Федеральное государственное бюджетное учреждение науки
Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук

На правах рукописи

Селиванова Ирина Вячеславовна

Методы тематической классификации научных текстов на основе теоретико-информационного подхода

05.13.17 – теоретические основы информатики

Диссертация на соискание ученой степени кандидата

технических наук

Научный руководитель

кандидат технических наук

Гуськов Андрей Евгеньевич

Новосибирск – 2020

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность проблемы. В последние десятилетия в связи с экспоненциальным ростом количества информации проблема классификации текстовых документов (текстов), то есть разделения текстов 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, О. В. Кукушкиной, А. А. Поликарповым, Д. В. Хмелёвым, Б. Я. Рябко 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.

Для достижения цели были поставлены следующие **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.

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».

Научная новизна работы заключается в следующем: 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.

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

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 [1–4], 1 опубликована в журнале, входящем в реферативную базу данных Scopus [5].

Внедрение результатов исследования. Теоретические и практические результаты диссертационного исследования были внедрены в процессе реализации базового проекта научно-исследовательских работ в Государственной публичной

научно-технической библиотеке Сибирского отделения Российской академии наук (ГПНТБ СО РАН), а также при тематическом анализе публикаций англоязычной версии журнала «Геология и геофизика».

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.

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.

ГЛАВА 1. КЛАССИФИКАЦИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ В НАУЧНОЙ СФЕРЕ

1.1. Методы классификации текстовых документов

В последние десятилетия в связи с экспоненциальным ростом количества информации проблема классификации текстовых документов (текстов), то есть разделения текстов на заранее заданное множество классов [6], становится особенно актуальной. Она возникает для текстов различного происхождения, для объемных или коротких сообщений и с разными целями.

В качестве источников данных для задач классификации часто используются художественные, поэтические, технические тексты или сообщения в СМИ. Так, в работе [7] задача классификации возникала для стихов Дикинсона и глав ранних американских романов. Показано, что в зависимости от типа данных лучшие результаты дают разные методы. Авторы работы [8] в качестве источника данных для тестирования различных алгоритмов классификации используют лицейскую поэзию Пушкина. Эксперименты проводились на текстах, представленных в виде векторов, составленных как из отдельных слов, так и из биграмм и триграмм. Было выявлено, какие из алгоритмов являются наиболее эффективными для автоматизации комплексного анализа русских поэтических текстов и могут упростить работу специалистов, исследующих русские поэтические стили и жанры. Поэтические тексты на османском языке классифицируются в исследовании [9]. В результате этой работы был выявлен алгоритм классификации и параметры, которые помогают проводить классификацию произведений на османском языке наиболее точно. В работе [10] приводится обзор формальных методов установления авторства, источником данных для которых выступают художественные тексты. В работе [11] задача классификации для устранения двусмысленности возникала для документов по разработке требований.

Работа [12] посвящена классификации журналистских статей. Для сравнения используются два алгоритма классификации, а также различные схемы представления данных. Авторы приходят к выводу, что для высокой точности автоматической классификации первоначальную классификацию документов эффективнее проводить с человеческим вмешательством.

В ряде исследований задачи классификации применяются к коротким сообщениям, например, СМС, твитам и т. д. В работе [13] разрабатывается метод онлайн-классификации, который тестируется на английских и китайских смс-сообщениях. В работе [14] в качестве исследуемых данных также выступают короткие сообщения, такие как СМС и твиты. В связи с тем, что в последнее время люди все чаще используют электронную почту, множество работ посвящено выделению среди общего потока нежелательных электронных писем. В работе [15] к фильтрации электронной почты предлагается подход, основанный на семантических методах. Традиционные методы классификации в применении к электронным письмам сравниваются в работе [16]. Новый метод классификации, основанный на анализе «серого списка» писем, то есть писем «неясного» статуса, предлагается авторами исследования [17].

Одной из часто встречаемых задач классификации текстовой информации является задача определения эмоций. Источником данных для таких работ в основном являются различные микроблоги и социальные сети. Так, в работе [18] предлагается метод классификации сообщений Twitter по различным классам эмоций, которые они отражают. Также сообщения Twitter исследуются на эмоциональный окрас и в работе [19]. Проблема выделения «токсичных» комментариев, то есть сообщений, содержащих угрозы, оскорбления и т.п., обсуждается в работе [20]. Исследование [21] посвящено классификации мнений различных пользователей социальных сетей, относящихся к чрезвычайным ситуациям или другим важным событиям. Однако задача определения эмоций, в виду ограниченной контекстной информации, которую обычно содержат короткие сообщения, является довольно сложно решаемой [22]. В связи с этим

в работе [23] разрабатывается метод классификации твитов, основанный не только на тексте сообщений, но и на информации, извлеченной из профилей авторов.

Методы классификации также эффективно применяются при определении авторства текстов. Это связано с тем, что каждый автор обладает уникальным стилем письма, который раскрывается путем анализа статистических особенностей его текста [24]. Например, в работе [25] традиционные методы классификации применяются для определения авторства поэм «Золотого лотоса». В работе [26] классификация текстов используется для определения пола автора литературных произведений.

