На правах рукописи

ГЛАЗКОВА Анна Валерьевна

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ КЛАССИФИКАЦИИ ОБЪЕКТОВ (НА ПРИМЕРЕ ОПРЕДЕЛЕНИЯ КАТЕГОРИИ ПОТЕНЦИАЛЬНЫХ АДРЕСАТОВ ТЕКСТА)

Специальность 05.13.18 — Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата технических наук

Тюмень–2016

Работа выполнена на кафедре программного обеспечения Федерального государственного бюджетного образовательного учреждения высшего образования «Тюменский государственный университет».

Научный руководитель: кандидат физико-математических наук, профессор **Захарова Ирина Гелиевна**

Официальные оппоненты: **Барахнин Владимир Борисович**, доктор технических наук, доцент, Федеральное государственное бюджетное учреждение науки Институт вычислительных технологий Сибирского отделения Российской академии наук (ИВТ СО РАН), ведущий научный сотрудник

Курушин Даниил Сергеевич, кандидат технических наук, Федеральное государственное бюджетное образовательное учреждение высшего образования «Пермский национальный исследовательский политехнический университет», доцент кафедры информационных технологий и автоматизированных систем

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Томский государственный университет систем управления и радиоэлектроники»

Защита диссертации состоится «20» октября 2016 года в 16-00 на заседании диссертационного совета Д 212.274.14 при ФБГОУ ВО «Тюменский государственный университет» по адресу: 625003, г. Тюмень, ул. Перекопская 15а, ауд. 410.

С диссертацией можно ознакомиться в библиотеке ФБГОУ ВО «Тюменский государственный университет» и на сайте diss.utmn.ru.

Автореферат разослан «___» _____2016 года.

Ученый секретарь диссертационного совета *Е. А. Оленников*

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. В условиях развития информационных ресурсов одним из ключевых направлений современной компьютерной науки является разработка методов систематизации и поиска информации. Процесс решения данных прикладных задач подразумевает, как правило, усовершенствование механизмов классификации текстов на естественном языке.

Вопросы классификации текстов рассматривались, в частности, Е.Д. Агафоновым, В.Б. Барахниным, Т.В. Батурой, К.В. Воронцовым, В.В. Гулиным, А.С. Епревым, Р.В. Мещеряковым, В.В. Поддубным, А.А. Роговым, А.С. Романовым, В.О. Толчеевым, Д.В. Хмелевым, О.Г. Шевелевым, S. Argamon, W. Cohen, T. Joachims, D. Nguyen, K. Santosh.

Одним из актуальных вопросов классификации документов является решение *задачи установления характеристик адресата текста*. Данная задача затрагивается преимущественно зарубежными исследователями. Так, работы R. Akker и D. Traum, D. Choi, H. Lee посвящены анализу признаков, характеризующих текст с точки зрения его ориентации на различные категории читателей. Использование данных признаков для текстов, написанных на русском языке, не представляется корректным в силу индивидуальных особенностей синтаксических структур каждого языка. Таким образом, для русскоязычных текстов в настоящее время не существует единого набора классификационных признаков, которые могли бы быть положены в основу определения возрастной аудитории текста.

В рамках данного исследования рассматривается *задача классификации текстов на примере их отнесения к той или иной возрастной категории адресатов.* Актуальность решения задач, связанных с идентификацией адресата текста, обоснована введением возрастных ограничений на контент интернет-ресурсов, развитием систем электронного обучения, а также малой освещенностью обозначенной проблемы в работах российских ученых. Возможность классифицировать тексты на основании групп адресатов способствует, в первую очередь, улучшению релевантности результатов информационного поиска. Также решение данной задачи позволяет усовершенствовать механизмы исключения из найденной выборки нежелательных ресурсов (например, сайтов, содержание которых рассчитано на пользователя иной категории).

Рассматриваемая задача относится к числу слабоформализуемых за счет сложности естественного языка и многообразия его коммуникативных форм, поиск путей ее решения требует построения адекватных математических моделей классификации.

Целью исследования является разработка математических методов моделирования отношений «текст-адресат» и алгоритмов классификации для

определения категории потенциальных адресатов текста, а также создание программного комплекса, реализующего данные методы и алгоритмы.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Проанализировать существующие методы и алгоритмы классификации текстов.
2. Разработать подход к математическому моделированию в задачах классификации текстов.
3. Разработать на основе полученного подхода методы и алгоритмы классификации.
4. Спроектировать и разработать программный комплекс, реализующий предложенные методы и алгоритмы.
5. Провести вычислительные эксперименты для тестирования разработанных методов и алгоритмов.

Объектом исследования являются математические методы моделирования задач классификации.

Предметом исследования являются методы и алгоритмы определения категории потенциальных адресатов текста на примере отнесения текстов к определенной возрастной аудитории, а также программная реализация предложенных методов и алгоритмов в рамках разработки интеллектуальной системы.

Методология и методы исследования. При проведении исследования применялись методы следующих областей знаний: математическое моделирование, теория множеств, математическая статистика, структурное проектирование информационных систем, объектно-ориентированное программирование, искусственный интеллект (искусственные нейронные сети).

