Национальный исследовательский ядерный университет «МИФИ»

На правах рукописи

Кузнецов Игорь Александрович

Методы и алгоритмы машинного обучения для предобработки и классификации слабоструктурированных текстовых данных в научных рекомендательных системах

Специальность: 05.13.01 – Системный анализ, управление и обработка информации (в информационных системах)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва – 2019

Работа выполнена в Национальном исследовательском ядерном университете «МИФИ»

Научный руководитель:

профессор, доктор технических наук, почетный работник сферы образования РФ **Гусева Анна Ивановна**	Национальный исследовательский ядерный университет «МИФИ»

Официальные оппоненты:

доктор технических наук **Захаров Виктор Николаевич**	Федеральный исследовательский центр «Информатика и управление» РАН, ученый секретарь
профессор, доктор технических наук, заслуженный работник высшей школы РФ **Пылькин Александр Николаевич**	Федеральное государственное бюджетное образовательное учреждение высшего образования «Рязанский государственный радиотехнический университет имени В.Ф. Уткина», декан факультета вычислительной техники
профессор, кандидат технических наук **Лупин Сергей Андреевич**	Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «МИЭТ», профессор Института микроприборов и систем управления

Защита диссертации состоится «20» ноября 2019 г. в 15 часов 00 минут на заседании диссертационного совета МИФИ.05.01 федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ» (115409, г. Москва, Каширское шоссе, 31).

С диссертацией можно ознакомиться в библиотеке и на сайте https://ds.mephi.ru федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ».

Автореферат разослан «__» _____ 2019 г.

Ученый секретарь
диссертационного совета, д.т.н. Леонова Наталия Михайловна

ВВЕДЕНИЕ

Диссертационная работа направлена на решение научно-технической задачи создания, применения и исследования эффективности методов и алгоритмов машинного обучения для обработки слабоструктурированных текстовых данных в научных рекомендательных системах.

Актуальность исследования. Переход от бумажных носителей информации к цифровым, вызванный повсеместным использованием информационных и телекоммуникационных технологий, открывает перспективы для работы с постоянно возрастающим объемом информации и возможностью извлечения знаний из слабоструктурированного массива данных. Несмотря на относительно небольшой срок активного применения в промышленности цифровых носителей, объем данных на них растет год от года в геометрической прогрессии, что является следствием преобразования и перевода данных из различных областей жизнедеятельности человека в цифровой вид. К 2016 году накопленное количество данных оценивалось в 16 зеттабайт, но по прогнозам аналитической компании IDC к 2025 году объем всех данных достигнет отметки в 163 зеттабайт. В связи с этим на первый план выходят способы хранения, обработки, поиска и извлечения знаний из данных. Эта технология получила название сквозной цифровой технологии «Большие данные».

К 2024 году государство намерено осуществить комплексную цифровую трансформацию экономики и социальной сферы России. Направление по обработке «больших данных» входит в зону ответственности ГК «Росатом», где в рамках создания цифровой платформы «Распределенная среда обработки больших данных» предусмотрены работы по развитию компонентов систем поддержки принятия решений, программных интерфейсов доступа к данным и др.

Ввиду возрастания объема данных и необходимости осуществлять поиск среди них, на первый план выходит задача определения релевантности и пертинентности информации. Понятие «релевантность» определяется как «соответствие полученной информации информационному запросу». Таким образом, релевантность определяется исключительно используемыми математическими моделями в конкретной информационно-поисковой системе. Под понятием «пертинентность» понимается «соответствие полученной информации информационной потребности», т.е. пертинентность – соответствие найденных информационно-поисковой системой документов информационным потребностям пользователя независимо от того, как полно и точно эта потребность выражена в форме запроса.

Один из подходов для повышения пертинентности информационного поиска – использование методов машинного обучения как частный случай применения интеллектуального анализа данных. Данное направление исследования лежит на пересечении целого ряда дисциплин: математики, информатики, статистики, теории вероятностей и др. Выделяют особый подкласс информационных систем, опирающийся на математические модели и позволяющий решать задачи по определению релевантности и пертинентности данных, который называется рекомендательными системами. Отсутствие универсальных подходов для решения задач предобработки и классификации в рекомендательных системах делает разработку методов и алгоритмов машинного обучения в этом случае особо актуальной.

Предложенные в диссертационной работе методы и алгоритмы направлены на увеличение точности и стабильности работы методов машинного обучения при предобработке и классификации больших объемов слабоструктурированных текстовых данных и повышение пертинентности информационного поиска в научных рекомендательных системах.

Степень разработанности проблемы. Исследованию, разработке и развитию математических моделей и алгоритмов классификации посвящены работы отечественных и зарубежных ученых Воронцова К.В., Дьяконова А.Г., Пылькина А.Н., Лупина С.А., Аматриаин К., М. де Геммис, Бреймана Л., Блей Д.

Вопросом повышения пертинентности информации и развитием рекомендательных систем занимались Пальчунов Д.Е, Захаров В.Н., Адомавичус Г., Бурке Р., Ступников С.А., Фелферниг А., Констан Д.

Развитием сквозной цифровой технологии «Большие данные» и интеллектуальным анализом данных занимаются Цветков В.Я., Бойд Д., Ву Ксиндонг, Линч Клиффорд, Кузнецов С.Д.

Объектом исследования диссертационной работы являются научные рекомендательные системы.

Предметом исследования выступают математические методы и алгоритмы машинного обучения, используемые в научных рекомендательных системах.

Теоретическую базу исследования составили фундаментальные научные труды российских и зарубежных авторов, касающиеся разработки методов и алгоритмов обработки данных, принципов формирования ансамблевых подходов, подходов работы с «большими данными», принципов Data Mining и Text Mining, теории автоматизированного управления, поддержки принятия решений и системного анализа.

Информационной базой исследования являются монографии, посвященные проблемам анализа и обработки информации; публикации в научных журналах; Российская государственная библиотека; Российская научно-техническая библиотека; научная база ВИНИТИ; научная электронная библиотека КиберЛенинка; материалы научно-практических конференций, сети Интернет.

Цель диссертационного исследования состоит в разработке математических методов и алгоритмов машинного обучения для обработки и анализа больших объемов слабоструктурированных текстовых данных в научных рекомендательных системах.

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

- исследовать существующие подходы к построению рекомендательных систем, выполнить системный анализ методов машинного обучения для решения задачи классификации, выявить наиболее перспективные методы для работы со слабоструктурированными текстовыми данными больших объемов;
- разработать и исследовать метод и алгоритмы обогащения признакового пространства в рамках решения задачи предобработки слабоструктурированного набора текстовых данных;
- разработать и исследовать ансамблевый метод и алгоритмы для решения задачи классификации при обработке слабоструктурированных текстовых данных;
- провести апробацию предложенных методов и алгоритмов в виде разработанных программных средств повышения пертинентности информации в научных рекомендательных системах;
- провести экспериментальную проверку эффективности использования предложенных методов и алгоритмов машинного обучения в научных рекомендательных системах.

Методы исследования. В работе использовались методы структурного системного анализа, методы теории управления и принятия решений, методы интеллектуального анализа данных и машинного обучения, методология объектно-ориентированного проектирования RUP (Rational Unified Process) и подходы, применяемые при разработке программного обеспечения.

Научная новизна заключается в следующем:

- обоснован параметрический подход для решения задачи предобработки слабоструктурированного набора текстовых данных на основе извлечения нового семантического параметра из текстовых публикаций – вида научного результата;
- на основе параметрического подхода предложен метод и разработаны алгоритмы предобработки текстовых публикаций в научных рекомендательных системах, позволяющие существенно повысить пертинентность информационного поиска;

- разработаны и исследованы новый метод и алгоритмы ансамблевой классификации, обладающие высокой прогнозной точностью и стабильностью, основанные на комбинации существующих алгоритмов с использованием энтропии в качестве меры взвешивания. Использование энтропии позволяет существенно повысить качество методов машинного обучения для решения задачи классификации по сравнению с аналогами;
- проведено экспериментальное исследование эффективности применения методов предобработки и ансамблевой классификации на слабоструктурированных текстовых научных данных.

Теоретическая и практическая значимость работы заключаются в следующем.

1. Проведенные исследования принципов формирования различных видов рекомендательных систем показывают, что использование контентной фильтрации способствует устранению таких проблем функционирования, как «холодный старт», «новый пользователь», и идентификации контекста для пользователя.

2. Предложенный параметрический подход, разработанные на его основе метод и алгоритмы обогащения признакового пространства слабоструктурированных научных данных позволяют выделить новый семантический параметр – вид научного результата, приводящий к расширению онтологии описания научной деятельности.

3. Проведенный системный анализ методов машинного обучения для решения задачи классификации показал, что наиболее острой проблемой при работе со слабоструктурированными данными является несоответствие между структурами данных обучающей выборки и генеральной совокупности, для уменьшения этого несоответствия предложено использовать ансамблевые методы.

4. Предложенный метод и разработанные алгоритмы ансамблевой классификации на основе энтропии позволяют повысить стабильность и точность работы ансамбля на больших объемах слабоструктурированных текстовых данных.

5. Разработанное инструментальное средство классификации данных защищено свидетельством о государственной регистрации и было успешно использовано в рамках выполнения проекта № 15-07-08742 РФФИ, что подтверждается соответствующим актом об использовании.

6. Разработанные базы данных учебно-методических материалов и научных публикаций защищены свидетельствами о госрегистрации и используются в течение ряда лет в учебной деятельности в НИЯУ МИФИ, что подтверждается соответствующим актом об использовании.

7. Разработанные методы и алгоритмы были реализованы в информационной системе Международный конгресс конференций «Информационные технологии в образовании» и показали свою эффективность, что подтверждается соответствующим актом о внедрении.

Достоверность полученных результатов. Научные положения и выводы, полученные в диссертационной работе, являются достоверными и обоснованными, что подтверждается использованием научной методологии исследования, достаточно большим объемом обработанных отечественных и зарубежных источников по теме исследования, последовательным подходом к решению поставленных задач, проведенными в работе экспериментами, анализом полученных результатов в сравнении с другими моделями, соответствующими актами об использовании, а также обсуждением основных положений диссертации на международных и российских научно-практических конференциях.

Научные результаты, полученные лично автором и выносимые на защиту.

1. Параметрический подход, метод и алгоритмы для извлечения нового семантического параметра – вида научного результата из слабоструктурированных текстовых данных научных публикаций, позволяющий повысить пертинентность информационного поиска.

