУРа по дисциплинам «Дискретная математика», «Теория вероятностей и математическая статистика».

Личный вклад. В диссертационной работе представлены результаты, в которых автору принадлежит определяющая роль. Часть опубликованных работ написана в соавторстве с научным руководителем.

Диссертант является автором метода определения искусственно созданных текстов, представленного в работе, а также разработчиком комплекса программ, состоящего из программы «Auth_stat» для расчета значений характеристик текста и программного средства «TextOrigin», реализующего фильтрацию искусственных текстов на основе разработанного автором метода.

Автором совместно с коллективами предприятий ООО «Агентство медиарешений» и ООО «Лингвистические и информационные технологии» проведены внедрение и апробация результатов работы. Постановка задачи исследования осуществлялась научным руководителем доктором технических наук, профессором Р.В. Мещеряковым.

Апробация работы. Основные результаты диссертационной работы докладывались на следующих конференциях:

1) Томские IEEE-семинары «Интеллектуальные системы моделирования, проектирования и управления», г. Томск, 2013–2016 гг.;
2) XIV Всероссийская конференция молодых ученых «Актуальные проблемы лингвистики и литературоведения», г. Томск, 2013 г.;
3) XIII Всероссийская конкурс-конференция студентов и аспирантов по информационной безопасности «SIBINFO-2013», г. Томск, 2013 г;
4) XVIII Всероссийская научно-техническая конференция студентов, аспирантов и молодых учёных «Научная сессия ТУСУР-2013», г. Томск, 2013 г.;
5) Межвузовская научно-практическая конференция «Актуальные проблемы инфосферы. Инфокоммуникации. Геоинформационные технологии. Информационная безопасность», г. Санкт-Петербург, 2013 г.;
6) Шестая международная конференция по когнитивной науке, г. Калининград, 2014 г.;
7) II Всероссийская научная интернет-конференция с международным участием «Современные системы искусственного интеллекта и их приложения в науке», г. Казань, 2014 г.;
8) 11 Международная научно-техническая конференция «Интерактивные системы: проблемы человеко-компьютерного взаимодействия», г. Ульяновск, 2015 г.;

9) Седьмая международная конференция по когнитивной науке, г. Светлогорск, 2016 г.

Результаты диссертационной работы использовались при выполнении проекта «Методы и алгоритмы идентификации моделей поведения объектов информационных инфраструктур для обеспечения безопасности государства, общества и личности», поддержанного грантом РФФИ № 16-47-700350.

Публикации по теме диссертации. Результаты диссертационной работы отражены в 11 публикациях, в том числе 5 публикаций в рецензируемых журналах из перечня ВАК, 6 публикаций в сборниках трудов конференций.

Объем и структура работы. Диссертация состоит из введения, трех глав, заключения. Полный объем диссертации составляет 123 страницы с 12 рисунками, 8 таблицами. Список использованных источников содержит 121 позицию.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационного исследования, сформулирована цель, определены основные задачи, научная новизна и практическая значимость полученных результатов, а также положения, выносимые на защиту.

В первой главе рассмотрены основные методы текстовой атрибуции, а также основанные на них методики определения авторства и определения искусственно созданных текстов. Проведен обзор методов, алгоритмов и программных решений для определения неестественных текстов, являющихся поисковым спамом и машинным переводом, выделены их особенности, оценена возможность их применение для определения простых интернет-сообщений, сгенерированных автоматически.

Задача определения искусственных текстов неразрывно связана с необходимостью выявления характерных языковых и стилистических особенностей в произведении, применением методов текстовой атрибуции. Определение авторства является наиболее исследованной областью применения текстовой атрибуции, поэтому используемые в ней методы и подходы могут быть применены для решения задачи данной диссертационной работы. В исследованиях по определению авторства долгое время использовались субъективные методы, например, оценка внешних деталей авторского стиля. При этом в первой половине XX в. доминировали филологические и историко-документальные методы исследования текстов. Однако середина прошлого столетия характеризуется появлением работ по определению авторства, в которых применялись математико-статистические методы. Развитие в конце 1970-х гг. методов автоматизации обработки текстов и расчета их численных характеристик ознаменовало новый этап в этом научном направлении и позволило добиться значительного повышения эффективности в определении авторства текстовых произведений.

Методы атрибуции в общем случае подразумевают исследование текстового произведения на пунктуационном, орфографическом, синтаксическом, стилистическом и лексико-фразеологическом уровнях. Это позволяет отразить основные свойства текста, выделяемые специалистами в области лингвистики:

символьные, лексические, синтаксические, семантические и тематические. Современные методы текстовой атрибуции текстов основаны на применении статистического анализа либо механизмов искусственного интеллекта.

По решению соответствующих задач системы атрибуции можно разделить на три класса:

– системы идентификации автора;

– системы диагностики (характеризации) автора;

– системы верификации автора (выявление подобия текстов).

В последнее десятилетие перечень задач, связанных с атрибуцией текстов, постоянно расширяется. Это обусловлено в первую очередь изменениями в способах создания, распространения и обработки текстового материала, а также его предназначения. На сегодняшний день среди множества научных работ не представлено общепризнанного и эффективного решения задачи определения искусственных текстов. Кроме того, рассматриваемая в данном исследовании задача однозначно не классифицирована. С одной стороны, она имеет схожие черты с задачей верификации авторства, так как требуется определить, имеет ли некоторый программный генератор отношение к созданию данного текста. С другой стороны, определение искусственно созданных текстов можно отнести к задаче идентификации авторства, так как в данном случае необходимо определить, каким из известных способов создан текст, либо сделать вывод о том, что это определить невозможно.

