На правах рукописи



СЕЛИВАНОВА Ирина Вячеславовна

# МЕТОДЫ ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ НАУЧНЫХ ТЕКСТОВ НА ОСНОВЕ ТЕОРЕТИКО-ИНФОРМАЦИОННОГО ПОДХОДА

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Новосибирск – 2020

Работа выполнена в Федеральном государственном бюджетном учреждении науки Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук (ГПНТБ СО РАН)

**Научный руководитель:**
*Гуськов Андрей Евгеньевич*
кандидат технических наук, директор, ГПНТБ СО РАН

**Официальные оппоненты:**
*Барахнин Владимир Борисович*
доктор технических наук, доцент, ведущий научный сотрудник лаборатории цифровых двойников и анализа больших данных, Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр информационных и вычислительных технологий»
*Загорулько Юрий Алексеевич*
кандидат технических наук, заведующий лабораторией искусственного интеллекта, Федеральное государственное бюджетное учреждение науки Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук

**Ведущая организация:**
Федеральное государственное бюджетное учреждение науки Библиотека по естественным наукам Российской академии наук

Защита состоится « 4 » марта 2021 года в 16 ч. 00 мин. на заседании диссертационного совета Д 219.005.02 в Федеральном государственном бюджетном образовательном учреждении высшего образования «Сибирский государственный университет телекоммуникаций и информатики» (СибГУТИ) по адресу 630102, г. Новосибирск, ул. Кирова, 86, ауд. 625.

С диссертацией можно ознакомиться в библиотеке СибГУТИ и на сайте: https://sibsutis.ru/science/postgraduate/dis_sovets/

Автореферат разослан «  »  г.

Ученый секретарь диссертационного совета
к.т.н., доцент  Нечта И.В.

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**Актуальность проблемы.** В последние десятилетия в связи с экспоненциальным ростом количества информации проблема классификации текстов, то есть разделения объектов на заранее заданное множество классов, становится особенно актуальной. Она возникает для текстов различного происхождения (художественных, поэтических, технических текстов или публикаций в СМИ), для объемных или коротких сообщений (СМС, твиты и комментарии в социальных сетях), с разными целями (анализ эмоций, определение авторства, тематическая кластеризация). Особую важность задача классификации играет в научной сфере, где в каждой дисциплине ежегодно добавляются десятки тысяч монографий, статей, препринтов и других видов публикаций. Эффективная обработка таких массивов, качество поиска в них материалов, релевантных конкретному направлению исследований, требуют точного соотнесения каждой публикации с ее тематической категорией.

Как правило, коды классификаторов научных работ либо определяются экспертами вручную, что требует больших трудозатрат, либо проставляются аналогично тематике журналов, в которых эти статьи опубликованы. Более того, в настоящий момент отсутствует единая система классификации, а существующие системы периодически пересматриваются. Их можно разбить на следующие группы: библиотечные классификаторы (УДК, ББК), национальные классификаторы (Field of research (FOR) из Australian and New Zealand Standard Research Classification, шифры научных специальностей Высшей аттестационной комиссии, Общероссийский классификатор специальностей по образованию, Государственный рубрикатор научно-технической информации), международные классификаторы (Field of science and technology OECD, номенклатура ЮНЕСКО для областей науки и техники), классификаторы в международных библиографических базах данных (All Science Journals Classification, классификатор в Web of Science). Таким образом, на точность проводимой классификации влияет не только метод, но и изначально выбранная система классификации.

Для решения задачи классификации текстовых документов применяется множество различных методов. Широко используемым

является метод k-ближайших соседей и его модификации, где классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки. Другим алгоритмом является байесовская классификация, которая работает на вычислении апостериорных вероятностей классов. Представителем линейных классификаторов является метод опорных векторов, который заключается в построении гиперплоскости, разделяющей объекты выборки наиболее оптимальным способом. В последнее время для решения задачи классификации все чаще применяются нейронные сети. В среднем точность различных алгоритмов классификации текстовой информации варьируется от 70 % до 90 % и зависит не только от алгоритмов классификации, но и от качества исходных данных.

Начиная с 2001 года к классификации различного рода информации применяются методы, основанные на теоретико-информационном подходе. Они базируются на алгоритмах компрессии, которые лучше сжимают тексты с близкими лексическими структурами. До настоящего момента подобные методы применялись R. Cilibrasi, M. B. Vitanyi, Кукушкиной О. В., Поликарповым А. А., Хмелёвым Д. В., Рябко Б. Я. и др. в задачах определения авторства текстов, языка, классификации литературных, музыкальных произведений и других.