2. Метод и алгоритмы машинного обучения с использованием энтропии в качестве меры взвешивания для ансамблевой классификации, отличающиеся стабильностью и точ-

ностью по сравнению с известными аналогами и позволяющий повысить пертинентность информационного поиска.

3. Результаты экспериментального исследования применимости алгоритмов предобработки и ансамблевой классификации на слабоструктурированных текстовых данных, показывающие прирост пертинентности при информационном поиске в научных рекомендательных системах.

4. Разработанное инструментальное средство для решения задачи классификации и набор баз данных для использования в научно-исследовательской и учебной деятельности.

5. Разработанная научная рекомендательная система для информационной системы Международного конгресса конференций «Информационные технологии в образовании».

Авторский вклад. Все результаты диссертационной работы получены либо лично автором, либо при его непосредственном участии.

Область исследования. Область диссертационного исследования соответствует по своему содержанию Паспорту научных специальностей ВАК Министерства образования РФ и науки РФ по специальности 05.13.01 «Системный анализ, управление и обработка информации (в информационных системах)»: п.1. «Теоретические основы и методы системного анализа, оптимизации, управления, принятия решений и обработки информации»; п.4. «Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации»; п.12. «Визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации».

Апробация результатов исследования. Основные положения и результаты диссертационного исследования были успешно доложены и обсуждены на XIX Международной телекоммуникационной конференции молодых ученых и студентов «Молодежь и наука» (Москва, 2015 г.); Международной научно-практической конференции «Информационные технологии в образовании XXI века» (Москва, 2015 г.); Международном научно-техническом семинаре «Современные технологии в задачах управления, автоматики и обработки информации» (Алушта, Республика Крым, 2016 г.); XVIII международной конференции DAMDID/RSDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных» (Ершово, Московская область, 2016 г.); XXII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов «Новые информационные технологии в научных исследованиях» НИТ-2017 (Рязань, 2017 г.); XIX Международной конференции DAMDID / RCDL'2017 «Аналитика и управление данными в областях с интенсивным использованием данных» (Москва, 2017 г.); Innovate-Data 2017 The 3rd International Conference on Big Data Innovations and Applications, IEEE-CS TCI (Prague, Czech Republic, 2017 г.); V Всероссийской научной конференции молодых ученых с международным участием «Информатика, Управление и Системный Анализ» ИУСА-2018 (Ростов, 2018 г.); XXIII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов «Новые информационные технологии в научных исследованиях» НИТ-2018 (Рязань, 2018 г.); 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (Москва, 2019 г.).

Внедрение результатов исследования. Результаты диссертационного исследования использовались при выполнении проектов:

- № 2014-14-576-0146 «Разработка метода и программно-технических решений повышения пертинентности информации в научных и аналитических рекомендательных системах» (2014-2016, ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2020 годы»);
- № 15-07-08742 «Принципы создания алгоритмического обеспечения для многомерной классификации на примере анализа научных направлений» (2015-2017, РФФИ).

Научные результаты успешно используются в информационной системе Международного конгресса конференций «Информационные Технологии в Образовании», что подтверждает соответствующий акт о внедрении. Научные и практические результаты дис-

сертационного исследования были использованы в учебном процессе НИЯУ МИФИ в рамках научно-практического семинара для магистров «Информационные технологии в науке и образовании», что подтверждается соответствующим актом об использовании.

Публикации. Основные результаты диссертации опубликованы в 25 печатных работах, из них 5 статей в периодических научных изданиях, рекомендованных ВАК РФ; 5 в журналах и сборниках трудов конференций, включенных в базу SCOPUS и/или Web of Science; главы в монографии и 5 работ в статьях и материалах конференций; 8 свидетельств о регистрации баз данных и программ для ЭВМ.

Структура и объем работы. Диссертация состоит из введения, пяти глав, заключения, списка литературы и трех приложений. Основной текст работы изложен на 116 страницах, приложения – на 11 страницах текста. Иллюстративный материал включает 34 рисунка и 12 таблиц. Список литературы содержит 103 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы работы, указаны объект и предмет исследования, цели и задачи работы, научная новизна, теоретическая и практическая значимость полученных результатов, выносимые на защиту положения, сведения об апробации и публикации результатов, структура и объем диссертации.

В **первой главе** представлен сравнительный анализ современных рекомендательных систем, рассматриваются основные проблемы и потребности в данной области. Формулируются основные проблемы рекомендательных систем. Проведено обобщение и систематизация математических алгоритмов классификации в рамках машинного обучения, выявлены их проблемы и пути развития, предложена авторская модель классификации на основе ансамбля алгоритмов.

Основным назначением рекомендательных систем является прогнозирование поведения пользователя по отношению к объекту информационного поиска, а также формирование последующей рекомендации схожего информационного объекта, с которым он ранее не встречался.

Формальная постановка задачи для рекомендаций выглядит следующим образом. Рассмотрим U – множество пользователей и D – множество объектов. Необходимо найти функцию r, $r : U \times D \rightarrow R$, которая формирует рекомендацию R таким образом, что для любого пользователя значение r между ним и объектом d максимально, т.е. является аргументом максимизации:

$$\forall u \in U, \qquad d_u = arg(\max_{d \epsilon D} r(u, d)). \tag{1}$$

Выделяют четыре вида рекомендательных систем:

- рекомендации, формируемые экспертным методом – связи между объектами устанавливаются вручную или на основе заранее определенных правил, но данный способ является актуальным только при небольшом перечне рекомендуемых объектов;
- коллаборативная фильтрация – рекомендации основываются на оценках пользователей по отношению к просмотренным объектам. Рекомендации могут строиться либо на основе поиска схожих пользователей по отношению к рассматриваемому пользователю (user-based), либо на основе поиска схожих объектов по отношению к объектам с выставленными ранее оценками, рассматриваемым пользователем (item-based);
- контентная фильтрация – рекомендации для рассматриваемого пользователя формируются на основе понравившихся ему объектов с учетом присвоенных каждому объекту набору параметров;

- гибридная фильтрация – используется комбинация подходов, основанных на контентной и коллаборативной фильтрации, что приводит к повышению качества формируемых рекомендаций.

Рекомендательные системы имеют несколько серьёзных недостатков, среди которых можно выделить сложность масштабирования, «холодный старт», новый пользователь, идентификация контекста для пользователя, небольшая шкала для выставления рейтингов, определение отрицательных предпочтений и временные характеристики. Построение рекомендательных систем на базе контентной фильтрации с использованием вспомогательных данных о пользователе и объекте рекомендаций способствуют устранению проблемы «холодного старта», нового пользователя, идентификации контекста для пользователя и проблемы временных характеристик.

Целью любой рекомендательной системы является рекомендация информационных объектов, ранее неизвестных пользователю, но полезных или интересных в текущем контексте. Рекомендательная система отвечает на вопрос о конкретном посетителе сайта: в каком информационном объекте этот посетитель заинтересован прямо сейчас? Общий принцип работы рекомендательной системы приведен на схеме (рис. 1).

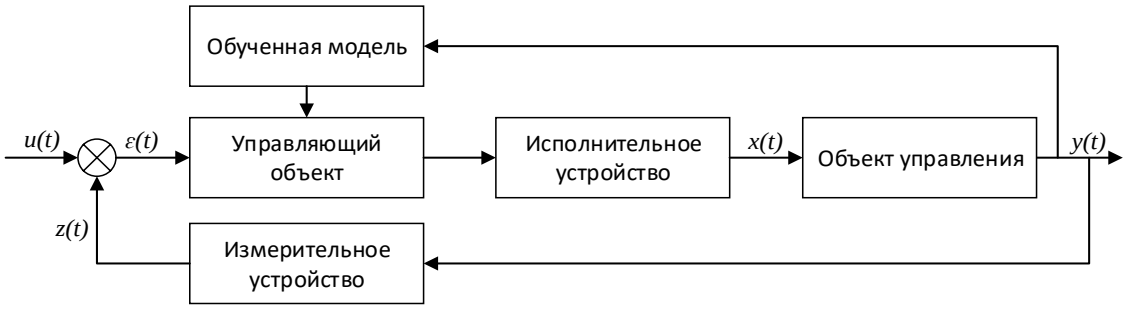

Рисунок 1. Принцип работы рекомендательных систем

Основными элементами функциональной схемы являются следующие объекты.

- *Объектом управления* является блок с перечнем рекомендаций, который должен соответствовать интересам пользователя в определенный момент времени.
- *Управляющим объектом* выступает набор алгоритмов классификации, которые позволяют на основе задающего воздействия $u(t)$ и сигнала от измерительного устройства формировать наиболее пертинентные значения.
- *Обученная модель* содержит представление рассматриваемых элементов и правила, по которым управляющий объект будет выполнять классификацию.
- Под *исполнительным устройством* подразумевается веб-сервер, который дополняет содержательной информацией из БД полученные идентификаторы значений от управляющего объекта и передает управляющее воздействие $x(t)$.
- *Измерительное устройство* выполняет сбор данных по реакции пользователя на полученные рекомендации и преобразует их в формализованные значения $z(t)$ поведения пользователя на странице интернет сайта.

Выходными параметрами данной схемы являются рекомендации, которые должны представлять наибольший интерес для пользователя согласно его интересам и потребностям. Применимость рекомендательных систем не ограничивается какими-то определенными сферами деятельности, а может быть использована во множестве различных областей, в частности для систем, содержащих научную информацию.

Научная рекомендательная система (НРС) представляет собой специальный модуль, который может быть установлен поверх базы данных системы с научной информацией и использоваться независимо от поисковой системы. К таким системам могут быть отнесены Google Scholar, Scopus, Web of Science, Mendeley, eLibrary, Cyberleninka и другие.

Системы подобного плана предназначаются для ученых, преподавателей, учителей, студентов и прочих заинтересованных лиц. Основная задача НРС – предсказать информационную потребность пользователя, где источником данных может являться сам профиль пользователя, его поведение на сайте и его поисковые запросы. На основе собранных данных формируется индивидуальный набор рекомендаций, который должен отражать текущую потребность пользователя.

Генерируемые данные в современном мире становятся активом, который способен приносить дополнительную пользу тем, кто умеет извлекать информацию и правильно ее использовать. Технология, которая отвечает за способы хранения и обработки, получила название сквозной цифровой технологии «Большие данные». Несмотря на полученное название, технология «Большие данные» характеризуется не только объемом данных, но также имеет ряд других значащих параметров. Наиболее полно данную технологию характеризуют так называемые V-модели, которые включают в себя различные параметры, описывающие данные. К наиболее часто используемым параметрам относятся объем данных (Volume), скорость прироста данных и скорость их обработки (Velocity), разнообразие данных (Variety). Количество «V», которые присутствуют в этих моделях постоянно увеличивается и иногда составляет более 40 параметров.