Автором предпринята попытка проанализировать возможность применения методов текстовой атрибуции для решения задачи определения искусственно созданных текстов.

Существующие решения в области определения неестественных текстов в первую очередь ориентированы на выявление поискового спама. Поисковым спамом называются приемы искусственного поднятия рейтинга и продвижения интернет-ресурсов. В основе подобных приемов лежат попытки обмана поисковых систем с целью повышения релевантности страниц и изменения упорядоченности результатов поиска. К инструментам создания поискового спама относят: дублирование веб-страниц, создание дорвеев, распространение ссылочной массы, манипуляции с текстом сайта путем применения алгоритмов массового порождения текстов.

Один из алгоритмов определения текстового спама, основанный на оценке разнообразия тематик документа, предложен в работах А.С. Павлова. В основе предложенной им методики лежит оценка характеристик текста, отличающих поисковый спам – большое количество ключевых слов, наличие скрытого текста, мелкого нечитаемого шрифта, применение тегов в заголовках и т.д. Использование данных характеристик не позволяет использовать эту методику для определения следов автоматического генератора на текстах, распространяемых как информационный контент, неблагоприятное воздействие которого направлено не на поисковые системы, а непосредственно на пользователей веб-ресурсов.

Также исследования веб-спама опубликованы в работах C. Castillo и D. Donato. В основе предложенной ими методики лежит анализ веб-контента на

наличие ссылок на определенные страницы и их распределения между страницами. В работах данных исследователей представлены схемы взаимодействия ссылок для паразитных «спам»-страниц и аналогичные – для «нормальных» страниц. Поскольку данная методика ориентирована на обнаружение веб-спама и имеет в своей основе особенности такого материала, ее применение для решения задачи данного диссертационного исследования не представляется возможным в связи с тем, что искусственный текст, предназначенный для прочтения человеком, не обладает указанными характеристиками.

В работе А.А. Зайцева, С.В. Кулешова и С.Н. Михайлова рассмотрен метод оценки качества текста на основе реферирования. Метод предполагает, что тексты, созданные с применением синонимизаторов или средств автоматического перевода, обладают меньшей тематической устойчивостью, поэтому при повышении компрессии «сжатия» реферата у них с большей скоростью уменьшается его размерность. В работе представлен ряд эвристических правил, которые позволяют исключить тексты низкого качества, тем самым доказывая, что тематические признаки текста позволяют выявить связные, информативные тексты.

В исследованиях текстов, созданных автоматически, часто уделяют внимание тематическим и семантическим свойствам текста. Для их оценки часто применяются для этого методы автоматической рубрикации. Например, в работе Е.В. Дунаева и А.А. Шелестова рассмотрены подходы к рубрикации текстовых данных на основе алгоритма PrTFID.

По итогам проведенного анализа существующих методов определения искусственно созданных текстов были сделаны следующие выводы:

1. Существующие методы определения неестественных текстов позволяют решать задачи выявления поискового спама, а также текстов, являющихся машинным переводом текста с одного естественного языка на другой.
2. Из-за наложенных ограничений и используемых особенностей рассмотренных текстов существующие методы не могут быть использованы в задаче определения текстов, созданных с помощью синонимизаторов или других средств автоматической генерации.
3. Для достижения большей точности в определении связности текста необходимо рассматривать все лингвистические уровни, в том числе синтаксический, на уровне предложений.
4. Качество текста напрямую связано с его тематическими свойствами, поэтому связанные с ними характеристики текста должны быть включены в инварианты.
5. При формировании набора текстовых характеристик, используемых для определения неестественных текстов, следует основываться на признаках текстов, обусловленных их лингвистическими особенностями, а также используемыми при генерации алгоритмами.

Таким образом, необходимо разработать метод определения текстов, сгенерированных автоматически, представляющих собой информационный веб-контент, а также алгоритм анализа характеристик данных текстов. Автором

предлагается использовать методы статистического анализа, так как они, в отличие от методов машинного обучения позволяют явно выделить устойчивость значения характеристик текста и применить к ним известные средства статистического анализа и в то же время являются эффективным инструментом текстовой атрибуции. При решении поставленной задачи необходимо учесть достижения исследований по оценке качества текстовых произведений и определению поискового спама. Также необходимо принять во внимание достижения в области выделения характерных черт естественных текстов и автоматической атрибуции текстов, поскольку они косвенно относятся к решению задачи определения массово порожденных текстов.

Вторая глава посвящена формированию инвариантов классов текстов. Дано формальное описание инварианта и задачи классификации текстов в целом. Предложен алгоритм формирования инвариантов, основанный на использовании качественных и уточняющих их количественных характеристик текста. В главе подробно описан процесс формирования инвариантов естественных и искусственных текстов, созданных с помощью синонимизации и метода Марковских цепей.

Классификация текстов, то есть определение соответствия его некоторому автору (группе авторов), классу, может быть представлена в формальном виде, основанном на теоретико-множественном подходе.

T – множество всех текстов.

$A = \{a_1, a_2, ..., a_n\}$ – множество инвариантов классов текстов в рамках решаемой задачи, $|A| = n$.

T_a – множество текстов, которым сопоставлен некоторый инвариант, то есть тексты известного авторства / класса.

T' – множество текстов, которым не сопоставлен инвариант, то есть совокупность текстов неизвестного авторства / класса.