Однако до сих пор этот подход не был использован при решении задач тематической классификации научных текстов, где он может оказаться перспективным, поскольку публикации из одной дисциплины обычно содержат много общих терминов и словосочетаний.

Таким образом, несмотря на большой практический интерес и научное значение, задача построения методов автоматической классификации научных текстов (статей, монографий и т. п.) далека от своего разрешения.

**Степень разработанности темы исследования.** Методы, основанные на сжатии данных, применялись для кластеризации и классификации литературных произведений, музыкальных файлов, вирусов и других групп живых существ, где в качестве текста был использован их геном, языков человека, компьютерных вирусов. Среди наиболее важных работ стоит отметить

исследования Б. Я. Рябко, О. В. Кукушкиной, А. А. Поликарпова, Д. В. Хмелёва, R. Cilibrasi, P. Vitányi и др.

**Целью работы** является разработка эффективного метода классификации научных текстов, основанного на теоретико-информационном подходе.

**Задачи исследования:**

1. Анализ и экспериментальное сравнение известных методов классификации для выявления их достоинств и недостатков.
2. Построение эффективного метода автоматической классификации научных текстов, базирующегося на теоретико-информационном подходе.
3. Применение разработанного метода для классификации основных типов научных текстов на русском и английском языках:
   А) полнотекстовых документов
   Б) аннотаций публикаций
   для экспериментального подтверждения эффективности метода.

**Объектом исследования** в диссертационной работе являются методы классификации научных текстов.

**Предметом исследования** в диссертационной работе является метод автоматической классификации научных текстов на основе сжатия данных.

**Методология и методы исследования.** В исследовании использовались методы теории информации, алгоритмы сжатия данных. Для программной реализации использовались методы объектно-ориентированного программирования.

**Основные положения, выносимые на защиту**, состоят в следующем:

1. Разработан метод тематической классификации научных текстов, основанный на алгоритмах сжатия потоков символов без потерь.
2. Разработаны два метода эффективного формирования обучающих выборок для классификации текстов, основанные на построении матрицы попарного сжатия и рейтинга цитирования.

3. Изучены свойства метода тематической классификации научных текстов в различных условиях: для массивов полнотекстовых документов и их аннотаций, для текстов на английском и русском языках, для различных алгоритмов сжатия, размеров обучающей выборки, количества классифицируемых категорий и способов предобработки текстов. На основе сравнительного анализа результатов классификации научных текстов традиционными алгоритмами классификации и методом на основе сжатия данных было доказано, что точность предложенного метода выше, чем у остальных.
4. Показаны возможности применения предложенного метода в задачах классификации массивов научных текстов, в том числе для определения тематик научных журналов или публикаций в библиографических базах данных и электронных архивах.

**Научная новизна работы.** Впервые построен метод автоматической классификации научных текстов, основанный на применении алгоритмов сжатия данных для сравнительного анализа «близости» текстов и не требующий дополнительного представления текстов в виде векторов, эффективность которого позволяет использовать его практически в научных библиотеках и других базах данных и знаний.

**Теоретическая и практическая значимость.** Результаты, полученные в диссертационной работе, могут быть использованы как для классификации вновь появляющихся научных публикаций в различных журналах, так и для оптимизации уже существующих систем классификации, например, в библиографических базах данных, таких как Scopus и Web of Science.

**Достоверность результатов** подтверждена экспериментальными исследованиями, основанными на реальных данных, полученных с архива научных публикаций arXiv.org и библиографической базы данных Scopus. Результаты исследования обсуждались на конференциях и семинарах и опубликованы в рецензируемых научных изданиях, рекомендованных ВАК и индексируемых в международных библиографических базах данных Web of Science и Scopus.

**Соответствие диссертации паспорту специальности.** Диссертация соответствует области исследований специальности 05.13.17 – Теоретические основы информатики по п. 2 «Исследование информационных структур, разработка и анализ моделей информационных процессов и структур»; п. 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений».

**Апробация работы.** Основные результаты работы докладывались на следующих российских и международных конференциях:

1. Международная научно-практическая конференция «Наука, технологии и информация в библиотеках (LIBWAY-2020)», 14–17 сентября 2020 г.
2. Распределенные информационно-вычислительные ресурсы. Цифровые двойники и большие данные (DICR-2019), 3–6 декабря 2019 г.
3. 54-я международная научная студенческая конференция МНСК–2016: Информационные технологии 2016, 16–20 апреля 2016 г.
4. International Symposium on Information Theory (ISIT 2017), 25-30 июня 2017 г.
5. International Conference «Mathematical and Information Technologies, MIT–2016», 28 августа–5 сентября 2016 г.