В результате систематизации основных подходов к реализации сквозной цифровой технологии «Большие данные» в диссертации был предложен набор основных параметров, с помощью которого можно охарактеризовать данные в рамках научных рекомендательных систем (табл. 1). Также приведена оценка важности того или иного параметра по шкале от 1 до 5, где 5 – максимальная важность. Оценка важности выделенных параметров была получена в рамках экспертной работы по гранту РФФИ № 15-07-08742.

Таблица 1. Основные параметры «больших данных» для НРС

Параметры	Описание	Важность
Объем (Volume)	Насколько много данных было накоплено за предыдущее время	2
Разнообразие (Variety)	Насколько структурированной является информация, на сколько много источников и как много форматов обрабатываемых данных	5
Скорость (Velocity)	Скорость создания и накопления данных, а также скорость обработки данных	4
Ценность (Value)	Насколько полезными являются накопленные данные	5
Достоверность (Veracity)	Насколько достоверны полученные данные	5

Сформированный индивидуальный набор рекомендаций для пользователя дает возможность повысить эффективность использования НРС, при этом возрастает удовлетворенность пользователя и сокращается время поиска необходимой информации. Эффективность можно оценить с помощью пертинентности, которая определяется субъективным восприятием человека. В формальном виде пертинентность определяется следующим образом:

$$Prt = \frac{1}{N}\sum_{i=1}^{N}\left|l_i \cap Q_{req}\right|, \quad l \in L_{sort}, N > 0, \tag{2}$$

где $l = \{w_1, w_2, \ldots, w_y\}$ – веб-страница в упорядоченном списке L_{sort}, выдаваемая поисковой системой, выраженная множеством слов; $Q_{req} = \{w_1, w_2, \ldots, w_y\}$ – эталонное множество слов, выражающее потребности пользователя; w – слово, приведенное к нормальной форме; N – количество элементов в упорядоченном списке поисковой выдачи.

Машинное обучение является подразделом в области искусственного интеллекта, направленным на изучение методов построения самообучающихся алгоритмов с целью

поиска закономерностей и последующим формированием прогнозов на их основе. Задача классификации относится к разделу машинного обучения, который называется «обучение с учителем». Одним из хорошо зарекомендовавших себя подходов к формированию алгоритмов классификации является комбинирование нескольких простых алгоритмов классификации в один более сильный. Такой подход называется ансамблем (комитетом) голосующих алгоритмов.

Формальная постановка задачи классификации выглядит следующим образом. Рассмотрим множество объектов (документов) $D = \{d_1, d_2, \ldots, d_i, \ldots, d_n\}$ и конечное множество меток классов (категорий) $C = \{c_1, c_2, \ldots, c_j, \ldots, c_k\}$. Существует такая функция α: $D \times C \rightarrow \{0,1\}$, которая задается следующим образом:

$$\alpha(d_i, c_j) = \begin{cases} 0, \text{если } d_i \notin c_j, \\ 1, \text{если } d_i \in c_j. \end{cases} \quad (3)$$

Таким образом, необходимо построить классификатор α', максимально близкий к α для каждого объекта из множества D к своему классу с наибольшей возможной точностью.

При решении задачи классификации часто возникает сложность с выбором методов машинного обучения, так как их количество велико, а проверить качество работы каждого в отдельности не представляется возможным. При этом каждый из методов обладает своими слабыми и сильными сторонами, имеет свою область применимости, где иногда на первое место выходит не качество работы, а способность интерпретации полученных результатов. Среди самых часто используемых и зарекомендовавших себя методов можно считать деревья решений, наивный Байес, логистическая регрессия, метод опорных векторов, метод k-ближайших соседей, нейронные сети, ансамбли алгоритмов.

В табл. 2 приведено сравнение основных методов классификации по нескольким параметрам, где для каждого из них была приведена оценка по шкале от 0 до 2, где 2 – максимальный балл.

Таблица 2. Сравнительная характеристика методов классификации

Методы машинного обучения / Параметры	Деревья решений	Наивный Байес	Алгоритм SVM	k-ближайших соседей	Случайный лес	Логистическая регрессия	Нейронные сети	Ансамблевые методы
Интерпретируемость результатов	2	1	1	2	0	1	0	1
Высокая скорость обучения	2	2	0	2	0	2	0	0
Высокая скорость классификации	2	2	2	1	1	2	2	2
Низкое количество параметров для настройки	1	2	1	2	1	2	0	1
Работа с малым объемом данных	1	2	0	1	0	2	0	1
Постоянство структуры вне зависимости от релевантности параметров	0	2	2	1	2	2	2	2

Представленные методы имеют разные области применимости, разные требования к набору входных данных и типу параметров, имеют разные вычислительные характеристи-

ки и уровень потребления памяти. Использование данных методов классификации в рекомендательных системах будет также давать различный результат, следовательно, для каждой задачи и каждого набора данных необходимо использовать свой метод классификации, который может основываться на ансамблевом подходе и включать в себя множество методов с последующим объединением результатов при принятии итогового решения.

Кроме того, построение ансамблей из различных алгоритмов позволяет устранить один из недостатков машинного обучения, когда структура небольшой обучающей выборки в силу своего размера мало соответствует структуре всей рассматриваемой совокупности данных, особенно на больших данных. В этом случае обучение каждого алгоритма может проводиться на разных обучающих выборках.

Формальная постановка задачи. Построение пертинентного индивидуального набора рекомендаций для пользователя можно представить с помощью принципа семантического эквивалентирования $\Xi = \langle \Psi_0, \Psi_1, P(\Psi_0, \Psi_1) \rangle$, где Ψ_0 – модель информационного поиска, Ψ_1 – модель автоматизированного формирования рекомендательного набора, а $P(\Psi_0, \Psi_1)$ – предикат функциональной целостности, отражающий правомерность перехода между ними.

Модель Ψ_0 совпадает со стандартной моделью информационного поиска и имеет вид $\Psi_0 = <M_0, S_0>$, где носитель M_0 включает следующие компоненты: тип (V), рубрикатор (Rub), ключевые слова (KW) и год публикации (Y), а также запрос пользователя (Z). Сигнатура S_0 отражает отношение идентичности.

Переход от модели информационного поиска Ψ_0 к модели автоматизированного построения рекомендаций $\Psi_1 = <M_1, S_1>$ происходит за счет расширения носителя и сигнатуры.

Целевая модель Ψ_1 расширяется за счет включения в носитель M_1 дополнительных компонентов, которые оказывают непосредственное влияние на формирование рекомендательного набора:

- идентификация схожих объектов (Ds) для текстовых данных включает в себя носитель, состоящий из набора слов, представленных в формате модели «Bag of Words», а сигнатурой выступает отношение принадлежности объекта к определенному классу;
- выявление пользовательской потребности на основе вида научного результата (T), где носителем является текстовая аннотация научных статей, а сигнатуру определяет отношение принадлежности объекта одному из видов научного результата.

Сигнатура S_1 расширяется за счет использования пертинентности (Prt) как критерия эффективности при оценке соответствия сформированного рекомендательного набора потребностям пользователя. В общем виде целевая модель будет иметь следующее представление $\Psi_1 = \langle \{M_0, Ds, T\}, \{S_0, Prt\} \rangle$.

Во второй главе обоснован параметрический подход для извлечения нового признака из текстовых публикаций – вида научного результата. Практическое назначение данного подхода основано на том, что при выполнении научного поиска каждый исследователь или ученый вынужден просматривать огромное количество материала, который не соответствует исходной потребности. Извлечение вида научного результата строится именно на потребности пользователя в контенте, обладающим определенным свойством. Можно сформулировать гипотезу о потребностях пользователя следующим образом.

- Пользователь хочет ознакомиться с предметной областью и ее основами.
- Пользователь хочет понять, в каком направлении развивается интересующая его область и какие есть проблемы.
- Пользователь хочет понять, каких результатов уже удалось добиться на текущий момент времени.
- Пользователь хочет понять, что уже апробировано и готово к использованию.

Данную гипотезу можно представить в виде схемы по видам научного результата исходя из потребности пользователя (рис. 2).

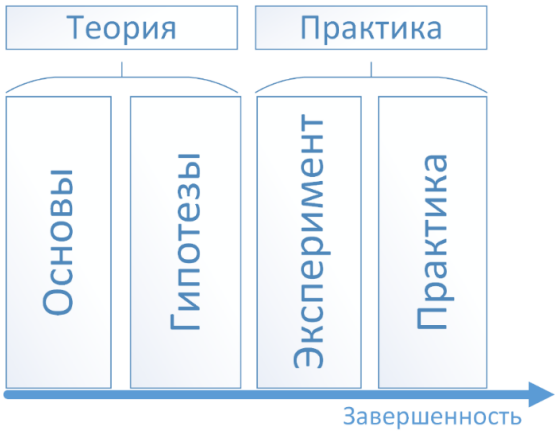

Рисунок 2. Гипотеза о видах научного результата

Предложенный подход позволяет выявить вид научного результата на основе текстовых данных научных статей и способствует определению схожести как между информационными объектами, так и между пользователями, чьи потребности могут быть выражены с помощью полученного признака.

При оценке степени схожести документов или пользователей, использовались различные метрики расстояния: квадрат евклидова расстояния, Манхэттенское расстояние, расстояние Чебышева и некоторые другие.

Метод предобработки слабоструктурированных текстовых данных научных статей для выявления вида научного результата имеет следующий вид.

Шаг 1. Выделение из текста значащих словосочетаний.

Основной задачей данного шага является формирование словосочетаний на основе значащих слов за счет доказательства того, что слова в тексте встречаются неслучайным образом. Дополнительным ограничением при формировании словосочетаний является обязательное наличие среди слов причастия в полной или краткой форме.

Нулевая гипотеза (H_0) предполагает, что встречающиеся в тексте два слова рядом не являются словосочетанием, т.е. появляются вместе случайно, альтернативная гипотеза (H_1) предполагает обратное:

$$\begin{aligned} H_0&: P(w_1, w_2) = P(w_1)P(w_2)\,, \\ H_1&: P(w_1, w_2) \neq P(w_1)P(w_2)\,, \end{aligned} \tag{4}$$

где $p(w_1)$, $p(w_2)$ – вероятность появления слов w_1 и w_2 в текстовом документе.