При этом выполняется: $T = T_a \cup T'$; $T_a \cap T' = \varnothing$.

X – множество исследуемых текстовых характеристик, множество определяется следующим образом: $X = \{x \mid x$ – изучаемая текстовая характеристика$\}$; $|X| = m$.

Инвариант $a_j \in A$ представляет собой массив упорядоченных пар вида: <текстовая характеристика $x_i \in X$; значение текстовой характеристики x_i для данного инварианта z_{ij}>:

$$a_j = (\langle x_1, z_{1j}\rangle, \ldots, \langle x_i, z_{ij}\rangle, \ldots, \langle x_m, z_{mj}\rangle),$$

где $i = 1..m$; $j = 1..n$; z_{ij} – некоторое числовое значение, формат и диапазон которого выбраны в соответствии с текстовой характеристикой, в отдельных случаях в качестве z_{ij} могут выступать диапазоны значения характеристики: $z_{ij} = [z_{ij\,min}; z_{ij\,max}]$.

Тогда на декартовом произведении множества текстов T и множества инвариантов A может быть задано бинарное отношение $R \subset T \times A$ такое, что выполняется tRa, если некоторый текст $t \in T$ соответствует инварианту $a \in A$, то есть текст t относится к классу, которому соответствует инвариант a, или текст t написан автором, которому соответствует инвариант a.

Учитывая все предшествующие обозначения, можно записать условие того, что некоторому входному тексту $t \in T$ соответствует инвариант $a \in A$. Выполняется отношение tRa, если значения текстовых характеристик исследуемого текста t соответствуют или приближены в определенной степени к значениям характеристик $x_i \in X$ инварианта a. При этом степень приближенности значений устанавливается автором соответствующего метода в каждом конкретном случае и должна быть обоснована экспериментальными данными.

В отличие от инварианта автора инвариант метода создания текста (генератора искусственных текстов) должен быть применим для искусственного текста, созданного на основе любого авторского материала. Особенность формирования инвариантов в задачах атрибуции текстов состоит в том, что наборы исследуемых характеристик прежде всего зависят от непосредственно решаемой задачи. Также следует отметить, что любая задача, связанная с атрибуцией текста, является междисциплинарной и связывает знания в области лингвистики, стилометрии, информатики, статистики и других областей знаний. Инвариант позволяет формализовать отличия текстов одного класса и представляет собой модель данного класса в виде набора значений определенных характеристик текста. В связи с этим формирование инварианта должно быть основано в первую очередь на лингвистических особенностях языка и текста.

В соответствии с приведенными особенностями исследования текстов автором предложен алгоритм формирования инвариантов классов текстов, который отличается наличием процедуры выделения качественных признаков, различающих исследуемые классы текстов. Другими словами, при формировании набора исследователь основывается на лингвистических особенностях рассматриваемых классов текстов. Алгоритм может быть применен для создания инвариантов при решении любой задачи, связанной с классификацией текстов.

Предложенный автором алгоритм позволяет сформировать множество количественных характеристик текста на основе качественных признаков, которые выделяются на основе лингвистических особенностей исследуемых классов текстов.

При этом входными данными для формирования инвариантов являются:

– множество классов текстов K, для которых необходимо сформировать инварианты; классы разделяют множество текстов T на соответствующие подмножества по некоторому устойчивому и однозначному свойству, например по способу создания, автору, некоторому признаку автора;

– наборы текстов T_i ($i = 1..n$) для каждого класса из множества K.

В ходе формирования инвариантов согласно предложенному алгоритму происходит последовательное выполнение следующих процедур.

1. Определение качественных признаков исследуемых классов текстов с учетом лингвистических особенностей языка. Данные признаки должны отражать основные отличия между текстами разных выборок.

2. Уточнение свойств текста, которые позволяют определить проявление выделенных качественных признаков. Данный этап также основан на использовании знаний в области лингвистики и подразумевает использование

свойств, которые обобщают особенности текста на уровне одних и тех же текстообразующих элементов.

3. Формирование множества количественных характеристик X, позволяющих оценить проявление выделенных на предыдущем этапе свойств текста.

Последующие этапы алгоритма носят расчетный характер. Расчеты производятся на основе математического аппарата, применимость которого должна быть обоснована опытом исследований в рассматриваемой области, а также логическими суждениями и внутренней непротиворечивостью производимых вычислений.

4. Расчет математического ожидания значения характеристик $x_j \in X$ для текстов из набора T_i производится циклично для каждой характеристики, для каждой выборки ($i = 1..n$, $j = 1..m$).

5. Анализ значений характеристик $x_j \in X$ на их различительную способность между исследуемыми наборами текстов производится в цикле при $j = 1..m$. Для измерения различительной способности применяется мера, позволяющая оценить различие между значениями характеристики для наборов текстов разных классов.

6. В случае если характеристика $x_j \in X$ не удовлетворяет условию различительной способности, она исключается из множества исследуемых характеристик.

7. Анализ характеристик на корреляционную зависимость проводится для всех $x_j \in X$ внутри каждого набора текстов одного класса.

8. Удаление характеристик, коррелирующих с другими характеристиками внутри наборов текстов, позволяет окончательно сформировать множество характеристик X.

Результатом выполнения всех этапов процесса являются инварианты классов текстов, объединенные множеством A и представляющие собой наборы численных значений характеристик текста множества X.