На защиту выносятся следующие результаты, соответствующие четырем пунктам паспорта специальности 05.13.18 — Математическое моделирование, численные методы и комплексы программ:

Пункт 1. Разработка новых математических методов моделирования объектов и явлений.

1. Подход к математическому моделированию классификации объектов (на примере отнесения текстов к той или иной возрастной категории адресатов), признакового пространства и зависимости классификационных признаков.

Пункт 3. Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий.

2. Численный метод классификации текстов, разработанный на основе разбиения множества текстов на классы эквивалентности.

Пункт 4. Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента.

3. Программный комплекс для автоматической классификации текстов. Проведен вычислительный эксперимент, использующий тексты, входящие в Национальный корпус русского языка. Результаты компьютерных экспериментов показали адекватность разработанных в диссертации математических моделей и методов. Получено свидетельство о регистрации программы для ЭВМ №2015616462.

Пункт 5. Комплексные исследования научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента.

4. Результаты моделирования процесса классификации текстов и описание на его основе зависимости классификационных признаков.

Таким образом, в соответствии с формулой специальности 05.13.18 в диссертации представлены оригинальные результаты одновременно из трех областей: математического моделирования, численных методов и комплексов программ.

Научная новизна исследования заключается в следующем:

1. Математическое моделирование

Разработан и обоснован подход к моделированию классификации текстов (на примере их отнесения к той или иной возрастной категории адресатов), развивающий существующие математические модели классификации за счет возможности учесть в процессе формализации особенности поставленной в работе задачи:

a) вложенность категорий;
b) пересечение категорий.

Показана возможность формализации постановки задачи классификации в общем виде.

2. Численные методы

Впервые для решения задачи отнесения текстов к той или иной возрастной категории адресатов предложен и обоснован численный метод классификации текстов, построенный на разбиении множества текстов на классы эквивалентности и позволяющий определить меру близости текстов как расстояние между векторами значений характеризующих их классификационных признаков.

3. Комплексы программ

Для тестирования предложенных моделей и методов создан программный комплекс — интеллектуальная система автоматической классификации текстов. Особенности модульной архитектуры программного комплекса позволяют проводить его гибкую интеграцию в системы работы с электронными документами. Работа модуля классификации текстов выполняется поэтапно. Эксперименты показали, что данный подход к реализации позволяет снизить временные затраты на обучение и работу модуля.

Теоретическая значимость работы заключается в следующем:

1. Предложенный подход, позволяющий формализовать постановку и этапы решения задачи классификации текстов на примере определения их предполагаемой возрастной аудитории, дает возможность получить формальное представление задач отнесения объектов к одной или нескольким пересекающимся или непересекающимся категориям и тем самым развивает теоретические основы формализации задачи классификации.
2. Разработанный и реализованный численный метод классификации текстов расширяет возможности применения численных методов для решения слабоформализуемых задач и позволяет определить расстояние между текстами на основании их представления в виде наборов значений признаков и соответствующих им весовых коэффициентов.
3. Предлагаемые в работе математические модели и методы имеют универсальный характер и могут применяться для классификации других видов объектов, модели которых могут быть описаны сходными классификационными признаками.

Практическая значимость работы. В целях тестирования разработанных методов и алгоритмов был реализован программный комплекс для автоматической классификации текстов. Программный комплекс оперирует знаниями в рассматриваемой области с целью отнесения текста к той или иной категории.

Разработанный программный комплекс может найти практическое применение в поисковых системах (для отбора релевантного контента), системах обучения, электронных библиотеках и каталогах, системах автоматического реферирования и рецензирования.

Достоверность изложенных в работе результатов подтверждается научно-теоретическим обоснованием избранного исследовательского направления; достаточным объемом обучающей и контрольной выборок для проведения вычислительного эксперимента; сравнением результатов вычислительного эксперимента с данными, полученными на основании мнений экспертов; всесторонним анализом полученных результатов и их широким обсуждением.

Внедрение результатов. Результаты диссертационного исследования получили практическое применение в некоммерческом партнерстве по содействию развитию науки и образования «Национальный корпус русского языка» и в негосударственном образовательном учреждении «Югорский учебный центр».

Апробация результатов. Основные результаты исследования докладывались на следующих конференциях и семинарах:

1. IEEE-семинар «Интеллектуальные системы моделирования, проектирования и управления» (г. Томск, Томский государственный университет систем управления и радиоэлектроники, 2016 г.).
2. 53 Международная научная студенческая конференция «МНСК-2015» (г. Новосибирск, Новосибирский государственный университет, 2015 г.).
3. V Международная научно-техническая конференция «Open Semantic Technologies for Intelligent Systems — OSTIS-2015» (Республика Беларусь, Минск, Белорусский государственный университет информатики и радиоэлектроники, 2015 г.).
4. VIII Международная научно-практическая конференция «Научное творчество XXI века» (г. Красноярск, Научно-исследовательский центр, 2014 г.).
5. III Международная научно-техническая конференция «Artificial Intelligence and Natural Language — AINL-2014» (г. Москва, Инновационный центр «Сколково», 2014 г.).
6. III Всеукраинская научно-практическая конференция «Интеллектуальные системы и прикладная лингвистика» (Украина, г. Харьков, Национальный технический университет «Харьковский политехнический институт», 2014 г.).
7. IV Международная научно-техническая конференция «Open Semantic Technologies for Intelligent Systems — OSTIS-2014» (Республика Беларусь, Минск, Белорусский государственный университет информатики и радиоэлектроники, 2014 г.).
8. XI Всероссийская конференция «Преподавание информационных технологий в Российской Федерации» (г. Воронеж, Воронежский государственный университет, 2013 г.).
9. VI Научно-практическая межрегиональная конференция «Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений» (г. Тюмень, Тюменский государственный университет, 2013 г.).

Публикации. Основные результаты диссертации опубликованы в 12 научных работах, в том числе в 3 статьях в рецензируемых научных изданиях, рекомендованных ВАК для представления основных научных результатов диссертаций на соискание ученой степени доктора или кандидата наук. Также получено свидетельство о государственной регистрации программы для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, трех глав, заключения, списка литературы и трех приложений. Общий объем работы составляет 141 страницу и включает в себя 26 рисунков и 44 таблицы. Список литературы содержит 154 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы исследования, ставятся цель и задачи исследования, приведены основные положения, выносимые на защиту, а также сведения о теоретической и практической значимости, научной новизне и апробации работы.

Первая глава посвящена описанию состояния и развития технологий автоматической обработки и классификации текстов. В главе также приведены постановка задачи классификации, этапы ее решения и наиболее часто применяемые методы классификации текстов.

Истоки исследований в области обработки естественного языка восходят к работам Н. Хомского по формализации структуры языка. Одновременно с этим значительное влияние на развитие данных технологий оказали различные работы в области искусственного интеллекта.

Задача классификации текстов подразделяется на две подзадачи: обучение классификатора и непосредственно выполнение классификации. Наибольшую трудность составляет первая подзадача, от успешности ее выполнения зависит достоверность проведенной классификации. В первую очередь проводится приведение документов к единому формату — построение моделей. Классифицируемые объекты необходимо представить в виде наборов признаков. Проблеме выделения признаков в задаче классификации посвящены работы Е.Д. Агафонова, С.И. Колесниковой, А.Е. Янковской, A. McCallum, Y. Yang и др. Вопросы моделирования структуры текстов различной специфики рассматривались А.Г. Варфоломеевым, Д.В. Кузнецовым, Д.С. Курушиным, Д.В. Ландэ, Н.Д. Москиным, Ю.А. Орловой и др.

На основе признаков, полученных для решения задачи, проводится обучение классификатора (с учителем, без учителя или смешанное обучение). Самым качественным считается обучение с учителем. Оно осуществимо только в том случае, когда возможно заранее получить выборку объектов со знанием их классов. Обучение с учителем использует байесовские или линейные методы, методы, построенные на применении деревьев принятия решений или нейронных сетей. Разработке, применению и сравнению различных типов классификаторов для задачи классификации текстов посвящены работы Т.В. Батуры, В.В. Гулина, Л.М. Ермаковой, А.В. Заболеевой-Зотовой, О.А. Невзоровой, В.О. Толчеева, А.Ф. Тузовского, W. Cohen, T. Joachims и др. Подходы к математическому моделированию задачи классификации предложены в работах В.Б. Барахнина и А.М. Федотова, А.С. Епрева, H. Nezreg и др. И.А. Ходашинским проведены исследования нечетких классификаторов; применение теории нечетких множеств к задаче классификации текстов описывается в работах О.В. Золотухина, С.В. Шпирко и др.

Исходя из специфики поставленной задачи, особый интерес для исследования представляли работы, посвященные извлечению из текста данных о его авторе или адресате. В ряде статей неоднократно рассматривались вопросы определения характеристик автора текста — его возраста, пола, типа личности и национальной принадлежности (K. Santosh и др., D. Nguyen и др., Е.А. Гречников и др., Р.В. Мещеряков, З.И. Резанова, А.С. Романов, О.Г. Шевелев и др.). D. Choi предложил подход к применению методов распознавания адресанта текста для поиска записей террористической тематики в Интернете. R. Akker и D. Traum, H. Lee рассматривали задачу создания диалоговых систем, в контексте которой анализировали признаки, характеризующие текст с точки зрения его ориентации на различных адресатов. М.Ф. Ашуровым и В.В. Поддубным проведена классификация текстов по их автору с использованием потоковых методов классификации. Подход к классификации поисковых запросов на основании оценки близости терминов предложен J. Attenberg и T. Suel.

Анализ показал, что большинство методов классификации текстов являются узкоспециальными. Это обусловлено тем, что выбор признаков, положенных в основу классификации, зависит от предметности документа и его особенностей, а следовательно, многие задачи классификации нуждаются в рассмотрении и исследовании.