Для проверки статистических гипотез были использованы t-критерий Стьюдента

$$t = \frac{|\bar{x} - m|}{\sqrt{\frac{s^2}{N}}}, \tag{5}$$

где $\bar{x}$ – выборочное среднее, m – мат. ожидание, s^2 – дисперсия, N – размер непустой выборки; и критерий Хи-квадрат

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{6}$$

где $O_{i,j}$ – фактическое количество словосочетаний, $E_{i,j}$ – ожидаемое количество словосочетаний.

Полученные значения по каждому словосочетанию позволяют принять или отвергнуть нулевую гипотезу путем сравнения с уровнем значимости, который составляет $\alpha = 0.005$. Нулевая гипотеза отвергается тогда, когда полученное значение критерия превышает значение из таблицы критических значений соответствующего критерия.

Шаг 2. Проведение нечеткой кластеризации.

Для каждого текстового документа формируется множество значащих словосочетаний, входящих в этот документ. Набор текстовых документов может быть поделен на несколько тематических классов с помощью алгоритма латентного размещения Дирихле.

Данный алгоритм позволит выполнить нечеткое разделение данных по заданным тематическим классам и определить вероятность отнесения текстового документа к одному из них. Алгоритм латентного размещения Дирихле основан на вероятностной модели:

$$p(d, w) = \sum_{t \epsilon T} p(d) \cdot p(w|t) \cdot p(t|d), \tag{7}$$

где d – документ, w – слово, t – тематический класс, $p(d)$ – априорное распределение на множестве документов, $p(w|t)$ – условное распределение слова w в тематическом классе t, $p(t|d)$ – условное распределение тематического класса t в документе d.

Схема обобщенного алгоритма представлена на рис. 3.

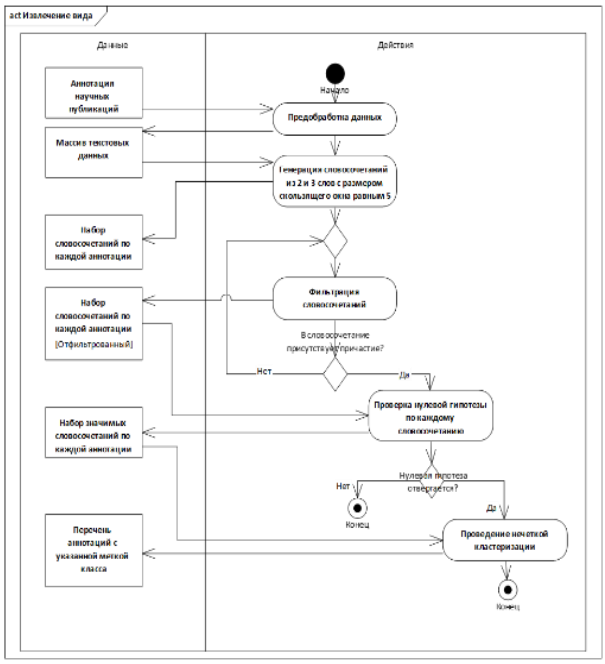

Рисунок 3. Схема алгоритма для извлечения вида научного результата (нотация UML)

Одним из внутренних способов оценки полученных результатов является оценка когерентности (согласованности), которая может быть рассчитана через поточечную взаимную информацию (pointwise mutual information) для наиболее часто встречаемых слов по каждому выделенному тематическому классу:

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i) \cdot P(w_j)}, \qquad (8)$$

где w_i и w_j – пара слов заданного тематического класса в порядке убывания по частоте.

В общем виде когерентность будет вычисляться следующим образом:

$$Coh = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j), \qquad N > 1, \qquad (9)$$

где w_i и w_j – пара слов заданного тематического класса в порядке убывания по частоте, N – количество наиболее часто встречающихся слов в тематическом классе.

В рамках развития научных рекомендательных систем предложенный метод был применен для определения вида научного результата на основе аннотаций научных публикаций, которые содержат основную суть работы. Источником данных выступили научные работы в количестве 5000 из рубрики «Информатика» базы данных ВИНИТИ.

В результате проведения предобработки и фильтрации данных, а также использовании фильтрации словосочетаний в соответствии с уровнем значимости для t-критерий Стьюдента и критерий Хи-квадрат α = 0,005, было получено 344 уникальных словосочетания. Аннотации научных статей описаны с помощью словосочетаний, где в среднем на каждую аннотацию приходится по 2,5 значимых словосочетания. Максимальное количество для одной аннотации составило 15 значимых словосочетаний.

Качество тематических моделей было оценено с помощью расчета когерентности для разного количества тематических классов, начиная от 2 и до 9 с шагом 1. Визуальное отражения значений когерентности для тематических классов (ТК) представлено на рис. 4. Максимальное значение когерентности соответствует четырем тематическим классам, но для трех из них значение когерентности также приемлемо. Построенную модель на основе четырех тематических классов можно перенести на двумерную плоскость для визуального анализа.

Используя дивергенцию Дженсена–Шеннона, можно визуализировать полученные тематические классы (рис. 5). Интерпретация полученных результатов позволяет говорить о том, что тематические классы «2», «3» и «4» расположены диаметрально противоположны. Тематически класс «1» имеет больше схожести с тематическим классом «2», что может объясняться близкой оценкой когерентности для трех и четырех тематических классов.

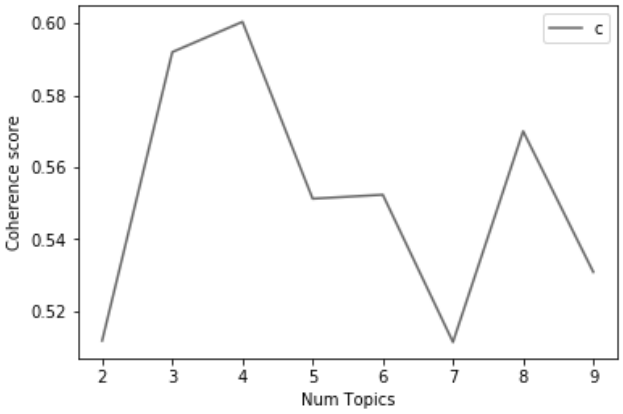

Рисунок 4. Когерентность

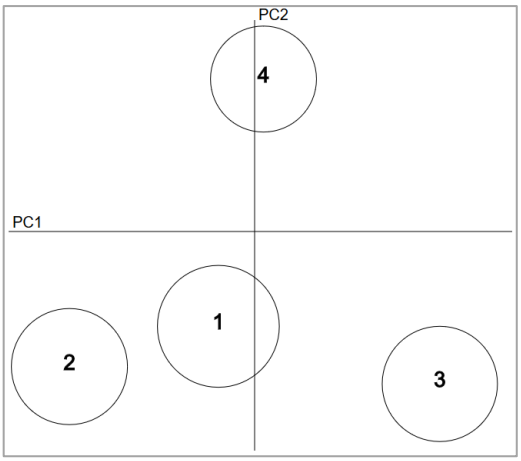

Рисунок 5. Распределение тематических классов

Каждый тематический класс выражается своим набором словосочетаний. В табл. 3 приведены часто встречаемые словосочетания по каждому полученному тематическому классу.

Таблица 3. Выделенные словосочетания для тематического класса «1», «2», «3» и «4»

№	*ТК 1*	*ТК 2*	*ТК 3*	*ТК 4*
1	Полученного_результат	Изложена_сведение	Быть_установлен	Предложен_методика
2	Приведен_пример	Представляющая_себя	Быть_разработан	Имеющая_отношение
3	Ограниченными_возможность	Быть_обсужден	Приведен_решение	Основа_положен
4	Разработан_алгоритм	Предложен_подход	Автоматизированная_система	Получившая_название
5	Получен_научный_результат	Предложен_метод	Быть_представлен	Проблема_возникавших
6	Сделан_вывод	Изложена_точка_зрение	Изложена_результат	Доклад_представлен
7	Быть_создавшая	Мочь_быть_использованными	Представлен_система	Даваться_развернутая
8	Представлен_использование	Представлен_анализ	Принят_решение	Обеспечивающая_ эффективный
9	Быть_проанализирован	Предпринята_попытка	Рассмотрен_возможность	Уделено_вопрос
10	Предложен_решение	Внимание_быть_уделено	Быть_выявлен	Обзор_посвящен

На основе анализа полученных значений и предложенного ранее подхода можно выполнить следующую интерпретацию словосочетаний по тематическим классам:

- ТК «1»: соответствует виду «Эксперимент», так как содержит большое количество словосочетаний, которые касаются полученных результатов и сделанных выводов.
- ТК «2»: соответствует виду «Гипотеза», так как содержит большое количество словосочетаний, касающихся возможных решений поставленных задач с помощью новых методов, основывающихся на предположениях авторов.
- ТК «3»: соответствует виду «Практика», так как содержатся словосочетания, обозначающие практический результат применимости разработанных методов и алгоритмов, реализованных в виде инструментальных средств или информационных систем.
- ТК «4»: соответствует виду «Основы», так как там содержатся основополагающие моменты рассматриваемых направлений, которые характеризуются соответствующим набором словосочетаний.

Имея полученную модель и зная интерпретацию каждого тематического класса, для любой текстовой аннотации можно выполнить классификацию и отнести ее к одному из представленных тематических классов. Помимо этого, стоит отметить, что, используя данную модель, можно получить вероятностное отношение аннотации к одному из тема-

тических классов. Таким образом, допускается вариант, когда одна аннотация может принадлежать одновременно нескольким тематическим классам.

Дальнейшие исследования показали, что пертинентность рекомендации, построенной с учетом вида научного результата, возрастает в среднем на 10–20 %, что связано с сокращением заведомо неинтересных пользователю документов.

В третьей главе предложен метод классификации данных на основе ансамблевого подхода. Зачастую наилучшие показатели качества достигают алгоритмы, использующие подход на основе комбинирования набора из простых базовых алгоритмов, где наилучший класс определяется путем «голосования» самими алгоритмами. Преимуществом такого подхода является возможность использовать «слабые» и нетребовательные к вычислительным ресурсам алгоритмы, которые можно выполнять параллельно, что приводит к сокращению временных затрат и повышению итогового качества работы. В рамках разработанного метода используется аналогичный подход, а сам процесс «голосования» будет строиться на основе использования энтропии как меры взвешивания.