Разработанный алгоритм формирования инвариантов классов текстов отличается от существующих использованием качественных и уточняющих их количественных текстовых характеристик, основанных на лингвистических особенностях языка. Блок-схема алгоритма представлена на рис. 1.

Известные аналогичные алгоритмы в качестве входных данных используют стандартные наборы количественных характеристик текстов, значения которых анализируются на степень различительной способности. При таком подходе значительно возрастает вычислительная сложность расчетов, так как количество всевозможных характеристик может составлять несколько тысяч. Также существует риск упущения текстовых характеристик, которые отсутствуют в стандартных наборах, но в конкретном случае могут обладать различительной способностью.

На основе предложенного алгоритма были выделены лингвистические особенности естественных текстов, отличающие их от искусственных экземпляров. Согласно всем шагам алгоритма были проведены формирование перечня количественных характеристик текста, а также оценка их различительной

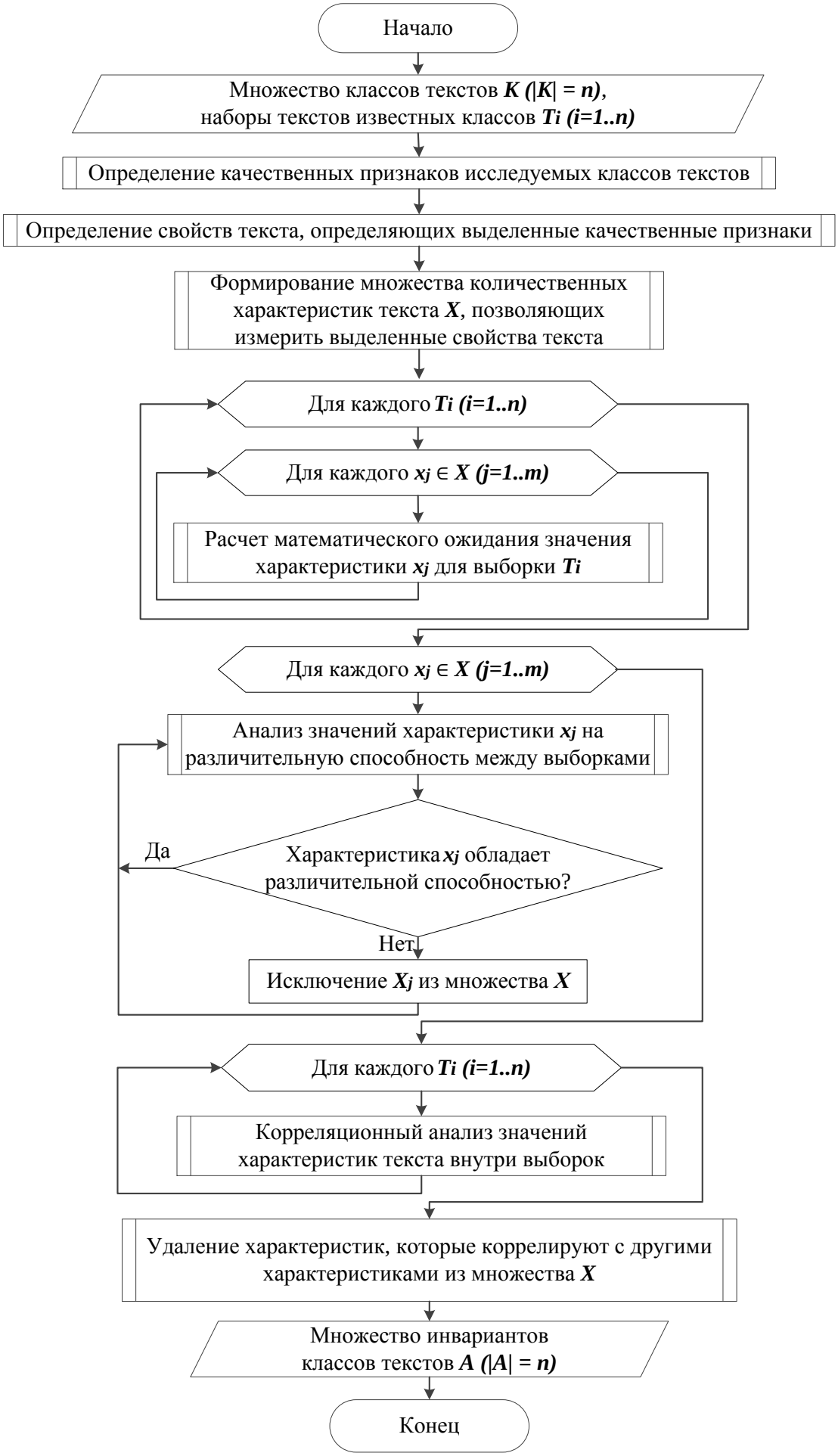

Рисунок 1 – Предложенный автором алгоритм формирования инвариантов

способности и корреляции между собой. На основе проведенных расчетов сформированы инварианты текстов, созданных искусственно с помощью синонимизации и метода цепей Маркова, которые позволяют представить различия данных классов текстов формально – в виде набора значений характеристик и использовать их при определении искусственных текстов.

В табл. 1 представлены разработанные автором новые инварианты для искусственных текстов, созданных с помощью синонимизации и метода Марковских цепей. Данные инварианты были получены с использованием предложенного автором алгоритма.