**Публикации.** Основные результаты диссертационного исследования изложены в 8 печатных работах, из которых 4 статьи опубликованы в журналах из списка ВАК РФ, 1 опубликована в журнале, входящем в реферативную базу данных Scopus.

**Личный вклад.** В работах, выполненных в соавторстве, вклад соискателя составляет не менее 80% и заключается в разработке метода классификации и методов формирования ядер, подготовке исходных данных для проведения экспериментов, формирования обучающих выборок, проведении экспериментальных исследований, анализе полученных результатов и подготовке текстов публикаций.

**Внедрение результатов исследования.** Теоретические и практические результаты диссертационного исследования были внедрены в процессе реализации базового проекта научно-исследовательских работ ГПНТБ СО РАН, а также при тематическом анализе публикаций англоязычной версии журнала «Геология и геофизика».

**Структура и объем диссертации.** Диссертация состоит из введения, четырех глав, заключения, списка литературы (128 наименований) и четырех приложений. Общий объем работы 116 страниц. В текст диссертации входят 25 иллюстраций, 18 таблиц.

## ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

**Во введении** обоснована актуальность темы, описаны цель, задачи исследования. Представлены основные положения диссертационной работы, выносимые на защиту.

**В первой главе** представлен обзор наиболее известных методов классификации текстов. Множество из них базируется на терминологической близости. Текст представляется в виде вектора в евклидовом пространстве, где оси координат – это термы, n-граммы или лексемы, выделяемые из текста, а координатой по оси является статистическая информация о них. Таким образом, текст может быть представлен в виде частотных векторов встречаемости слов на основе схем tf, tf*idf, tf*CHI и других. При классификации различными методами (например, kNN, основанными на теории графов или SVM) между векторами рассчитывается мера близости, при этом ее выбор оказывает значительное влияние на качество классификации. В последнее время для классификации текстов применяются нейронные сети, которые показывают высокую точность классификации.

Отдельно рассмотрены подходы, применяемые к классификации научных текстов. Часть из них базируется на цитировании, которое включает в себя прямое цитирование, ко-цитирование и библиографическое сочетание. В других используются стандартные методы классификации текстов, но применяются к спискам соавторов, названиям публикаций, ключевым словам, аннотациям, полным текстам и др. Для улучшения точности классификации применяется комбинация этих двух подходов. Отмечается, что важную роль для улучшения точности классификации научных документов играет выбранная система классификации.

Также приводится обзор работ, где в качестве метода классификации используется метод, базирующийся на сжатии данных. Этот метод показал высокую эффективность при классификации и кластеризации различного вида информации, например, такой как вирусы, литературные произведения, музыкальные файлы, а также определении авторства и языка, на котором написан тот или иной текст.

Применение подхода, основанного на сжатии данных, к классификации научных текстов было бы особенно перспективно, поскольку в научных текстах, относящихся к одной тематике, используются близкие термины и обороты. Но до настоящего момента таких работ проведено не было.

**Во второй главе** рассматривается теория и обоснование работы метода классификации, основанного на сжатии данных. Показано, что на качество классификации влияет размер обучающей выборки, ее состав, архиватор, алгоритм и параметры сжатия.

В общем виде задача классификации формулируется следующим образом[1]. Пусть есть множество описаний объектов X, разделенных некоторым образом на классы. Также имеется некоторое конечное множество объектов Y, для которых известно, к каким классам они относятся. Классовая принадлежность остальных объектов не известна. Требуется построить функцию $\alpha: X \to Y$, способную классифицировать произвольный объект $x \in X$.

Задача классификации текстовых документов строится на том, что в текстах, относящихся к одному классу, присутствует больше общих понятий, терминов и оборотов, чем в текстах из различных классов. Для определения классифицирующей функции $\alpha: X \to Y$ строится мера «лексической близости» текстов.

Для решения задачи классификации научных текстов в диссертационном исследовании используется теоретико-информационный подход. Его идея состоит в том, что степень «лексической близости» текстов оценивается при помощи методов компрессии – сжатия текстов архиваторами: текст, который надо отнести к одному из классов, будет принадлежать к тому из них, с которым он «лучше», или «сильнее», сжимается, то есть к классу, с которым его статистические свойства совпадают лучше всего.