В классическом понимании энтропия является показателем определенности или мерой однородности. Практический смысл при использовании энтропии применительно к алгоритмам классификации следующий: определить, насколько однородными являются предсказанные значения классов по отношению к истинным значениям в наборе данных

Оценка качества работы алгоритмов машинного обучения будет выполняться с помощью F-меры:

$$F = 2 \cdot \frac{\text{Точность} \cdot \text{Полнота}}{\text{Точность} + \text{Полнота}}, \tag{10}$$

где *точность* – доля объектов, названных классификатором положительными и при этом действительно являющимися положительными; *полнота* – доля объектов положительного класса из всех объектов положительного класса, которая была найдена алгоритмом.

Для построения метода классификации слабоструктурированных данных с использованием энтропии должны быть выполнены следующие шаги.

Шаг 1. Предобработка данных и формирование выборки.

Формирование признакового пространства и преобразование исходного массива текстовых данных (D) в цифровой вид происходит при помощи представления текста в виде матрицы слов. В данном случае Bag of words используется как наиболее часто используемый вид представления слов, но это не ограничивает возможность использования TF-IDF, n-грамм и т.д. Предобработка данных выполняется путем удаления стоп-слов и символов пунктуации, приведения символов к нижнему регистру и слов к нормальной форме. Далее все слова обрабатываются в нормализованном виде. Массив данных будет представлять собой метку класса (C) и набор признаков, описывающий объект.

Разбиение общего массива данных на выборки происходит случайным образом, но в заранее заданных пропорциях: 75% массива данных отводится на обучающую выборку ($D_{об}$), 25% на тестовую выборку (D_T). Распределение объектов по классам в исходном массиве данных является равномерным. Дополнительно от обучающей выборки ($D_{об}$) выделяется в пропорции 50:50 на выборку для дополнительного обучения ($D_{до}$). Таким образом исходный массив данных будет иметь вид $D = \{D_{об}, D_{до}, D_{т}\}$.

Шаг 2. Расчет энтропии.

На основе выборки для обучения ($D_{об}$) происходит обучение базового набора алгоритмов $A = \{a_1, a_2, \ldots, a_b, \ldots, a_n\}$ и формирование обученных моделей по каждому алгоритму $F = \{f_{a1}, f_{a2}, \ldots, f_{ab}, \ldots, f_{an}\}$. На основе полученных предсказаний обученных моделей алгоритмов (F) строятся матрицы неточностей для каждого из алгоритмов (A):

$$\begin{pmatrix} a_b x_{11} & a_b x_{12} & \dots & a_b x_{1C} \\ a_b x_{21} & a_b x_{22} & \dots & a_b x_{2C} \\ \dots & \dots & a_b x_{ij} & \dots \\ a_b x_{C1} & a_b x_{C2} & \dots & a_b x_{CC} \end{pmatrix}, \quad (11)$$

где $a_b x_{ij}$ – количество объектов принадлежащий классу i, но классифицированных алгоритмом a_b как класс j; c – количество классов.

На основе представленной матрицы неточности считается энтропия для каждого класса, которому соответствует столбец из матрицы (11). Формула для расчета энтропии имеет вид:

$$H(a_b x_{.j}) = -\sum_{i=1}^{c} p(a_b x_{ij}) log_2 p(a_b x_{ij}), \quad (12)$$

где $x_{.j}$ – j-й столбец со значениями класса матрицы неточностей; a_b – алгоритм классификации; c – количество классов; $p(a_b x_{ij})$ – вероятность встретить элемент в j-ом столбце.

Рассчитанные значения энтропии для каждого класса представлены в матрице

$$\begin{pmatrix} ha_1 c_1 & ha_1 c_2 & \dots & ha_1 c_k \\ ha_2 c_1 & ha_2 c_2 & \dots & ha_2 c_k \\ \vdots & \vdots & ha_i c_j & \vdots \\ ha_n c_1 & ha_n c_2 & \dots & ha_n c_k \end{pmatrix}, \quad (13)$$

где a_n – перечень используемых алгоритмов, c_k – перечень классов, $ha_i c_j$ – энтропия по алгоритму a класса c.

Шаг 3. Выбор финального результата.

На основе обученных моделей алгоритмов (F) выполняется расчет предсказаний класса (C) по каждому объекту из тестовой выборки с определенной вероятностью (P). Для каждого объекта происходит расчет показателя значимости (H') для определения принадлежности к классу следующим образом:

$$H'(x) = \left(1 - 0{,}5 \cdot \frac{h(x_i) - h(x)_{min}}{h(x)_{max} - h(x)_{min}}\right) \cdot p(x_i) + p(x_i), \quad (14)$$

где x_i – входной объект, p – вероятность на отрезке [0,1], h – энтропия.

Получив новый показатель значимости по каждому классу для каждого объекта выбор итогового класса происходит путем определения максимального количества голосующих алгоритмов за конкретный класс. Если выбрать победителя большинством не удалось, то победителем становится класс с наибольшим показателем значимости. Обобщенная схема построения алгоритма классификации представлена на рис. 6 и 7.

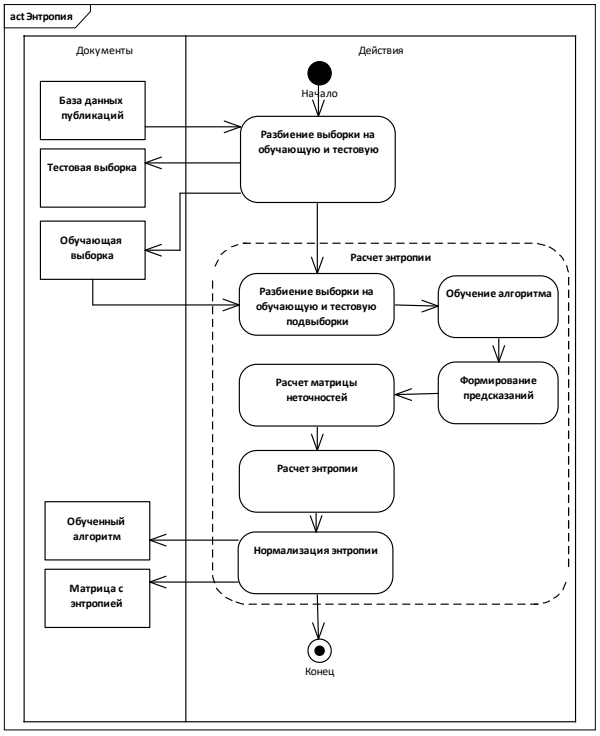

Рисунок 6. Схема расчета энтропии для построения алгоритма классификации (нотация UML)

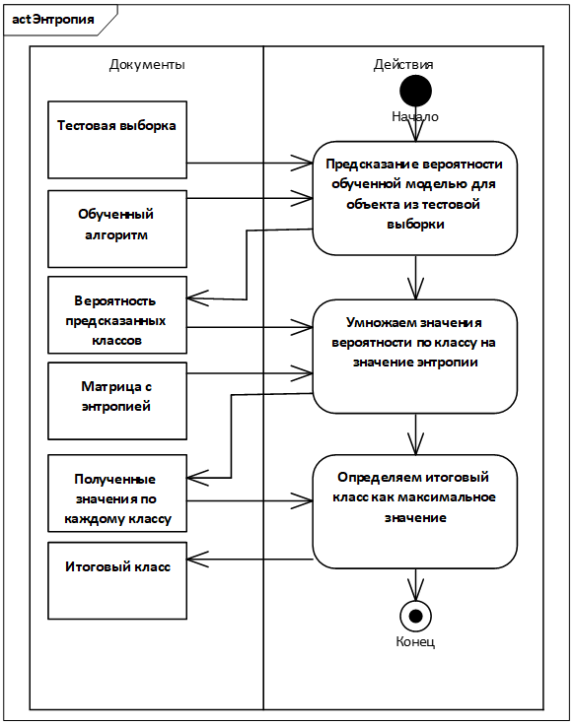

Рисунок 7. Схема применения энтропии при построении алгоритма классификации (нотация UML)

Апробация предложенного алгоритма выполнялась на основе научных публикаций базы данных ВИНИТИ в количестве 5000 объектов. Было подготовлено три отдельные выборки, включающие в себя 5, 10 и 15 научных направлений соответственно. Для сравнительного анализа в представленном исследовании в качестве базовых алгоритмов были использованы алгоритмы с максимально различной природой происхождения и способом определения верного класса: RandomForest (RF), Stochastic gradient descent (SGD), Naive Bayes classifier for multinomial models (MNB), Logistic Regression (LR), Multi-layer Perceptron classifier (NN), а также базовых ансамблей Bagging и AdaBoost.

Применив к представленным алгоритмам меру энтропии, получим следующие значения точности для 5, 10 и 15 классов (рис.8), где для расчета показателя точности использовалась метрика F-мера.

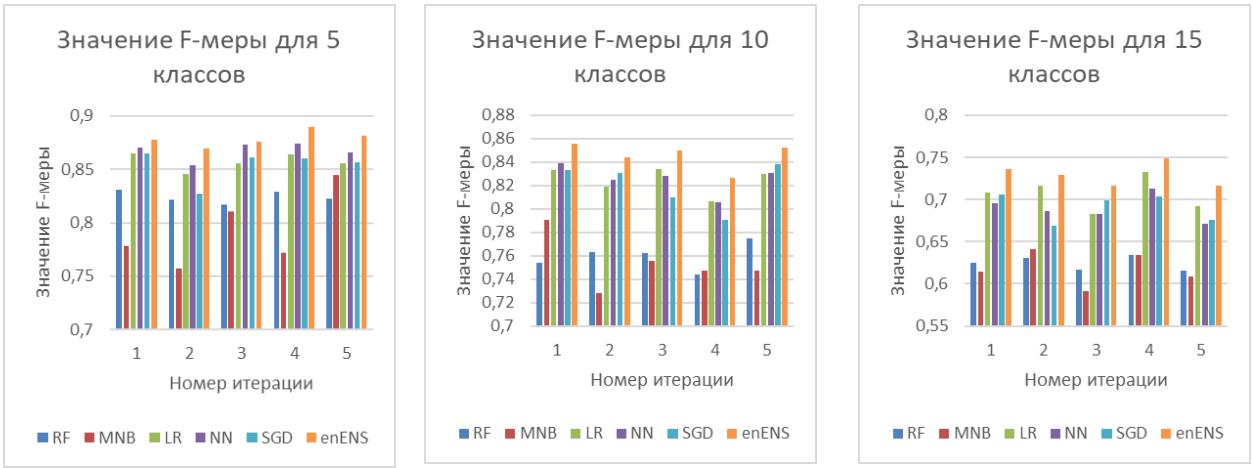

Рисунок 8. Качество работы базовых алгоритмов при 5, 10 и 15 классах соответственно

На основе представленных результатов можно говорить о высокой стабильности работы ансамбля при проведенных пяти итерациях (кросс-валидации). В то время как базо-

вые алгоритмы имеют большой разброс показателя F-меры, разброс ансамбля существенно ниже, но по-прежнему связан с качеством базовых алгоритмов и коррелирует с ними. Так, можно отметить, что прирост F-меры по сравнению с базовыми алгоритмами достигает в среднем 4–7 %.