Таблица 1 – Полученные инварианты для определения искусственных текстов

j	Текстовая характеристика (x_j)	Значения характеристик для инвариантов		
		Естественные тексты (z_{j1})	Искусственные тексты (синонимизация) (z_{j2})	Искусственные тексты (метод цепей Маркова) (z_{j3})
1	Среднее количество знаков пунктуации на 1000 символов	31,742	29,035	21,545
2	Частота 100 популярных биграмм букв на 1000 символов	201,269	112,562	184,360
3	Частота служебных слов на 1000 символов	34,691	25,702	31,865
4	Количество уникальных слов на 1000 символов	64,804	101,659	66,155
5	Среднее число слов в предложении	9,113	9,987	16,549
6	Количество грамматических ошибок на 1000 символов	0,010	6,215	1,002
7	Количество предложений в тексте на 1000 символов	109,812	100,200	60,356
8	Количество сложноподчиненных предложений на 1000 символов	68,655	62,082	31,658
9	Количество вопросительных предложений на 1000 символов	1,414	1,358	0,787
10	Частота 100 популярных слов на 1000 символов	49,001	32,882	37,104
11	Частота 100 популярных 2-грамм слов на 1000 символов	9,100	3,554	4,196
12	Количество слов в семантическом ядре	66,025	95,645	75,268
13	Наличие единства тематики в разных частях текста	2,862	1,135	1,578

Таким образом, с помощью авторского алгоритма формирования инвариантов классов текстов был получен набор характеристик текстов, обладающих различительной способностью в решении задачи определения происхождения текста, а именно определения, написан ли текст человеком или создан автоматически с помощью программного генератора.

В третьей главе предложен метод определения искусственных текстов, отличающийся использованием меры принадлежности входного текста к известным классам текстов и позволяющий принять решение о способе создания текста. Приведены алгоритмическое описание метода, описание программного средства, реализующего авторский метод, а также результаты его экспериментальной апробации.

Метод предназначен для определения происхождения текста, а именно определения, создан входной текст человеком или искусственно.

Область применения метода:

– поисковые системы;

– системы поиска спама, в том числе в составе сложных систем (sms, почтовые сервисы, социальные сети, форумы);

– системы определения плагиата;

– другие системы и сервисы, нуждающиеся в определении происхождения текста.

На вход метода подается текст на русском языке, длиной от 1 200 символов. Выходными данными метода являются заключение о происхождении текста (текст написан человеком или сгенерирован автоматически) и точность, с которой выдано заключение.

Обозначения, используемые в методе:

T – множество текстов, которые могут быть исследованы;

$X = \{x_1, x_2, \ldots, x_m\}$ – множество исследуемых характеристик текста, $|X| = m$;

$A = \{a_1, a_2, \ldots, a_n\}$ – множество инвариантов классов текстов, разделенных по происхождению, то есть множество наборов усредненных значений характеристик текста $x_j \in X$ для n классов текстов.

a_i – инвариант i-го класса текста; $a_i = (a_{i1}, a_{i2}, \ldots, a_{im})$, где a_{ij} – усредненное информативное значение j-й характеристики текста i-го инварианта, $i = 1..n,\ j = 1..m$.

a' – набор рассчитанных значений характеристик текста некоторого входного текста t, происхождение которого требуется определить, $t \in T$;

a'_j – рассчитанное значение j-ой текстовой характеристики входного текста, $j = 1..m$;

$V(a', A)$ – мера оценки принадлежности входного текста к классу текстов с известным происхождением;

$D(a', a_i)$ – мера расстояния между входным текстом и i-м классом текстов известного авторства, представляемая как мера расстояния между векторами a' и a_i;

l – пороговое значение расстояния между вектором значений текстовых характеристик входного текста a' и вектора-инварианта i-го класса текста с известным происхождением a_i такое, что значение максимальное значение меры $D(a', a_i)$ не должно превышать l при $i = 1..n$;

R_A – точность решения о способе создания текста, которое принимается с помощью метода.

Метод определения происхождения текста представляет собой последовательность действий, связанных с анализом характеристик текста:

Шаг 1. Рассчитать числовые значения текстовых характеристик $x_i \in X$ входного текста, $i = 1..n$.

Шаг 2. Сформировать вектор a' как набор рассчитанных значений характеристик. $a' = (a'_1, a'_2, ..., a'_m)$.

Шаг 3. Рассчитать меры расстояний $D(a', a_i)$ между векторами a' и $a_i \in A$, $i = 1..n$.

Шаг 4. Рассчитать меру принадлежности входного текста к известным классам. Мера принадлежности определяется как выбор наименьшей из мер расстояний, рассчитанных на шаге 3:

$$V(a', A) = \min[D(a', a_i)];\ i = 1...n.$$

Шаг 5. Сравнить меру принадлежности, полученную на шаге 4, с заданным пороговым значением l.

Шаг 6. Принять решение о происхождении текста.

Принимается решение о том, что входной текст отнесен к i-му классу текстов, если выполняются следующие соотношения:

$$V(a', A) \equiv D(a', a_i);$$
$$V(a', A) \le l.$$

Принимается решение о том, что входной текст не может быть отнесен ни к одному из классов текстов известного происхождения, если выполняется соотношение:

$$V(a', A) > l.$$

Шаг 7. Рассчитать точность заключения о происхождении текста R_A по формуле:

$$R_A = 1 - \frac{V(a', A)}{D(a_x, a_y)},$$

где $a_x, a_y \in A$ – два вектора-инварианта, наиболее приближенных (согласно мере расстояния) к a'.