На основе этого подхода соискателем предложен следующий метод. Пусть есть n научных тематик: $X_i$, i=1,…,n. Для каждой $X_i$ назовем тематическим *ядром* множество типичных для нее текстов

---

[1] Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.
URL: http://www.machinelearning.ru/wiki/index.php?title=Классификация (дата обращения: 19.06.2020)

$\{x_j^i;\ i = 1, \dots, n; j = 1, \dots, m_i\}$. Пусть $y$ – это текст, тематику которого нужно определить, причем известно, что он точно относится к одной из указанных тематик. Пусть $\varphi: \{t_1, t_2, \dots\} \to Z$ – это функция компрессии, которая сжимает текстовые данные в закодированное сообщение. Определим функцию $C(y/x_1,\dots,x_k) = (|\varphi(x_1,\dots,x_k, y)| - |\varphi(x_1,\dots,x_k)|) / |y|$ – степень сжатия (в %) текста $y$ с множеством текстов $(x_1,\dots,x_k)$, где $|y|$ – размер текста. Наименьшие значения этой функции соответствуют наилучшему сжатию. Таким образом, полагаем, что текст $y$ принадлежит тематике $X_j$, то есть $C(y/ x_1^j, x_2^j, \dots, x_{m_j}^j) = \min_{i=1,\dots,n} C\left(y/ x_1^i, x_2^i, \dots, x_{m_i}^i\right)$.

Совместно с Б. Я. Рябко было показано, что метод является состоятельным (вероятность ошибки классификации стремится к нулю), если код $\varphi$ является универсальным, то есть по определению сжимает (в пределе) последовательности, порождаемые стационарным эргодическим источником, до минимально возможного значения – энтропии источника (на символ сообщения).

Для проверки работы метода соискателем проведены следующие тесты на динамику сжатия (ТДС). Ядро произвольной научной тематики разбивается на группы, состоящие из одного, двух и т.д. текстов, причем каждая предыдущая группа является подгруппой следующей. Произвольно выбранный тестовый файл последовательно сжимается с каждой из этих групп, и определяется зависимость между количеством текстов в ядре и степенью сжатия текста, которая должна представлять собой монотонно убывающую кривую (рис. 1). Это связано с тем, что при наполнении последовательности символов необходимым количеством символов скорость убывания кривой должна уменьшаться, то есть научный текст у с добавлением новых файлов в тематическое ядро должен все больше быть «похожим» на него.

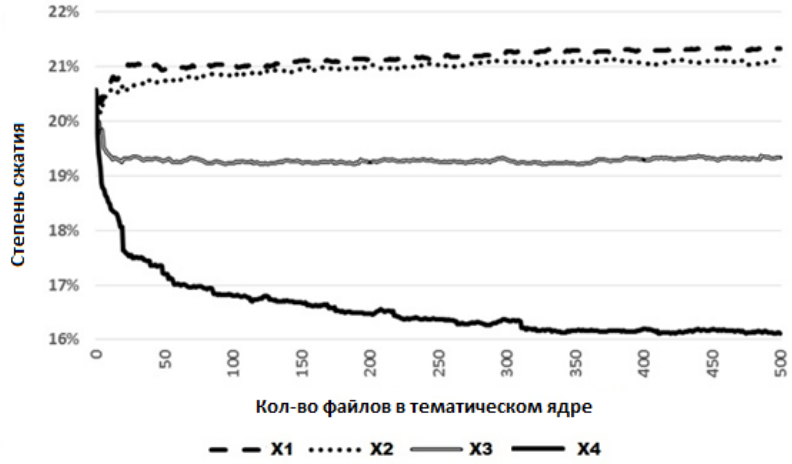



Рисунок 1. ТДС тестового файла из научной тематики $X_4$

Также в главе приводятся результаты выбора архиватора и оптимального размера тематического ядра. Проведенные эксперименты с находящимися в открытом доступе файловыми архиваторами WinRAR, 7z, PeaZIP и различными применяющимися в них алгоритмами сжатия показали, что архиватор WinRAR при максимальном значении памяти 128 Мбайт (память – параметр архиватора) и действующий в нем алгоритм PPMd дает наилучшие результаты (рис. 2).