Сравнительный анализ работы разработанного ансамбля и базовых ансамблей Bagging и AdaBoost показал, что разработанный ансамбль позволяет существенно повысить качество классификации в среднем от 6–8 % до 16–20 % в зависимости от числа классов.

Предложенный ансамбль алгоритмов машинного обучения на базе энтропии был реализован в рамках автоматизированной системы многомерной классификации данных (свидетельство о государственной регистрации программы для ЭВМ № 2017612002 от 14.02.2017). Разработанная программа для ЭВМ содержит визуальный интерфейс для управления реализованными алгоритмами (рис. 9).

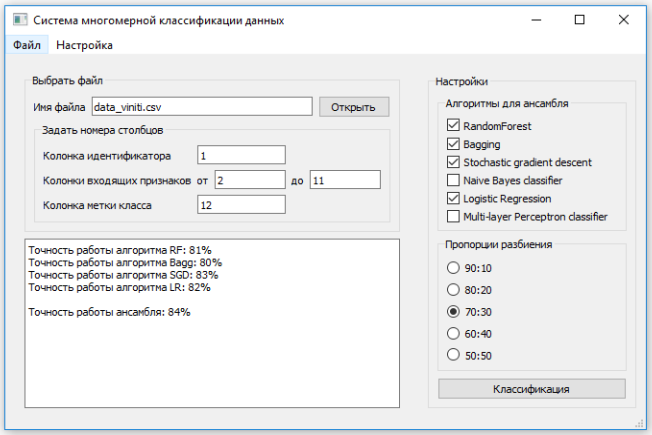

Рисунок 9. Главное окно автоматизированной системы многомерной классификации

Разработанная программа для ЭВМ позволяет на основе набора входных научных текстовых данных выполнить операцию классификации и определить точность полученной модели. Данная программа может рассматриваться как целостное решение, которое может использоваться в качестве рабочего инструмента исследователя или аналитика, когда необходимо в кратчайшие сроки провести классификацию данных или определить наилучший алгоритм для конкретного набора данных.

Автоматизированная система позволяет с помощью пользовательского интерфейса настраивать набор базовых алгоритмов для использования в ансамбле, также задавать пропорции разбиения обучающей и тестовой выборки. Система реализована с использованием языка программирования Python. Использовались библиотеки scikit-learn, Numpy, Scipy, gensim, NLTK, pyQT5. Система состоит из трех подсистем и содержит более 5000 строк кода.

Автоматизированная система многомерной классификации данных использовалась при проведении исследований по проекту РФФИ № 15-07-08742 «Принципы создания алгоритмического обеспечения для многомерной классификации на примере анализа научных направлений» (2015–2017 гг.).

Кроме того, в рамках данного проекта была разработана и зарегистрирована еще одна программа для ЭВМ № 2015610216, предназначенная для анализа поведения интернет-пользователей в форме оценки тональности информационных ресурсов (на примере реализации проектов в атомной промышленности). Научные и практические результаты диссертационного исследования вошли в состав пяти баз данных научных публикаций и учебно-методических материалов и используются в учебном процессе НИЯУ МИФИ в

рамках научно-практического семинара для магистров «Информационные технологии в науке и образовании», что подтверждается соответствующим актом об использовании.

В четвертой главе представлено программно-техническое решение повышения пертинентности информации в научных и аналитических рекомендательных системах на базе алгоритма классификации данных с использованием энтропии, который был апробирован в рамках проекта № 2014-14-576-0146 ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2020 годы» (свидетельство о государственной регистрации программы для ЭВМ № 2015662714 от 30.11.2015). Разработанное программно-техническое решение предназначено для прогнозирования поведения пользователя в отношении объекта информационного поиска и формирования рекомендаций для объектов, с которыми он еще не встречался.

Метод, повышающий пертинентность информации, реализуется экспериментальным образцом программного комплекса повышения пертинентности информации. Программа позволяет выполнить сбор и хранение поведенческих данных, установить корреляционные зависимости между пользовательскими профилями и сформировать информационные предложения на основе анализа поведенческих данных. На основе предложенных методов и алгоритмов были реализованы модули: «2. Анализ пользователя», «3. Анализ информационной единицы», «4. Вывод информационной единицы», принцип взаимодействия которых представлен на рис. 10.

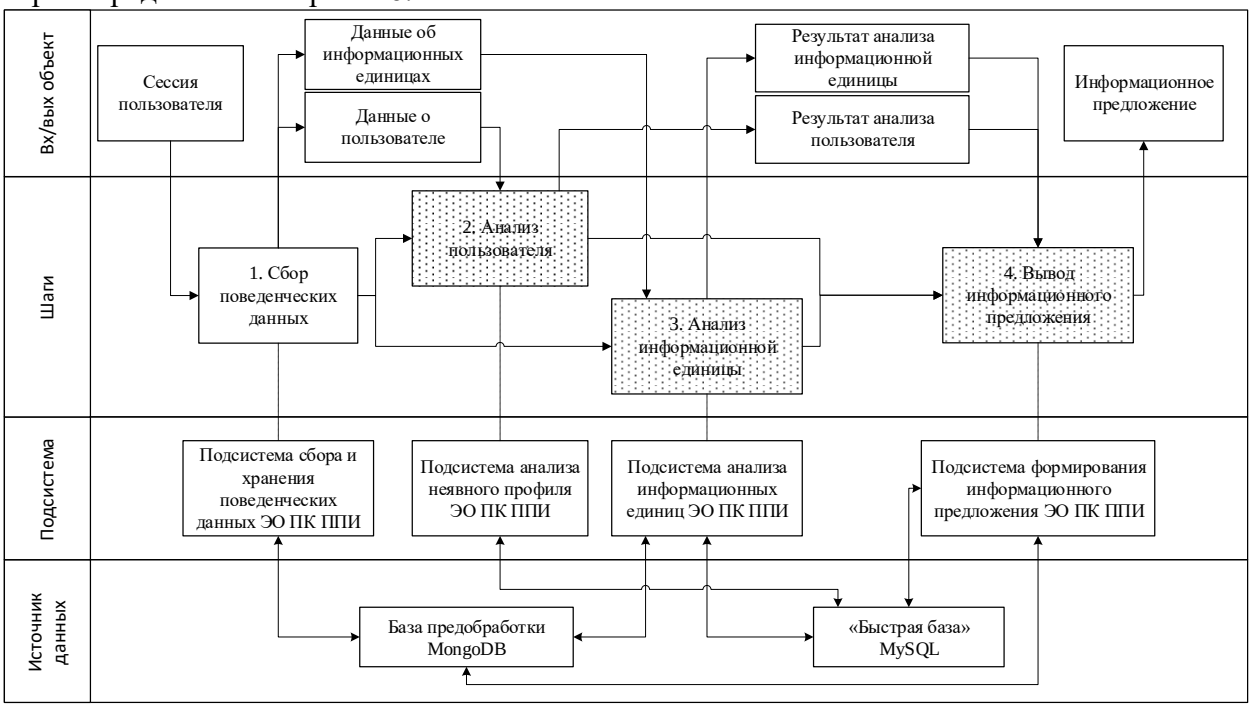

Рисунок 10. Взаимодействие подсистем в программно-техническом решении повышения пертинентности

Данная система позволяет пользователю получать персонифицированные рекомендации, которые строятся на основе целого набора данных. Научные и аналитические рекомендательные системы имеют свою специфику, а пользователь таких систем взаимодействует с информационными единицами – научными результатами. При работе с научной рекомендательной системой пользователь совершает различные действия, которые фиксируются в логе системы (создаёт, читает, присваивает оценку, скачивает и т.д.). Каждое действие имеет также пространственно-временные характеристики (где и когда произведено действие). На основе этих данных система пытается предсказать информационную потребность у пользователя в конкретном научном результате.

Повышение эффективности формируемых рекомендательных наборов и сокращение временных затрат при информационном поиске являются основными достоинствами рекомендательных систем.

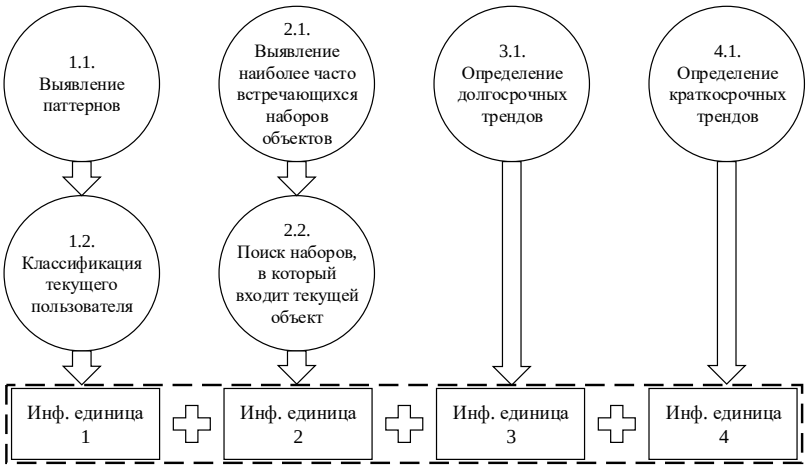

Рисунок 11. Схема формирования кортежа информационных предложений

Для вывода информационного предложения был разработан алгоритм формирования кортежа информационных предложений пользователю. Схема формирования финального кортежа представлена на рис. 11.

Кортеж состоит из информационных единиц (ИЕ) общей длиной 7 ± 2 объектов, чтобы учитывать особенности человеческого восприятия и включает следующие части:

1. ИЕ, полученные по пользователям, похожим на текущего;
2. ИЕ, полученные из наиболее часто встречающихся наборов;
3. ИЕ, относящиеся к долговременным трендам («топы»);
4. ИЕ, относящиеся к кратковременным трендам, актуальным («тренды»).