Пояснения к методу:

1. Для расчета меры расстояния между векторами может быть использована любая метрика, позволяющая количественно оценить расстояние между двумя точками в k-мерном пространстве. Среди возможных метрик – общеизвестные евклидова метрика и метрика Махаланобиса. Нужно отметить, что выбор метрики должен отвечать требованиям поставленной задачи.

Так как для решения задачи определения происхождения текста необходим единый масштаб получаемых значений меры $D(a', a_i)$, для ее расчета автором предлагается использовать расстояние Махаланобиса. Данная метрика обобщает понятие расстояния Евклида, учитывает корреляции между переменными и инвариантно к масштабу. Она широко используется в кластерном анализе и методах классификации.

2. Пороговое значение l ограничено сверху величиной, равной половине меры расстояния между двумя наиболее приближенными векторами-инвариантами $a_x, a_y \in A$ и может быть скорректировано в меньшую сторону на основе экспериментальных расчетов:

$$l \leq \frac{D(a_x; a_y)}{2}.$$

3. Значение величины точности заключения о происхождении текста (соответствии текста некоторому инварианту из множества A) R_A лежит в интервале [0;1]. Ее значение тем выше, чем меньше расстояние между вектором значения характеристик входного текста с вектором-инвариантом, с которым он был соотнесен.

Программное средство определения искусственно созданных текстов

На основе предложенного метода было разработано специализированное программное средство «TextOrigin», предназначенное для автоматизации процесса обнаружения и последующей фильтрации искусственно сгенерированных текстов.

Для разработки программного средства была выбрана веб-ориентированная технология, позволяющая создавать кроссплатформенные решения, а также использовать для написания сервиса и клиентской части разные языки программирования. Кроме того, базовые возможности программного средства были реализованы в виде модулей для популярных систем управления контентом. Функционал программного средства позволяет интегрировать его в действующие веб-ресурсы для проведения анализа входных данных по критерию искусственности их происхождения. Практическая ценность разработанного средства заключается в возможности использования его в качестве фильтра для модулей загрузки «авторских» новостей в СМИ, приема заявок или обращений онлайн, загрузки постов в сообществах социальных сетей.

Для реализации программного средства использовались сформированные и приведенные во второй главе инварианты естественных и искусственно созданных текстов. Кроссплатформенность и кроссбраузерность обеспечиваются использованием современных веб-технологий. Структура программного средства определения искусственно созданных текстов приведена на рис. 2, где выделены его основные модули.

Программное средство может обрабатывать входные данные, представленные множеством широко используемых текстовых форматов. Отличительной особенностью системы является возможность обработки документов, использующих языки разметки, такие как html-тэги и xml-сущности. Структуризация входных данных осуществляется модулем нормализации текста.