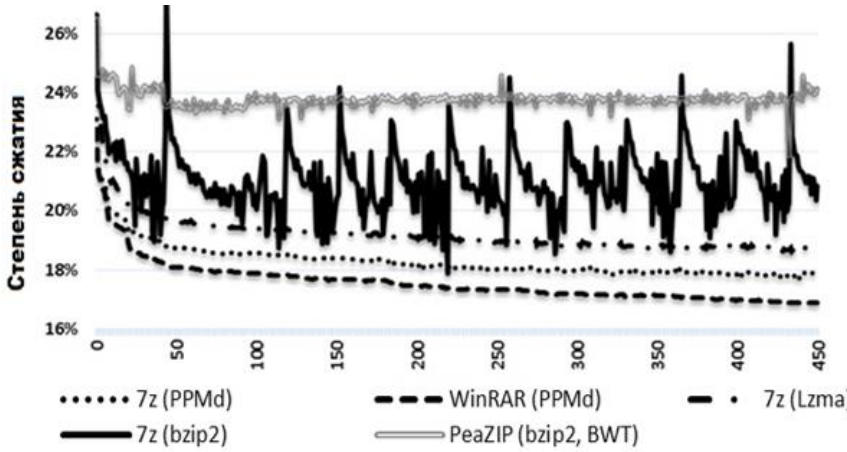



Рисунок 2. Выбор архиватора

Экспериментально доказано, что при уменьшении размера тематического ядра количество ошибок классификации возрастает, а время обработки сокращается. Также показано, что степень сжатия текста зависит от объема тематического ядра и определен его оптимальный объем. При объеме тематического ядра в 100 файлов время обработки одной научной тематики в 8 раз меньше, при этом общее количество ошибок незначительно больше, чем при объеме ядра в 1000 файлов. Поэтому рекомендуемое количество файлов для ядра – не менее 100 текстов, но если мощность вычислительных ресурсов позволяет быстро работать, то его можно увеличить до 1000 (рис. 3, 4).




Рисунок 1. Зависимость времени обработки одной категории (Intel® Core™ i5, RAM 8 Гб) от размера тематического ядра

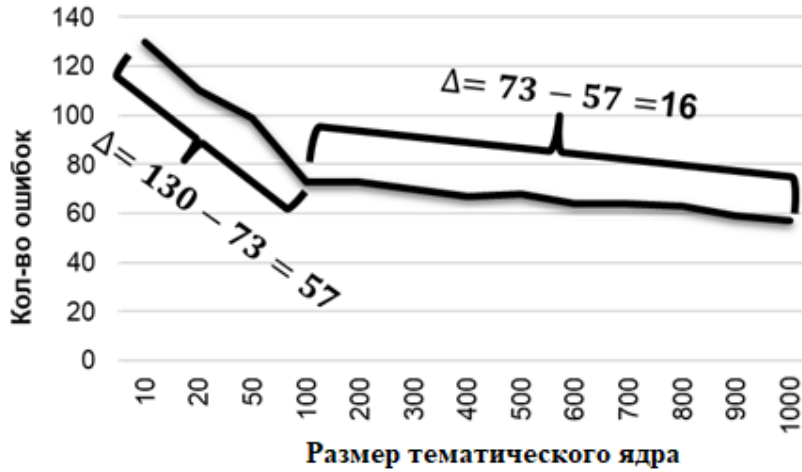



Рисунок 2. Зависимость числа ошибок от размера тематического ядра

Одним из важных параметров работы метода является состав тематического ядра. Соискателем предложены и изучены три метода формирования тематических ядер.

*Метод «Случайный выбор».* В ядро попадают произвольные тексты из научной тематики.

*Метод «Матрица сжатия».* Пусть тексты $x_1^1$, $x_2^1$, …, $x_k^1$ принадлежат научной тематике $X_1$. Возьмем файл $x_1^1$ и вычислим последовательно $C(x_1^1/x_2^1)$, $C(x_1^1/x_3^1)$,…, $C(x_1^1/x_k^1)$. Далее проведем подобную процедуру для каждого текста. Из полученных величин сформируем матрицу сжатия, у которой на пересечении строк и столбцов будет указана степень сжатия каждого текста друг с другом.

Далее для каждого столбца посчитаем величины $S_i = \frac{C(x_i^1)}{k}$, где i = 1, …, k, после чего удалим из строк и столбцов тот текст, где величина $S_i$ будет минимальной (этот текст уже точно войдет в ядро). Это связано с тем фактом, что чем меньше степень сжатия, тем сжатие «лучше». Повторим процедуру столько раз, сколько текстов должно быть в тематическом ядре. Иными словами, в ядре окажутся тексты, несущие наибольший объем информации для других файлов этой научной тематики.