Первые две части кортежа относятся к восстанавливаемой информационной потребности и предполагаются вносящими наибольший вклад в рекомендованный набор (порядка 60-70%). Последние две части кортежа относятся к формируемой информационной потребности и должны занимать 30-40% от объёма кортежа.

Система разрабатывалась с использованием языка программирования Ruby и платформы создания приложений RubyonRails. В качестве СУБД были использованы MongoDB и MySQL. Система состоит из четырех подсистем и содержит около 40 функций. Для управления разработанной системой было предусмотрено разграничение прав пользователей по ролям: «Оператор» и «Администратор». Управление системой происходит через набор конфигурационных файлов, доступ к которым реализован с помощью ряда скриптов, а взаимодействие «Оператора» со скриптом происходит через консоль.

Работа пользователя с правами доступа «Администратор» включает в себя функции по генерации данных для тестирования системы, которые позволят проверить все функции работы системы и убедиться в корректности их работы.

Представленное программно-техническое решение было использовано в рамках разработки научной рекомендательной системы для информационной системы Международный конгресс конференций «Информационные технологии в образовании», что подтверждается соответствующим актом о внедрении.

В пятой главе проводится оценка эффективности использования предложенных алгоритмов применительно к работе поискового модуля информационной системы Международного конгресса конференций «Информационные технологии в образовании». Информационная система содержит архив материалов с 1994 года. Согласно данным поисковых систем Яндекса и Google количество проиндексированных страниц составляет 27000 и 23000 соответственно. Посещаемость интернет-портала составляет в среднем около 1000-1200 посетителей в неделю.

Большинство посетителей выполняют только просмотр одной страницы, а затем покидают интернет-портал. Однако часть пользователей выполняет вплоть до 60 переходов, но процент пользователей с таким высоким количеством просмотренных страниц составляет менее 1 %. Согласно сервису Яндекс.Метрика глубина просмотра страниц и среднее время, проведенное на интернет-портале конгресса конференций, имеет следующую зависимость:

- для 2-3 просмотров – время составляет 7 минут 4 секунды;
- для 4-7 просмотров – время составляет 10 минут 24 секунды;
- для 8-15 просмотров – время составляет 13 минут 15 секунд;
- для 16-31 просмотров – время составляет 24 минуты 35 секунд.

В рамках работы с НРС выделяются два ключевых элемента: пользователь и объект (веб-страница, содержащая научную статью или тезисы конференций). Базовый профиль пользователя НРС может состоять из двух составных частей: статической и динамической. Данные части профиля соответствуют составленной онтологии научной деятельности, но их использование в рамках научных информационных систем может быть лишь частичным. К статической части относятся параметры: образование, ученая степень, перечень научных работ, интересы и прочее. Динамическая часть профиля формируется путем анализа работы пользователя на странице научной информационной системы и совершении им каких-то действий по отношению к просматриваемому на странице объекту. Объект содержит в базовом профиле параметры: тип, название, ключевые слова, авторы, дата и прочее.

Формирование профиля пользователя $U = <Ds_u, T_u>$ на основе просмотренного им набора объектов позволяет расширить базовый профиль. Профиль пользователя дополнительно будет включать параметр $Ds_u = \{c_1, c_2, \dots c_n\}$, который представляет собой множество классов, к которым были отнесены просмотренные пользователем объекты, и называется «Рубрикатор». Параметр $T_u = (t_1, t_2, t_3, t_4)$ определяет вес каждого вида научного результата, к которым были отнесены просматриваемые объекты и называется «Вид».

Неявный профиль пользователя (без учета основных базовых параметров профиля) может иметь следующий вид: $u_i = <[4,6], \{0.49;0.31;0.12;0.08\}>$. Неявный профиль объекта будет уникален, но при этом также будет состоять из параметра «Рубрикатор» и параметра «Вид», например, может иметь вид: $d_i = <[6], \{0.72;0.21;0.05;0.02\}>$.

Пользователь с профилем u_1 выполняя на сайте НРС поисковый запрос получает релевантный набор документов, состоящий из объектов, каждый из которых имеет свой профиль. Однако, поисковый модуль не учитывает профиль пользователя. Использование параметра «Вид» сводится к тому, что необходимо выполнить поиск наиболее похожих объектов среди поисковой выдачи на значения из профиля пользователя. Использование параметра «Рубрикатор» сводится к тому, чтобы найти пересечение между данным значением из профиля пользователя и значением из профиля объекта.

На представленном примере (рис.12) показана трансформация поисковой выдачи, где упорядоченные по номерам объекты сначала были отсортированы в соответствии со значением из поля «Вид», а затем объект № 4 был перенесен в конец списка, так как его рубрикатор не содержится в профиле пользователя.

Модель информационного поиска		
	Параметры	
№	Вид научного результата	Рубрикатор
1	{0.39, 0.05, 0.21, 0.35}	4
2	{0.14, 0.53, 0.11, 0.22}	6
3	{0.34, 0.43, 0.15, 0.08}	6
4	{0.29, 0.13, 0.42, 0.16}	11
5	{0.72, 0.21, 0.05, 0.02}	6

Модель построения рекомендаций		
	+ производные параметры	
№	Схожесть по "Виду"	Рубрикатор
3	0.19	6
5	0.26	6
1	0.39	4
2	0.43	6
4	0.41	11

Рисунок 12. Принцип ранжирования поисковой выдачи в зависимости от профиля пользователя

Представленные параметры, а также остальные параметры профиля пользователя и объекта могут быть применены к поисковой системе и учитываться при формировании

поисковой выдачи. Для оценки эффективности полученных параметров «Вид» и «Рубрикатор» с помощью разработанных алгоритмов был проведен расчет пертинентности по нескольким сценариям.

Первый сценарий предполагает использование в рамках НРС классического полнотекстового поискового движка Sphinx, где в качестве поисковых полей используется название, аннотация и ключевые слова научных статей. Данный сценарий направлен на воссоздание максимально реальных условий работы платформы Bitrix.

Второй сценарий включает в себя использование вида научного результата как дополнительного параметра «Вид», учитываемого при ранжировании поисковой выдачи. Похожесть профиля пользователя и объектов поисковой выдачи осуществляется за счет использования метрики Евклидово расстояния. На основе полученных значений выполняется сортировка по возрастанию и формируется поисковое предложение.

Третий сценарий включает в себя использование рубрик научных статей – параметр «Рубрикатор». Статья, рубрика которой не представлена в профиле пользователя, переносится в поисковой выдаче в конец списка с сохранением последовательности.

Четвертый сценарий предполагает одновременное использование дополнительных параметров из двух сценариев.

В рамках предлагаемого подхода выполняется трансформация поисковой выдачи. Принцип работы предлагаемого решения состоит в следующем. Используя поисковый модуль на странице НРС, пользователь вводит интересующий его запрос. На основе алгоритмов ранжирования из встроенного в НРС поискового движка Sphinx пользователь получает набор поисковых предложений.

Для оценки эффективности работы описанных сценариев были привлечены эксперты, которые оценивали соответствие перечня полученных рекомендаций и их потребности. В состав экспертной комиссии вошли разные группы исследователей: 19 – сложившиеся ученые, 9 – молодые ученые, 23 – аспиранты, 25 – магистры. Для проведения экспертизы использовалась диалоговая панель научной рекомендательной системы, состоящая из базы данных с набором тезисов конференций и поискового движка Sphinx, формирующим поисковую выдачу. На основе введенного запроса диалоговая панель предлагала наиболее релевантный набор документов, которые эксперт оценивает по принципу «соответствует | не соответствует». Для формирования первичного профиля пользователя выполняется несколько итераций аналогичным способом.

В первом сценарии было выявлено, что на основе стандартного поискового механизма уровень удовлетворенности пользователя без учета его профиля в среднем составлял от 40 до 60 %, что означает, что количество предложенных документов примерно наполовину соответствовали исходным потребностям.

Во втором сценарии уровень удовлетворенности пользователя возрастал в среднем на 10–20 %, что связано с сокращением заведомо неинтересных пользователю документов в поисковой выдаче.

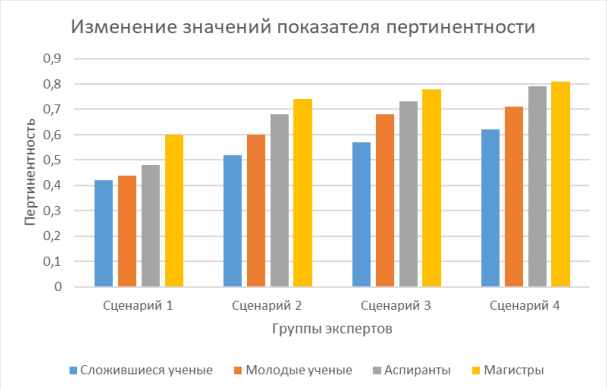

Рисунок 13. Изменение значений показателя пертинентности в зависимости от использования предложенных методов

В третьем сценарии уровень удовлетворенности пользователя возрастал в среднем на 15–25 % по той причине, что была скорректирована направленность предлагаемых документов. В четвертом сценарии совместное использование двух алгоритмов позволило повысить уровень удовлетворенности пользователя в среднем на 20–30 %.

На основе проведенного эксперимента (рис. 13) можно сделать вывод, что использование обратной связи от пользователя, а также вспомогательной информации о профиле пользователя и неявных признаках

просматриваемых документов позволяют существенно увеличить эффективности итогового набора поискового предложения и повысить удовлетворенность пользователя.

В **заключение** подведены итоги и приведены основные результаты, полученные в рамках диссертационной работы.

В **приложении** представлены копии свидетельств о государственной регистрации программ для ЭВМ и акты о внедрении результатов диссертационного исследования.