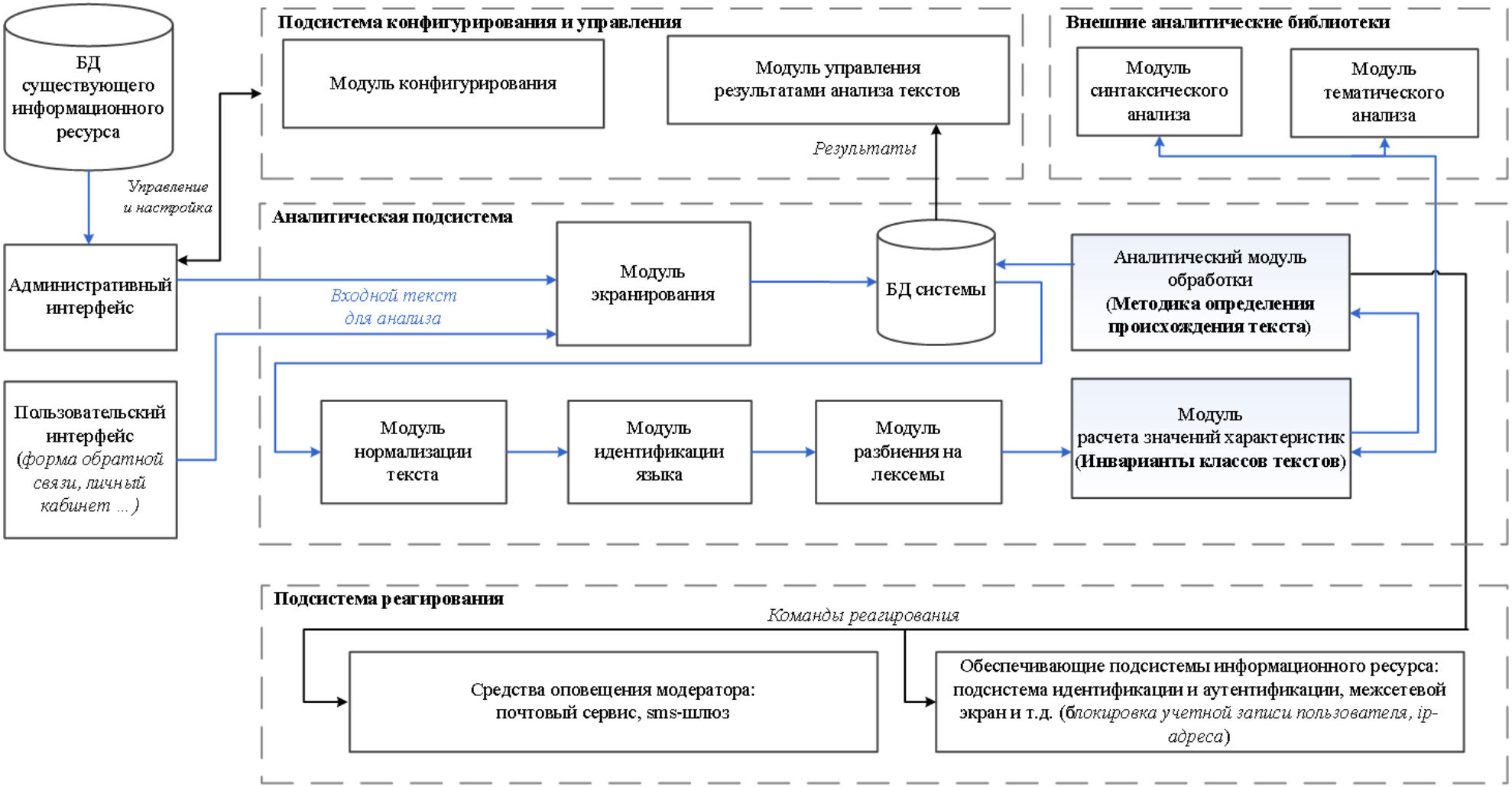

Рисунок 2 – Структура программного средства определения искусственно созданных текстов

Тестирование разработанного метода и программного средства

Для оценки эффективности предложенного метода определения искусственно созданных текстов были проведены экспериментальные расчеты меры принадлежности некоторых входных текстов известного происхождения к исследуемым классам текстов. В экспериментах были использованы тексты различного происхождения, собранные автором работы в популярных социальных сетях, а также предоставленные компаниями-партнерами для апробации. Объем каждой статьи составлял от 1 200 до 5 200 символов. Общий объем исходной выборки естественных текстов составил 1 000 текстов или 1,965 млн символов. Статьи отбирались по следующим тематикам: общество, политика, финансы, власть, армия, силовые структуры, наука и техника и смежные с ними.

Целью проведения серии экспериментов является получение объективных сведений об эффективности предложенного автором метода определения искусственно созданных текстов.

Естественные тексты подверглись действию автоматических генераторов на основе синонимизации и метода Марковских цепей для того, чтобы создать искусственные экземпляры. Для дополнительной оценки влияния объема словаря синонимом на результат определения происхождения текста были использованы 2 словаря: с 700 тыс. синонимов и с 130 тыс. синонимов. Среднее значение уникальности искусственных текстов, созданных с помощью словаря с меньшим объемом, рассчитанное с помощью алгоритма шинглов, составило 36,5%, тогда как для словаря с бóльшим объемом это значение составило 69,1%.

Таким образом, в эксперименте использовано 4 выборки по 1 000 текстов, объединенных одним из способов создания:

– естественные тексты, написанные человеком или несколькими людьми;

– искусственные тексты, созданные с помощью метода Марковских цепей;

– искусственные тексты, созданные с помощью синонимизации с помощью словаря из 700 тысяч синонимов;

– искусственные тексты, созданные с помощью синонимизации с помощью словаря из 130 тысяч синонимов.

В табл. 2 приведены показатели ошибок 1-го и 2-го рода на основе полученных результатов проведенных вычислений для полных выборок.

Под ошибками 1-го рода в данном случае понимаются случаи, когда естественный текст был принят за искусственный (ложноположительное событие, или «ложная тревога» для пользователя системы). Ошибки 2-го рода указывают на случаи, когда искусственный текст не был распознан системой и был принят за естественный экземпляр (ложноотрицательное событие, или «пропуск события»).

Средний показатель ошибок 1-го рода по определению искусственных текстов в целом составил 4,27 %, ошибок 2 рода – 2,40 %.

Таблица 2 – Показатели ошибок 1-го и 2-го рода метода определения искусственно созданных текстов

Показатель	Ошибки 1-го рода	Ошибки 2-го рода
Определение текста, созданного с помощью синонимизации (любительский словарь – среднее значение уникальности 36 %)	3,2 %	6,1%
Определение текста, созданного с помощью синонимизации (дополненный словарь – среднее значение уникальности 69 %)		2,8%
Определение текста, созданного с помощью алгоритма на основе цепей Маркова		2,6%

Показатель ошибок 1-го рода по определению искусственных текстов в целом (для текстов уникальностью выше 50 %) составил не более 4 %, ошибок 2-го рода – не более 3 %. Высокие результаты определения искусственно созданных текстов показывают эффективность применения разработанного автором метода, а также инвариантов искусственных текстов.

По показателю ошибок 1-го рода разработанный метод эффективнее методики А.С. Павлова обнаружения искусственных текстов, представляющих собой поисковый спам, которая имеет показатель ошибок свыше 5 %. По показателям ошибок 2-го рода разработанный метод имеет меньшую эффективность. Указанная методика имеет показатель менее 2 %, однако достигнутый результат считается умеренным для систем текстовой атрибуции. Также следует отметить, что метод показывает бóльшую эффективность при определении искусственных текстов, которые обладают уникальностью выше 50%.

В заключении приведены основные результаты и выводы по проделанной работе.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В работе приведены результаты комплексного исследования по решению задачи определения искусственно созданных текстов. Данная задача является междисциплинарной и охватывает научные знания прикладной лингвистики, текстовой атрибуции, стилометрии, статистического анализа.