Также необходимо отметить, что большинство существующих методов проводят классификацию только на основе лексических единиц текста. Методы, рассматривающие особенности синтаксической структуры предложений, как правило, учитывают специфику синтаксиса текста в виде подсчета количества лексем, относящихся к той или иной синтаксической категории. Таким образом, важной задачей автоматической обработки текстов является интеграция имеющихся математических и статистических методов с методами анализа документов, которые в полной мере учитывали бы их структурные и синтаксические особенности.

Кроме того, на основе проведенного анализа методов и алгоритмов классификации текстов сделан вывод о возможности и перспективности применения нейросетевого подхода к решению данного рода задач. При этом отмечено, что обученная сеть представляет собой «черный ящик», а следовательно, логика получения результатов ее работы не понятна пользователю. Интеграция детерминированных и нейросетевых методов классификации обеспечивает большую прозрачность процесса классификации, что не только оказывает положительное влияние на точность соотнесения объектов категориям, но и позволяет пользователю проанализировать степень зависимости результатов классификации от значений различных классификационных признаков.

Вторая глава посвящена математическому моделированию процесса классификации текстов на примере их отнесения к определенной возрастной

категории, что в рамках данного исследования подразумевало решение следующих подзадач:

1. Построение подхода к формализации задачи классификации текстов и моделированию процесса классификации.
2. Формализованное описание признакового пространства и зависимости классификационных признаков.
3. Разработка подхода к классификации текстов.

Под *категорией* в данной работе понимается класс текстов, выделенный на основе значений классификационных признаков входящих в него объектов (текстов). Категория K_i в таком случае может быть определена следующим образом:

$$K_i=\{q_j^C, V_j^i, w_j\},\ 1\leqslant j\leqslant L, \qquad (1)$$

где q_j^C — идентификатор классификационного признака; V_j^i — критическое значение j-го признака из категории K_i; $w_j\in[0,1]$, $\sum_{j=1}^{L} w_j=1$ — весовой коэффициент классификационного признака; L — общее число классификационных признаков.

Критическое значение признака представляет собой значение, которое признак принимает для данной категории K_i, оно может быть задано в общем случае в виде интервальной оценки:

$$V_j^i \in (l_j^i,\ r_j^i),\ 1\leqslant j\leqslant L, \qquad (2)$$

где l_j^i, r_j^i — границы критического интервала.

Весовой коэффициент классификационного признака характеризует значимость признака q_j^C в сравнении с другими признаками, которая может быть определена на основании экспертных оценок или в ходе обучения системы.

Для каждого признака q_j^C имеем f_j — отображение множества текстов $\mathfrak{I}$ во множество допустимых значений признака Q_f (Q_f относится к определенному признаку):

$$f_j: \mathfrak{I} \rightarrow Q_f. \qquad (3)$$

Таким образом, *категория определяется набором критических значений и весовых коэффициентов*, которые соответствуют классификационным признакам, а *текст характеризуется своим признаковым описанием* — набором классификационных признаков q_j^T и их значений a_j для данного текста $T=T_i$ из множества $\mathfrak{I}=\{T_1, T_2, \ldots, T_n\},\ 1\leqslant i\leqslant n$:

$$T=\{q_j^T,\ a_j\},\ 1\leqslant j\leqslant L. \qquad (4)$$

Отображение текста T в его признаковое описание допустимо записать в виде:

$$\varphi: T \rightarrow F_T. \quad (5)$$

При этом признаковое описание, которое в контексте данной задачи возможно отождествлять с самим текстом, может быть представлено в виде вектора F_T:

$$F_T=(f_1(T),f_2(T),...,f_L(T)). \quad (6)$$

Обозначим через $\mathbb{K}$ множество категорий, по которым проводится классификация. Тогда существует отображение:

$$\gamma : \mathfrak{I} \rightarrow \mathbb{K}, \quad (7)$$

ставящее в соответствие любому тексту T из множества $\mathfrak{I}=\{T_1,T_2,...,T_n\}$ категорию, к которой относится данный текст.

Возможно существование нескольких подходов к формализации задачи классификации текстов.

В общем виде задача классификации текстов состоит в следующем. Имеется текст T и множество категорий $\mathbb{K}=\{K_1,K_2,...,K_n\}$, с которыми данный текст должен быть сопоставлен. Задача сводится к тому, чтобы выбрать категорию, к которой относится текст T:

$$T \sim K_i,\ K_i \in \mathbb{K},\ 1 \leqslant i \leqslant n. \quad (8)$$

В данной работе знак (~) означает принадлежность текста категории или множеству категорий.

Рассматриваемый подход (8) позволяет в том числе однозначно отнести текст к одной из существующих категорий. В частности, данный подход уместен при проведении возрастной классификации текстовых ресурсов. Подобная формальная постановка задачи позволяет выбрать из множества категорий информационной продукции ту, с которой может быть соотнесен классифицируемый текст.

При рассмотрении задачи классификации в общем виде (8) считалось, что категории $K_1,K_2,...,K_n$ являются независимыми. Следовательно, отнесение текста к категории K_i означало, что он не может быть причислен к прочим категориям из множества $\mathbb{K}$. В то же время данное представление не всегда соответствует целям проводимой классификации.