*Метод «Рейтинг цитирования»* применяется для публикаций, источником которых служат ББД, такие как Web of Science, Scopus, РИНЦ и др. Он заключается в том, что в тематическое ядро включаются самые высокоцитируемые публикации из рассматриваемой научной тематики. Такой подход улучшает качество тематического ядра из-за того, что публикации из той же научной тематики, цитирующие отобранные в ядро статьи, с высокой долей вероятности наследуют и характерную лексику. Для улучшения качества классификации в ядра следует включать тексты только с одной тематикой, исключая мультидисциплинарные.

**В третьей главе** рассматриваются результаты классификации научных текстов трех типов:

1. Полные англоязычные тексты (на примере научных текстов, полученных с архива научных текстов arXiv.org);
2. Полные русскоязычные тексты (на примере научных публикаций, полученных из научной электронной библиотеки «КиберЛенинка»);

3. Аннотации публикаций (на примере публикаций, индексируемых в ББД Scopus).

Ввиду того, что полные тексты получены в формате .pdf, из них извлекается текстовый слой и удаляются символы, получившиеся при преобразовании математических формул, цифр, знаков препинания при помощи регулярного выражения. Эксперимент на изучение влияния стоп-слов на качество классификации полнотекстовых документов показывает, что сжатие происходит лучше при их отсутствии (рис. 5).

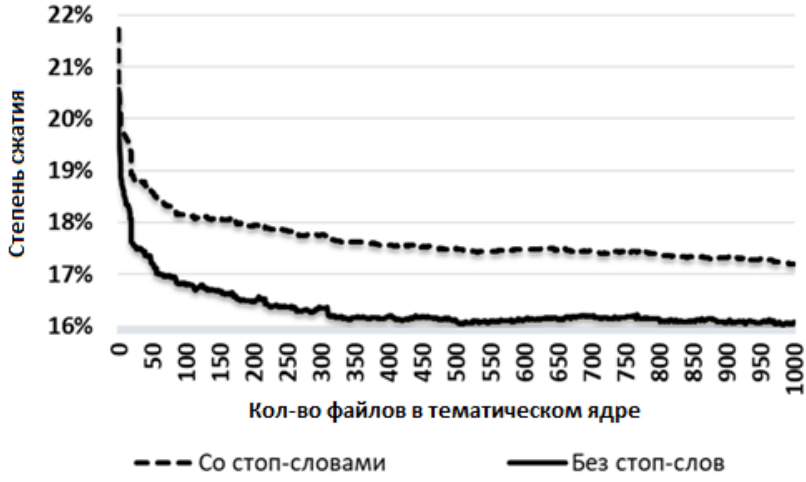



Рисунок 5. Влияние стоп-слов на качество классификации

В открытом архиве научных препринтов arXiv.org используются шесть научных областей (Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistic), в каждой из которых определены тематические категории. Для эксперимента были использованы 30 категорий с наибольшим числом публикаций, в каждой категории было отобрано 20 случайных публикаций-тестов. Из указанных автором категорий первую будем называть главной, а остальные – второстепенными. В ядра вошли по 100 научных текстов, отобранных методами «Случайный выбор» и «Матрица сжатия». Результаты

классификации показали, что при формировании ядер «Случайным выбором» доля ошибок от общего количества тестовых файлов составила 11 %.

При применении метода формирования ядер на основе матрицы сжатия количество ошибок сократилось до 8%. Изначальная классификации текстов на arXiv.org позволила разделить ошибки по типам определения:

- ложный выбор главной категории (3%);
- ложный выбор категории внутри области наук (4%);
- ложный выбор области наук (1%).

Классификация русскоязычных полных текстов в научной электронной библиотеке «КиберЛенинка» основана на 80 категориях Государственного рубрикатора научно-технической информации (ГРНТИ). Произвольным образом были отобраны 400 тестовых файлов, по 20 для каждой области. В результате при произвольном выборе ядра неправильно было определено 47% тестов. При более детальном изучении текстов, полученных из «Киберленинки», было выявлено низкое качество существующей классификации: у большой доли публикаций заведомо неверно была указана научная область. Применение метода «матрица сжатия» позволило сократить количество ошибок более чем в 1,5 раза.

В качестве классификационной системы в библиографической базе данных Scopus используется All Science Journals Classification (ASJC), которая включает четыре общих научных направления (биологические науки, физические науки, медицина, социальные и гуманитарные науки), разделенные на 27 областей наук и более 300 категорий. Классификация публикаций происходит только на уровне журнала (каждой публикации присваиваются все те же категории, что были у журнала), что является значительным недостатком, особенно для мультидисциплинарных журналов.