По итогам проведенного диссертационного исследования была достигнута цель работы: повышена точность определения искусственно созданных текстов за счет того, что впервые были исследованы тексты, представляющие собой информационный контент. Отличием данных текстов является их предназначение – быть прочитанными пользователями, в отличие от поискового спама, который в первую очередь направлен на взаимодействие с поисковой машиной и зачастую не предназначен для прочтения пользователем. Значимость решения данной

задачи подчеркивается ростом популярности интернет-ресурсов, которые все чаще становятся платформой для высказывания собственного мнения, агитации, рекламы, а также широко используются в качестве средств массовой информации.

В ходе решения поставленных задач получены следующие основные результаты:

1. Проведенный анализ современных подходов, используемых в задачах классификации текстов, показал, что существующие решения имеют ряд недостатков и не позволяют эффективно определять искусственно созданные тексты, представляющие собой информационный интернет-контент.

2. Получены инварианты, позволяющие определять искусственные тексты, созданные с помощью синонимизации и метода Марковских цепей.

3. Разработан метод определения искусственно созданных текстов, включающий в себя алгоритм расчета меры принадлежности входного текста к базовым классам.

4. Успешная экспериментальная апробация разработанной методики подтвердила возможность ее применения в задаче определения происхождения текстов: ошибки 1-го рода составили не более 4 %, ошибки 2-го рода – не более 3 %.

5. Разработано программное средство фильтрации искусственно созданных текстов, позволяющее в автоматическом режиме определять контент, созданный с помощью программных генераторов.

Результаты экспериментов показали, что разработанные автором метод и программное средство определения искусственно созданных текстов в 93% случаев позволили успешно определить искусственные тексты, сгенерированные на основе синонимизации и Марковских цепей.

Использование разработанного программного средства в ООО «Агентство медиарешений» и ООО «Лингвистические и информационные технологии» позволило снизить трудозатраты модератора контента за счет использования автоматической проверки текстов, а также повысить качество проводимого автоматизированного анализа текстового материала на русском языке, что подтверждается актами внедрения. Результаты диссертационной работы также были внедрены в учебную деятельность Томского государственного университета систем управления и радиоэлектроники.

Основные результаты диссертационного исследования опубликованы в следующих работах:

Статьи, опубликованные в журналах, включенных ВАК в перечень ведущих рецензируемых научных журналов и изданий:

1. Шумская (Исхакова) А.О. Идентифицирующие признаки текстовых сообщений при установлении автора / А.О. Шумская // Ползуновский вестник. – 2013. – № 2. – С. 265–266.

2. Шумская (Исхакова) А.О. Выбор параметров для идентификации искусственно созданных текстов / А.О. Шумская // Доклады ТУСУРа. – 2013. – № 2 (28). – С. 126-128.

3. Шумская (Исхакова) А.О. Оценка эффективности метрик расстояния Евклида и расстояния Махаланобиса в задачах идентификации происхождения текста / А.О. Шумская // Доклады ТУСУРа. – 2013. – № 3 (29). – С. 141–145.

4. Исхакова А.О. Модель процесса формирования инвариантов классов текстов / А.О. Исхакова // Доклады ТУСУРа. – 2016. – № 3. – С. 65–69.

5. Исхакова А.О. Метод определения искусственных текстов на основе расчета меры принадлежности к инвариантам / А.О. Исхакова // Труды СПИИРАН. – 2016. – № 6 (49). – С. 124–140.

Публикации в других научных изданиях:

6. Шумская (Исхакова) А.О. Анализ текстовых признаков искусственных текстов, созданных на основе синонимизации / А.О Шумская // Научная сессия ТУСУР – 2013 : материалы Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых, Томск, 16–18 мая 2013 г. : в 5 ч. – Томск : В-Спектр, 2012. – Ч. 4. – С. 226–228.

7. Шумская (Исхакова) А.О. Задачи идентификации искусственных текстов / А.О. Шумская // Научная сессия ТУСУР – 2013 : материалы Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых, Томск, 16–18 мая 2013 г. : в 5 ч. – Томск : В-Спектр, 2012. – Ч. 4. – С. 224–226.

8. Shumskaya (Iskhakova) A.O. / The effectiveness of the Euclidean and Mahalanobis distances while solving the problem of the text origin identification / A.O. Shumskaya // Modern informatization problems : Proceedings of the XIX International Open Science Conference (Yelm, WA, Usa, January 2014) / ed. O.Ja. Kravets. – Yelm, WA : Science Book Publishing House, 2014. – P. 73–78.

9. Шумская (Исхакова) А.О. Об идентификации искусственно созданных текстов // А.О. Шумская, Р.В. Мещеряков / Шестая международная конференция по когнитивной науке : тезисы докладов. – Калининград, 2014. – С. 648–649.

10. Shumskaya (Iskhakova) A.O. Using Euclidean and Mahalanobis distances while solving the problem of the text origin identification, scientific paper / A.O. Shumskaya // Interactive systems: Problems of Human-Computer Interaction : Collection of scientific papers. – Ulyanovsk : USTU, 2015. – P. 211–217.

11. Шумская (Исхакова) А.О. Определение искусственных текстов на основе поиска часто употребляемых слов и устойчивых словосочетаний / А.О. Шумская // Седьмая международная конференция по когнитивной науке : тезисы докладов. – Светлогорск, 2016. – С. 647–648.

Свидетельство о регистрации программ для ЭВМ:

12. Auth_stat. Программное обеспечение для расчета статистических значений текстовых характеристик : свидетельство о гос. регистрации программы для ЭВМ № 2015663136 Российская Федерация / Шумская (Исхакова) А.О., Мещеряков Р.В; правообладатель ТУСУР. Зарегистрировано в Реестре программ для ЭВМ 11.12.2015.