Принимая во внимание некоторые особенности предметной области, в контексте данной задачи имеет смысл говорить о вложенности категорий. Очевидно, что текст, адресованный некоторой возрастной аудитории, может предназначаться и другим возрастным группам. Так, принадлежность текста некой категории подразумевает также то, что он будет понятен читателям старших возрастов.

Учитывая эту особенность, отношения между категориями можно представить в виде $K_1 \subset K_2 \subset ... \subset K_n$, тогда:

$$T \sim K_i \Rightarrow T \sim K_j,\ i \leqslant j \leqslant n. \qquad (9)$$

Обозначенный подход к моделированию предметной области позволяет принять во внимание то, что текст из категории K_i принадлежит также $K_{i+1}, K_{i+2}, \ldots, K_n$.

В качестве особенности предложенного пути формализации (9) следует отметить, что речь в предыдущем примере идет преимущественно не об адресованности текста определенной аудитории, а о его понятности представителям той или иной возрастной группы. В нашем же примере особый интерес вызывает то, что содержание и структура текста, адресованного читателям самого младшего возраста, хотя и будут понятны другим категориям реципиентов, могут не соответствовать уровню коммуникативного развития адресатов, относящихся к другим категориям.

Таким образом, в процессе формализации данной задачи имеет смысл предусмотреть возможность причислить текст к ряду пересекающихся категорий, но при этом учесть, что эти категории не всегда будут вложенными друг в друга. Тогда на основании различных наборов классификационных признаков и в зависимости от цели классификации появится возможность отнести текст к различному ряду категорий.

Учитывая вышесказанное, задача классификации может быть сформулирована следующим образом. Пусть дан текст T и множество категорий $\mathcal{K}=\{K_1, K_2, \ldots, K_n\}$. Необходимо найти подмножество $\mathcal{K}_T$, состоящее из категорий, которым может принадлежать текст:

$$T \sim \mathcal{K}_T,\ \mathcal{K}_T=\{K_i : T \sim K_i\},\ 1 \leqslant i \leqslant n,\ i=j_1, j_2, \ldots, j_m. \qquad (10)$$

Подход к формальной постановке задачи классификации текстов с учетом вложенности категорий (9) в контексте решаемой задачи может быть представлен в виде (8). В случае, когда принадлежность текста некой категории подразумевает также то, что он будет понятен читателям старших возрастов, требуется найти только ту возрастную категорию, которой адресован текст (минимальную возрастную категорию).

Подход (10), учитывающий пересечение категорий текстов, также может быть представлен в виде (8). В случае, когда множество категорий содержит пересекающиеся категории (существуют тексты, которые относятся более чем к одной категории), данные пересекающиеся категории могут быть разделены на большее число описанных по отдельности непересекающихся категорий.

Таким образом, на вход системы классификации подается F_T — вектор значений признаков, характеризующих текст (признаковое описание текста). Выходом является идентификатор категории K_i — описание класса из набора $\mathcal{K}=\{K_1, K_2, \ldots, K_n\}$.

Отличие данного подхода от предложенных ранее состоит в возможности учесть ряд особенностей поставленной в работе задачи (вложенность категорий, пересечение категорий). Предложены три пути формализации задачи классификации. Показана возможность приведения всех путей формализации к общему виду.

Для разных классификационных признаков в соответствии с возможным диапазоном их значений (в зависимости от множества Q_f) предусмотрена возможность использования различных шкал. Таким образом, выделены следующие типы признаков:

1. Бинарные: {0,1} (например, наличие/отсутствие специальной лексики в тексте).
2. Номинальные: конечное множество значений (литературная форма — рассказ, повесть, роман; жанр).
3. Порядковые: конечное упорядоченное множество значений (период создания; уровень образования аудитории).
4. Интервальные: интервальное значение (число сложных синтаксических конструкций; число предложений).

Некоторые из бинарных, номинальных и порядковых признаков могут не влиять на принадлежность текста категории (например, структурный тип текста — проза или поэзия). В то время как влияющие признаки данных типов могут либо представлять собой маркеры, ограничивающие круг категорий, с которыми сопоставляется текст, либо подразумевать наличие дополнительных уточняющих признаков. Так, в частности, присутствие ненормативной лексики в тексте однозначно говорит о том, что данный текст не предназначен читателям младших возрастных групп. С другой стороны, бинарный признак, характеризующий наличие иллюстраций в документе, нуждается в уточнениях, касающихся типа изображений. С большой долей вероятности текст, содержащий графики, не может быть адресован младшей возрастной аудитории. Присутствие маркеров в признаковом описании текста позволяет проводить классификацию поэтапно:

Этап 1. Определение условий классификации (уточнение степени влияния бинарных, номинальных и порядковых признаков).

Этап 2. Проверка наличия маркеров (определяющих категорию или ограничивающих круг категорий).

Этап 3. Сопоставление текста категориям на основании интервальных признаков.