Для проведения экспериментов были выбраны случайные 30 категорий, у которых за период 2009–2018 гг. было не менее 300 публикаций в одной категории с ненулевым числом цитирований. Было использовано 2 типа ядер:

- сформированные случайным методом;
- сформированные методом на основе цитирования.

В каждой категории было отобрано 20 случайных публикаций-тестов, у которых была указана только эта единственная категория.

При произвольных аннотациях публикаций, индексируемых в ББД SCOPUS, в ядре количество ошибок классификации составило 32%. Подбор ядер из аннотаций высокоцитируемых публикаций позволил сократить их количество до 12%. Разделение ошибок по классификации ASJC показало, что чаще всего возникают следующие ошибки определения категории:

- ложный выбор категории внутри области наук (6%);
- ложный выбор области наук внутри научного направления (5%);
- ложный выбор научного направления (1%).

Показано, что автоматическое создание обучающих выборок для всех категорий классификатора ASJC невозможно из-за:

1. ограничений на выгрузку данных, установленного в Scopus, и отсутствием названий категорий в Scopus Search API.
2. отсутствия в некоторых категориях журналов и, соответственно, публикаций, у которых указана единственная категория.

Сделан вывод о том, что применение метода ко всем 26 областям наук невозможно из-за начальной классификации Scopus: в разных областях наук находятся терминологически близкие категории, что затрудняет отнесение публикации к верной области. Например, категории «Language and Linguistics» и «Linguistics and Language» областей «Arts and Humanities» и «Social Sciences» соответственно, две категории «Archaeology» в этих же областях, три категории «Pharmacology» в областях «Nursing», «Pharmacology, Toxicology and Pharmaceutics» и «Medicine» и др.

Соискателем выявлено, что помимо состава ядер на качество классификации влияет и изначальное количество категорий: чем меньше категорий участвуют при классификации и чем больше терминологическое различие между ними, тем выше ее качество.

В главе также рассматривается применение метода классификации, основанного на сжатии данных, к классификации публикаций из журнала «Геология и геофизика». В качестве ядер используются 12 категорий области «Earth and Planetary Science», обозначенной в классификации ASJC. Изначально этому журналу присвоено только две категории: «Geology» и «Geophysics». Результаты классификации сравниваются с экспертной оценкой. Классификация показала, что в журнале «Геология и геофизика», помимо указанных в Scopus категорий «Geology» и «Geophysics», также присутствуют и публикации других категорий. Так, одними из дополнительных категорий, выявленных методом классификации, основанным на сжатии данных, являются «Stratigraphy» и «Geochemistry and Petrology». В рейтинге Scimago Journal & Country Rank журнал «Геология и геофизика» присутствует только в тех же двух направлениях, что и в Scopus. Scimago Journal Rank (аналог Journal Impact Factor) журнала равен 0.633. Если бы журнал вошел в этих рейтингах дополнительно в категории «Stratigraphy» и «Geochemistry and Petrology», то есть в направления, указанные на сайте издательства, то он бы мог попасть на 16 место в Q2 и 59 место в Q2 соответственно.

В **четвертой главе** проводится сравнение метода, основанного на сжатии данных, с пятью методами классификации:

- Logistic regression (LR),
- k-nearest neighbors (KNN),
- Naive Bayes classifier (NB),
- Random forest (RF),
- Support vector machine (SVM).

Методы были реализованы при помощи бесплатной библиотеки scikit-learn для языка программирования Python. Сравнение было проведено для полнотекстовых документов и аннотаций публикаций.

В зависимости от типа данных лучшую точность классификации показали разные методы. Полнотекстовые документы лучше всего были классифицированы методом на основе сжатия данных, его точность достигла 91%. Остальные методы показали точность ниже 90%. Аннотации публикаций с точностью

89% классифицировал SVM, метод на основе сжатия и LR показали близкую точность в 88%.

Таким образом, метод на основе сжатия данных показывает точность классификации, сопоставимую с известными алгоритмами классификации, однако не требует предварительного представления текста в виде вектора, что значительно упрощает работу.