Выводы и результаты

1. Проведено исследование современных подходов к построению рекомендательных систем и существующих методов машинного обучения. Показано, что предпочтительно использовать ансамбли (комитеты) алгоритмов, дающие следующие преимущества:
   - использовать «слабые» и нетребовательные к вычислительным ресурсам алгоритмы, которые можно выполнять параллельно, что приводит к сокращению временных затрат и повышению итогового качества работы;
   - «обучать» разные алгоритмы на разных обучающих выборках для уменьшения несоответствия между структурами их данных и генеральной совокупности.
2. Предложен параметрический подход, на его основе разработаны метод и алгоритмы обогащения признакового пространства для слабоструктурированных текстовых научных данных и выделен новый параметр – вид научного результат. Проведенные экспериментальные исследования показали, что использование данного параметра в научных рекомендательных системах позволяет повысить пертинентность рекомендаций в среднем на 10–20 %.
3. Разработаны и исследованы ансамблевые метод и алгоритмы для решения задачи классификации при обработке слабоструктурированных текстовых данных. Ансамблевый метод на основе энтропии позволяет повысить точность и стабильность классификации за счет использования внутри ансамбля различных базовых алгоритмов машинного обучения. Проведенные эксперименты показывают, что пертинентность рекомендаций возрастает в среднем на 15–25 %.
4. Получены экспериментальные результаты применения предложенных методов в научных рекомендательных системах. Проведенный эксперимент показал, что одновременное использование двух предложенных методов позволяет повысить пертинентность рекомендаций в среднем на 20–30 %.
5. Проведена апробация предложенных методов и алгоритмов в виде разработанных программных средств повышения пертинентности информации для информационной системы Международный конгресс конференций «Информационные технологии в образовании» (подтверждено соответствующим актом о внедрении).
6. Научные и практические результаты диссертационного исследования вошли в состав пяти разработанных баз данных учебно-методических материалов и научных публикаций, а также были использованы в рамках научной и учебной деятельности в НИЯУ МИФИ, что подтверждается соответствующим актом.

Основные публикации по теме диссертации

I. Публикации, представленные в международных базах цитирования Scopus и Web of Science:

1. Kuznetsov, I. A. Scientific and Educational Recommender Systems / A. I. Guseva, V. S. Kireev, P. V. Bochkarev, I. A. Kuznetsov, S. A. Philippov // 2017 Information Technologies in Education of the XXI Century (ITE-XXI), AIP Conf. Proc. – 2017. – Volume 1797. – Issue 1. DOI: 10.1063/1.4972422 (Web of Science и Scopus).
2. Kuznetsov, I. A. The use of entropy measure for higher quality machine learning algorithms in text data processing / A. I. Guseva, I. A. Kuznetsov // Proceedings - 2017 5th Interna-

tional Conference on Future Internet of Things and Cloud Workshops (FiCloudW). – 2017. P. 47-52 (Web of Science и Scopus)

3. Kuznetsov, I. A. Development of algorithms ensemble in case of the solution of the task of statistical classification in recommender systems / V. S. Kireev, I. A. Kuznetsov // International Journal of Applied Engineering Research. – 2016. – Volume 11. – № 9. P. 6613-6618 (Scopus).

4. Kuznetsov, I. A. Development of an ensemble of classification algorithms using the entropy quality measure for solving the problem of behavioral scoring / I. A. Kuznetsov, V. S. Kireev // CEUR Workshop Proceedings. – 2016. – Volume 1752. P. 37-43 (Scopus).

5. Kuznetsov, I. A. A method for obtaining a type of scientific result from the text of an article abstract to improve the quality of recommender systems / Igor A. Kuznetsov, Anna I. Guseva // Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2019. – 2019. – P. 1888-1891 DOI: 10.1109/EIConRus.2019.8656806 (Web of Science и Scopus).

II. Публикации в журналах, включенные в перечень периодических изданий ВАК Российской Федерации:

6. Кузнецов И. А. Метод автоматизированной классификации научных статей по типу результата в научных аналитических системах / И. А. Кузнецов // Современные наукоемкие технологии. – 2018. – № 2. – С. 59-63.

7. Кузнецов И. А. Разработка модели пользователя научных сетей на основе концепции OPENSCIENCE / В. С. Киреев, И. А. Кузнецов, П. В. Бочкарёв, А. И. Гусева, С. А. Филиппов // Фундаментальные исследования. – 2015. – № 12, часть 5. – С. 907-913.

8. Кузнецов И. А. Исследование алгоритмов многомерной классификации научных данных / А. И. Гусева, В. С. Киреев, И. А. Кузнецов, П. В. Бочкарёв // Фундаментальные исследования. – 2015. – № 11, часть 5. – С. 868-874.

9. Кузнецов И. А. Модель научного направления на основе интеграции объектно-ориентированного, наукометрического и экспертного подходов / П. В. Бочкарёв, А. И. Гусева, В. С. Киреев, И. А. Кузнецов, С. А. Филиппов // Фундаментальные исследования. – 2015. – № 12, часть 6. – С. 1095-1102.

10. Кузнецов И. А. Подходы и их реализация при анализе данных общественного мнения о развитии атомного промышленного комплекса / И. А. Кузнецов, М. В. Коптелов // Научное обозрение. – 2014. – № 6. – С. 112-114.

III. Другие публикации:

11. Кузнецов И. А. Обзор современных архитектур хранения и обработки больших данных / А. И. Гусева, В. С. Киреев, П. В. Бочкарёв, И. А. Кузнецов // Цифровые платформы управления жизненным циклом комплексных систем. Под общ. ред. д.э.н., проф. В. А. Тупчиенко. – М.: Издательство «Научный консультант». – 2018. – С. 125-158.

12. Кузнецов И. А. Применение искусственного интеллекта для персонализации потребительских услуг / В. С. Киреев, И. А. Кузнецов / Цифровые платформы управления жизненным циклом комплексных систем. Под общ. ред. д.э.н., проф. В. А. Тупчиенко. – М.: Издательство «Научный консультант». – 2018. – С. 181-204.

13. Кузнецов И. А. Формирование и применение неявного профиля пользователя в научных аналитических системах // Новые информационные технологии в научных исследованиях: материалы XXIII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов: в 2 томах. Рязань, РГРТУ, 12-14 декабря 2018 г.: сб. докладов, том 1. – С. 176-178.

14. Кузнецов И. А. Реализация модуля персонифицированных рекомендаций в системе многомерной классификации научных данных // В сборнике: Информатика, управление и системный анализ. Труды V Всероссийской научной конференции молодых ученых с международным участием. – 2018. – С. 247-254.

15. Кузнецов И. А. Автоматизация процесса формирования онтологии на основе классов пользователей в научных рекомендательных системах // Новые информационные технологии в научных исследованиях: материалы XXII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов, Рязань, РГРТУ, 15-17 ноября 2017 г.: сб. докладов. – С. 34-36.

16. Кузнецов И. А. Повышение пертинентности информации в научных рекомендательных системах с использованием ансамблей алгоритмов машинного обучения / И. А. Кузнецов // Международный научно-технический семинар «Современные технологии в задачах управления, автоматики и обработки информации» Алушта, Республика Крым, Российская Федерация, 14-20 сентября 2016 г.: сб. докладов. – С.160-161.

17. Кузнецов И.А. Предобработка данных, выбор и формирование признаков при анализе данных [Электронный ресурс] / И.А. Кузнецов // 19-я Международная телекоммуникационная конференция молодых ученых и студентов «Молодежь и наука» (Москва, 1 окт. 10 дек. 2015 г.): материалы конференции. URL:http://mn.mephi.ru/articles/1538 (дата обращения 21.01.2015)

IV. Свидетельство о регистрации программ для ЭВМ:

18. Кузнецов И. А. Автоматизированная система многомерной классификации данных с использованием ансамбля алгоритмов машинного обучения на базе энтропии / Кузнецов И. А., Гусева А. И., Киреев В. С., Гудков П. Г., Филиппов С. А.// Свидетельство о государственной регистрации программы для ЭВМ РФ № 2017612002. Правообладатель НИЯУ МИФИ (Россия). 2017. Бюл. № 2.

19. Кузнецов И. А. Программа классификации неявных профилей пользователей научных и аналитических рекомендательных систем на основе комбинированного правила голосования/ Гусева А. И., Киреев В. С., Кузнецов И. А., Бочкарёв П. В., Коптелов М. В. // Свидетельство о государственной регистрации программы для ЭВМ РФ № 2015662714. Правообладатель Общество с ограниченной ответственностью «Социальные конференционные технологии» (Россия). 2015. Бюл. № 12.

20. Кузнецов И. А. База данных публикаций по тематике «Методы обработки больших данных (BigData) в научных и социальных сетях, включая методы классификации с учителем и без»/ Гусева А. И., Киреев В. С., Филиппов С. А., Бочкарёв П. В., Кузнецов И. А., Гаврось Л. В., Гудков П. Г.// Свидетельство о государственной регистрации базы данных РФ № 2015621524. Правообладатель НИЯУ МИФИ (Россия). 2015. Бюл. № 11.

21. Кузнецов И. А. База данных публикаций по тематике «Исследование поведенческих профилей пользователей научных и социальных сетей»/ Гусева А. И., Киреев В. С., Филиппов С. А., Бочкарёв П. В., Кузнецов И. А., Сомова О. А.// Свидетельство о государственной регистрации базы данных РФ № 2015621512. Правообладатель НИЯУ МИФИ (Россия). 2015. Бюл. № 10.

22. Кузнецов И.А. База данных публикаций по тематике «Дифференциация поведенческих профилей пользователей научных и социальных сетей с учетом фактора ботов» / Гусева А. И., Киреев В. С., Филиппов С. А., Бочкарёв П. В., Кузнецов И. А., Кузьмин Д. С.// Свидетельство о государственной регистрации базы данных РФ № 2015621457. Правообладатель НИЯУ МИФИ (Россия). 2015. Бюл. № 10.

23. Кузнецов И. А. База данных учебно-методических материалов по направлению подготовки «Бизнес-информатика»/ Гусева А. И., Тихомирова А. Н., Коровкина Л. Н., Киреев В. С., Цыганов А. А., Филиппов С. А., Матросова Е. В., Кузнецов И. А., Кирьяков И. Л., Маслий Н. П.// Свидетельство о государственной регистрации базы данных РФ № 2015620231. Правообладатель НИЯУ МИФИ (Россия). 2015. Бюл. № 3.

24. Кузнецов И. А. База данных учебно-методических материалов по направлению подготовки «Прикладная информатика»/ Гусева А. И., Золотухина Е. Б., Киреев В. С., Путилов А. В., Тихомирова А. Н., Филиппов С. А., Шнырев С. Л., Кузнецов И. А., Гриб И.

И., Маслий Н. П. // Свидетельство о государственной регистрации базы данных РФ № 2015620233. Правообладатель НИЯУ МИФИ (Россия). 2015. Бюл. № 3.

25. Кузнецов И. А. Программа для ЭВМ «Автоматизированная система оценки тональности информационных ресурсов о реализации проектов в атомной промышленности» / Кузнецов И. А., Коптелов М. В. // Свидетельство о государственной регистрации программы для ЭВМ РФ № 2015610216. Правообладатель НИЯУ МИФИ (Россия). 2016. Бюл. № 2.