Рассматриваемая задача относится к задачам многомерной классификации, где некоторые признаки являются зависимыми от других (например, длина предложений в тексте и сложность их грамматических конструкций связаны друг с другом). Тогда для $a_1,a_2,\ldots,a_L$ — значений признаков $q_1^T,q_1^T,\ldots,q_L^T$ — существует функция:

$$a_j=g_j(a_{j_1},a_{j_2},...,a_{j_p})+a_{j_0},\ 0\leqslant p\leqslant L,\ 1\leqslant j_k\leqslant L. \qquad (11)$$

Функция $g_j(a_{j_1},a_{j_2},...,a_{j_p})$ может быть получена на основе экспертных оценок и дополнительного статистического анализа возможных зависимостей.

Для интервальных признаков также можно предположить, что существует некоторый порог влияния значений $a_{j_1},a_{j_2},...,a_{j_p}$ на a_{j_0}. Тогда при некоторых значениях признаков $a_{j_1},a_{j_2},...,a_{j_p}$, отличных от нулевых, $g_j(a_{j_1},a_{j_2},...,a_{j_p})=0$. В таком случае состояния каждого признака текста можно представить в виде вектора:

$$Q_j^T=(a_{j_0},g_j(a_{j_1},a_{j_2},...,a_{j_p}),a_{j_{max}},a_{j_{min}}), \qquad (12)$$

где a_{j_0} показывает значение признака при отсутствии влияния на него других признаков; $a_{j_{max}}$ и $a_{j_{min}}$ — соответственно максимальное и минимальное значения, которых можно достичь, изменяя значения признаков $a_{j_1},a_{j_2},...,a_{j_p};a_{j_{max}},a_{j_{min}} \in Q_f$.

В данной работе тексты T_i и T_j называются *принадлежащими к одному таксономическому виду*, если:

$$\gamma(T_i)=\gamma(T_j). \qquad (13)$$

Таким образом, тексты относятся к одному таксономическому виду, если они относятся к одним и тем же категориям. Множество текстов можно разбить на непересекающиеся классы эквивалентности, которые в контексте рассматриваемой задачи совпадают с категориями текстов.

Анализ данных показал, что признаки объектов, между которыми устанавливается мера сходства, являются статистически зависимыми. При этом числовая оценка их значимости определяется весовыми коэффициентами. В этом случае в качестве меры близости текстов может быть принято расстояние Махаланобиса. Тогда расстояние между текстом и центром масс категории R, представленным в виде вектора средневзвешенных значений признаков, определяется следующим образом:

$$\rho(F_{T_i},R)=\sqrt{(F_{T_i}-R)^T\Lambda^T C^{-1}(F_{T_i}-R)},$$
$$R=\frac{\sum_{j=1}^{M} k_j F_{T_j}}{M}, \qquad (14)$$

где Λ — матрица весовых коэффициентов; C — матрица ковариации; R — вектор, характеризующий расположение центра масс категорий; M — число текстов данной категории, входящих в обучающую выборку, $1\leqslant M\leqslant L$; k_j — весовой коэффициент доверия тексту обучающей выборки, $k_j>0$, $\sum_{j=1}^{L} k_j=1$.

Пусть k_{min} — пороговое значение весового коэффициента k_j. В случае, когда $k_j< k_{min}$, корректировка значений, составляющих вектор R, не будет произво-

диться. Следовательно, $k_j < k_{min}$ при проведении классификации на контрольной выборке. Таким образом, центр масс категорий для множества L текстов вычисляется следующим образом:

$$R_{(L)} = \begin{cases} \dfrac{(L-1)R_{(L-1)} + k_L F_{T_L}}{L}, & k_L \geq k_{min} \\ R_{(L-1)}, & k_L < k_{min} \end{cases} \qquad (15)$$

$$R_{(1)} = k_1 F_{T_1}.$$

В общем случае, когда значения признаков представлены не только интервальными величинами, в качестве меры близости может быть использовано расстояние хи-квадрат, полученное на основе таблицы сопряженности.

Третья глава содержит описание программного комплекса для автоматической классификации текстов, реализующего предложенные математические модели и методы.

Программный комплекс состоит из трех подсистем (рисунок 1): модуля семантико-синтаксического анализа, модуля классификации текстов и модуля хранения текстов.

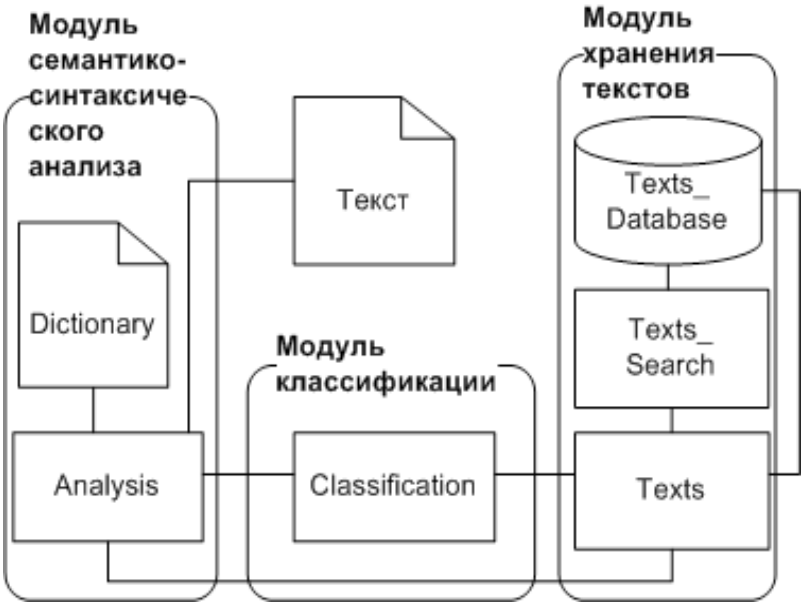

Рисунок 1 — Архитектура программного комплекса

Модуль классификации осуществляет:

1. Определение соответствия текста существующим категориям.
2. Соотнесение текста с категориями на основе классификационных признаков, полученных в ходе работы модуля семантико-синтаксического анализа.

Модуль классификации предоставляет возможность вести работу с программным комплексом в двух режимах: в режиме обучения системы и в контрольном режиме. Работа модуля классификации текстов выполняется поэтапно. Эксперименты показали, что данный подход к реализации позволяет снизить временные затраты на обучение и работу модуля классификации.

Модуль реализует классификацию двумя способами: с помощью описанного в работе численного метода и с использованием нейронной сети типа «многослойный персептрон». Использование данного типа сети обусловлено его способностью к решению слабоформализуемых задач на основании имеющихся примеров и выявлению закономерностей в связи входных и выходных данных. К другим достоинствам многослойного персептрона можно отнести его универсальность и относительную простоту структуры, которые в сочетании с возможностью совершения любых отображений входных векторов значений в выходные обеспечивают достаточную вычислительную мощность.

Модуль хранения текстов предназначен для добавления информации и организации хранения данных в реляционных таблицах базы данных. В базе данных предусмотрена возможность хранения значений признаков различных типов.

Данный модуль выполняет следующие функции:

1. Взаимодействие модулей программного комплекса.
2. Хранение данных о текстах, поступающих в систему, и назначенных им категориях в базе данных.
3. Поиск информации о текстах по заданным критериям, организованный при помощи SQL-запросов.

Программный комплекс разработан на языке C# в среде Visual Studio 2010. Модуль хранения текстов использует СУБД Microsoft SQL Server 2012 Express.

Особенности модульной архитектуры позволяют гибко интегрировать данный программный комплекс в системы работы с электронными документами для решения задач анализа, классификации и хранения текстов. В разработанной системе предусмотрены следующие текстовые форматы обмена данными: .xls, .xlsx, .xml, .txt.

В ходе вычислительного эксперимента, а также разработки и тестирования программного комплекса использовались база данных «Морфологический стандарт Национального корпуса русского языка» и «База данных метатек-

стовой разметки Национального корпуса русского языка» (коллекция детской литературы)». Оба источника данных содержат заведомо качественные и максимально разнообразные тексты на русском языке, возрастная категория потенциальных читателей которых — взрослая или детская — определена на основании мнений экспертов. Объем выборки — 532 текста художественной литературы и 510 текстов детской литературы.

Точность классификации для метода классификации текстов, представленного в работе, составила 74,16% (среднеквадратическое отклонение — 5,88%), для нейросетевого метода — 72,07% (среднеквадратическое отклонение — 6,62%).

Если считать целью классификации фильтрацию текстов, не предназначенных детской возрастной аудитории (отсечение текстов, адресованных взрослым читателям), то в данном эксперименте можно рассмотреть ошибки двух типов: ошибка первого рода (доля случаев, когда текст, адресованный детской возрастной группе, не был отнесен к категории детских текстов); ошибка второго рода (доля случаев, когда текст, адресованный взрослой возрастной группе, был отнесен к категории детских текстов). Величина ошибки первого рода для предложенного метода классификации текстов составила 26,67%, для нейросетевого метода — 28,43%. Величина ошибки второго рода — 24,81% и 27,44% соответственно.

Учитывая объем выборки, использовавшейся для проведения вычислительного эксперимента, может быть сделан вывод о том, что метод классификации текстов, предложенный в работе, показал эффективность в сравнении с нейросетевым методом.

В **заключении** сформулированы основные результаты диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ И ВЫВОДЫ

1. Предложенный подход к моделированию классификации текстов (на примере их отнесения к той или иной возрастной категории адресатов) развивает существующие математические модели классификации за счет возможности учесть особенности поставленной в работе задачи при ее формализации (вложенность категорий, пересечение категорий). На основе предложенного подхода показана возможность формализации постановки задачи классификации в общем виде.
2. Построенная на основании предложенного подхода математическая модель процесса классификации текстов развивает возможности математического моделирования для постановки и решения слабоформализуемых задач (в частности, задач классификации), поскольку имеет

универсальный характер и может применяться для классификации других видов объектов, модели которых могут быть описаны сходными классификационными признаками.