**В заключении** приводятся основные результаты, полученные в диссертационной работе.

## ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Метод классификации научных текстов на основе алгоритмов сжатия данных является высокоточным при практическом использовании. Его точность зависит от ряда факторов: размера обучающей выборки, ее состава, изначального количества ядер, их тематической близости, архиватора, алгоритма и параметров сжатия.
2. Предложенные в работе методы формирования ядер на основе матрицы сжатия и рейтинга цитирования позволяют улучшить точность работы метода классификации.
3. Метод применен для классификации научных документов, основанных на полных текстах, одной из крупнейших мировых баз данных arXiv.org. Показано, что метод правильно классифицирует до 92% тестовых файлов, что сравнимо с точностью «ручной» классификации специалистами. Анализ ошибок показал, что в 3% случаев у тестового файла определяется второстепенная категория вместо главной, в 4% - определяется другая категория из научной области. Самый серьезный тип ошибок, когда вместо указанной научной области определяется другая, встречается только у 1% файлов.
4. Метод применен для классификации научных документов на основе аннотаций, полученных с библиографической базы данных Scopus. Метод правильно классифицирует до 88% тестовых файлов в зависимости от состава ядра.
5. Эффективность работы метода доказана как экспертной оценкой, так и сравнением его с другими известными методами классификации.

6. Показана возможность практического применения метода в задачах классификации массивов научных текстов, в том числе для уточнения тематик научных публикаций в мультидисциплинарных журналах.

В дальнейшем предложенный алгоритм может быть доработан следующим образом. Как уже упоминалось, большое влияние на точность классификации оказывает выбранный алгоритм сжатия и архиватор. В работе рассматривались архиваторы, находящиеся в открытом доступе, но с закрытым исходным кодом, что могло вызывать некоторые неточности при классификации. Изучение и разработка собственного алгоритма, с нашей точки зрения, могли бы значительно улучшить качество классификаций.

Другим этапом развития работы является разработка системы классификации для русскоязычных научных текстов, поскольку существующие в настоящий момент системы часто работают неверно.

## СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ИССЛЕДОВАНИЯ

*Статьи в журналах из перечня ВАК РФ*

1. Селиванова И. В. Ограничения применения метода на основе сжатия данных к классификации аннотаций публикаций, индексируемых в Scopus / И. В. Селиванова // Вестник НГУ. Серия: Информационные технологии. – 2020. – Т. 18. –№ 3. – С. 57–68.
2. Селиванова И. В. Классификация посредством компрессии: применение методов теории информации для определения тематики научных текстов / И. В. Селиванова, Б. Я. Рябко, А. Е. Гуськов // НТИ. Сер. 2. Информ. процессы и системы. - 2017. - N 6. - С.8-15
3. Рябко Б.Я. Теоретико-информационный метод классификации текстов / Б. Я. Рябко, А. Е. Гуськов, И. В. Селиванова // Проблемы передачи информации. - 2017. - Т.53, N 3. - С.100-111.
4. Селиванова И. В. Классификация посредством компрессии: применение методов теории информации для определения тематики научных текстов / И. В. Селиванова, Д. В. Косяков,

А. Е. Гуськов // НТИ. Сер. 2. Информ. процессы и системы. - 2019. - N 12. - С.25-38

*Статьи в изданиях, индексируемых в Scopus*

5. Ryabko B. Using data-compressors for statistical analysis of problems on homogeneity testing and classification / B. Ryabko, A. Guskov, I. Selivanova // Information Theory (ISIT), 2017 IEEE International Symposium on, 25-30 June 2017, Germany. - Aachen, 2017. - P.121-125.
6. Guskov A. Information-theoretic approach to classification of scientific documents/ A. Guskov, B. Ryabko, I. Selivanova // Математические и информационные технологии MIT-2016. - 2016. - С. 137-138.

*Прочие публикации*

7. Селиванова И.В. Классификация научных текстов посредством сжатия аннотаций на примере публикаций, индексируемых в библиографической базе данных Scopus/ И. В. Селиванова // Распределенные информационно-вычислительные ресурсы. Цифровые двойники и большие данные. (DICR-2019) Труды XVII Международной конференции. - Новосибирск, 2019. - С. 178-184.
8. Селиванова И.В. Метод классификации научных текстов на основе алгоритмов компрессии / И. В. Селиванова // Материалы 54-й международной научной студенческой конференции МНСК-2016 16-20 апреля 2016 г. Серия: Информационные технологии. - Новосибирск, 2016. - С.232.

Селиванова Ирина Вячеславовна

# МЕТОДЫ ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ НАУЧНЫХ ТЕКСТОВ НА ОСНОВЕ ТЕОРЕТИКО-ИНФОРМАЦИОННОГО ПОДХОДА

*Автореферат*

Подписано в печать 24.12.2020. Формат 60x84/16
Усл. п. л. 1,2. Уч.-изд. л. 0,8. Тир. 110 экз. Заказ №272
Полиграфический участок ГПНТБ СО РАН
ул. Восход, 15