3. Принцип разбиения множества текстов на классы эквивалентности позволил реализовать численный метод классификации текстов, дающий возможность определить меру близости текстов как расстояние между векторами значений характеризующих их классификационных признаков.
4. В рамках задачи классификации текстов на основании их возрастной аудитории осуществлен подбор информативных классификационных признаков для экспериментальной проверки предложенных подходов и методов.
5. Разработанный программный комплекс — интеллектуальная система для автоматической классификации текстов — получил практическое применение в некоммерческом партнерстве по содействию развитию науки и образования «Национальный корпус русского языка» и в негосударственном образовательном учреждении «Югорский учебный центр», а также в перспективе может быть внедрен в различные системы, осуществляющие обработку текстов на естественном языке.
6. Экспериментальная проверка результатов с использованием текстов, включенных в Национальный корпус русского языка, подтвердила адекватность разработанных моделей и методов автоматической классификации текстов, а также преимущество предложенного численного метода, состоящее в возможности анализа степени зависимости результатов классификации от значений различных классификационных признаков.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в периодических изданиях, рекомендованных ВАК

1. Глазкова А.В. Подход к моделированию задачи автоматической классификации текстов (на примере их отнесения к определенной возрастной аудитории) [Текст] / А.В. Глазкова, И.Г. Захарова // Вестник Тюменского государственного университета. — 2014, №7. — С. 205-211.
2. Глазкова А.В. Подход к формализованному описанию зависимости признаков в задаче классификации текстов [Текст] / А.В. Глазкова // В мире научных открытий. — 2014, №12.2. — С. 717-726.
3. Глазкова А.В. Оценка степени близости категорий текстов при решении задач классификации электронных документов [Текст] / А.В. Глазкова // Вестник Томского государственного университета. — 2015, №2. — С. 18-25.

Публикации в других изданиях

4. Кружинова А.В. Разработка и применение информационной системы синтаксического анализа для лингвистических исследований [Текст] / Ю.В. Бидуля, Ю.В. Глухова, А.В. Кружинова // Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений: сб. тр. четвёртой научно-практической региональной конференции. — Тюмень: Издательство ТюмГУ, 2011. — С. 15-19.
5. Глазкова А.В. Задачи и методы автоматической классификации текстов [Текст] / А.В. Глазкова // Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений: материалы VI научно-практической региональной конференции. — Тюмень: Издательство ТюмГУ, 2013. — С. 109-114.
6. Глазкова А.В. Основные подходы к решению задач автоматической классификации документов [Текст] / А.В. Глазкова // Современные проблемы математического и информационного моделирования. Перспективы разработки и внедрения инновационных IT-решений: материалы VI научно-практической региональной конференции. — Тюмень: Издательство ТюмГУ, 2013. — С. 114-119.
7. Глазкова А.В. Роль и перспективы развития технологий автоматической обработки текстов в современном учебном процессе [Текст] / А.В. Глазкова // Преподавание информационных технологий в Российской Федерации: материалы XI открытой всероссийской конференции. — Воронеж: Воронежский государственный университет, 2013. — С. 149-150.
8. Глазкова А.В. Возможность автоматического определения адресата на основе семантико-синтаксических особенностей текста [Текст] / А.В. Глазкова // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2014): мате-

риалы IV международной научно-технической конференции. — Минск: Белорусский государственный университет информатики и радиоэлектроники, 2014. — С. 509-513.
9. Глазкова А.В. Некоторые аспекты осуществления автоматической классификации текстов на основе распознавания их адресатов [Текст] / А.В. Глазкова // Интеллектуальные системы и прикладная лингвистика: материалы III всеукраинской научно-практической конференции. — Харьков: Национальный технический университет «Харьковский политехнический институт», 2014. — С. 26-28.
10. Глазкова А.В. Проверка информативности классификационных признаков в задаче автоматической классификации текстов на естественном языке [Текст] / А.В. Глазкова // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2015): материалы V международной научно-технической конференции. — Минск: Белорусский государственный университет информатики и радиоэлектроники, 2015. — С. 541-544.
11. Глазкова А.В. Создание прототипа программного комплекса для классификации текстов на естественном языке [Текст] / А.В. Глазкова // Материалы 53-й Международной научной студенческой конференции МНСК-2015: Информационные технологии. — Новосибирск: Новосибирский государственный университет, 2015. — С. 159.
12. Глазкова А.В. Использование морфологических характеристик слов текста в качестве классификационных признаков [Текст] / А.В. Глазкова // Математическое и информационное моделирование: сборник научных трудов. Вып. 14. — Тюмень: Издательство ТюмГУ, 2015. — С. 59-62.

Свидетельство о регистрации программ для ЭВМ

13. Глазкова А.В. Расчёт оценки степени близости категорий текстов при решении задач классификации электронных документов. РОСПАТЕНТ. Свидетельство №2015616462 от 10.06.2015.

Подписано в печать 24.06.2016. Тираж 120 экз.
Объем 1,0 уч.-изд. л. Формат 60×84/16. Заказ 582.

Издательство Тюменского государственного университета
625003, г. Тюмень, ул. Семакова, 10
Тел./факс: (3452) 59-74-81, 59-74-68
E-mail: izdatelstvo@utmn.ru