Другой задачей классификации является определение тематики текстов. Так, в работе [27] решается задача классификации документов из Википедии по 34 предметным областям.

Множество существующих в настоящий момент методов классификации текстов базируется на терминологической близости. Текст представляется в виде вектора в евклидовом пространстве, где оси координат – это термы, n-граммы [28] или лексемы, выделяемые из текста, а координатой по оси является статистическая информация о них [29]. Таким образом, текст может быть представлен в виде частотных векторов встречаемости слов [30, 31] на основе схем tf, tf*idf, tf*CHI и других [32]. Впервые идея о том, что значимость слова в статье зависит от частоты его встречаемости, была высказана Х. П. Луном в работе [33] в 1958 г. В большинстве случаев из текстов также удаляются стоп-слова [34, 35], то есть слова, которые не несут никакого информационного смысла (предлоги, артикли, местоимения и т.д.), но могут повлиять на качество проводимой классификации. Однако к выбору стоп-слов стоит подходить с особой аккуратностью, т. к. в некоторых задачах, например, при определении типа или авторства текста они могут исказить стилевой окрас произведения, тем самым ухудшив результаты классификации [7].

Другим важным параметром в классификации текстов является мера близости, которая рассчитывается между векторами. При этом ее выбор оказывает значительное влияние на качество классификации [36, 37]. Наиболее известными

метриками являются [38–41]: расстояние Евклида, расстояние Минковского, коэффициент Отиаи, коэффициент Жаккара, проекционное расстояние и др.

Рассмотрим подробнее основные методы, применяемые при классификации текстов. Эти методы относятся к методам машинного обучения с учителем.

К метрическим методам классификации относится метод k-ближайших соседей, где классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки [42]. У классического алгоритма k-ближайших соседей имеется множество модификаций. Это связано с высокой вычислительной сложностью алгоритма и низкой скоростью классификации [43, 44]. В работе [45] приведено сравнение результатов классификации текстов университета Фудань пятью методами: классическим методом k-ближайших соседей [46], k взвешенных ближайших соседей [47], нечетким методом k-ближайших соседей [48], методом k-ближайших соседей, основанном на теории Демпстера – Шафера [49], и k-ближайших соседей, основанном на нечетком интеграле. Показано, что наилучшую точность, 86 %, показывает алгоритм, основанный на нечетком интеграле, тогда как точность классическим алгоритмом k-ближайших соседей составляет только 78 %.

Другой группой классификаторов являются вероятностные [50]. Широко используемым алгоритмом, относящимся к этому классу, является наивная байесовская классификация. Она представляет наиболее простую вариацию байесовских классификаторов – наивный байесовский классификатор, основанный на предположении о независимости признаков. В связи с тем, что в классическом подходе к наивной байесовской классификации часто не включаются веса изученных признаков в оценке условной вероятности, Liangxiao Jiang и соавторы в работе [51] предлагают наивную байесовскую классификацию с глубоким взвешиванием признаков, в которой вычисляются взвешенные характеристики по частотам на основе обучающих данных, а затем эти веса учитываются при расчете вероятности. В работе [52] наивная байесовская классификация применяется при определении авторства текстов. В зависимости от представления текста, например, в виде n-грамм, точность метода в применении к этой задаче показала результаты

от 40 % (при три- и тетраграммах) до 96,67 % (при термах). В работе [53] обнаружена проблема в процессе оценки параметров, которая может влиять на точность наивной байесовской классификации текстовой информации. Для ее устранения авторы предлагают проводить для каждого документа нормализацию текста и использовать метод взвешивания признаков. Для повышения производительности метода наивной байесовской классификации также используется метод вспомогательных функций [54], между словами рассчитывается расстояние Кульбака – Лейблера [55], строят наивные байесовские деревья [56], проводят полиномиальную наивную байесовскую классификацию [57–59], наивную байесовскую классификацию Бернулли [60], гауссовскую наивную байесовскую классификацию [61] и др. В работе [62] показано, что полиномиальная наивная байесовская классификация дает лучший результат при классификации текстов (хотя ее точность составляет только 73,4 %), чем наивная байесовская классификация Бернулли (ее точность – 69,15 %). В работе [63] при сравнении трех методов, основанных на наивной байесовской классификации, показано, что наивная байесовская классификации Бернулли сравнима по результатам с классической, тогда как гауссовский наивный байесовский классификатор дает самую лучшую точность классификации.

Одним из представителей линейных классификаторов является метод опорных векторов, который заключается в построении гиперплоскости, разделяющей объекты выборки наиболее оптимальным способом [64]. В работе [65] предложена модификация метода опорных векторов, в которой выбор характеристик происходит с использованием схемы взвешенных энтропий. В работе [66] в качестве метода классификации текста и метода организации знаний используется комбинация метода опорных векторов и стратегии, организованной онтологией и пользовательскими базами знаний.

Также существует классификация, базирующаяся на методах теории графов. К ней относится, например, метод «случайный лес» (random forest). Он заключается в построении ансамбля параллельно обучаемых независимых деревьев решений [67]. В ряде исследований приводятся способы улучшения работы метода

случайного леса. В работе [68] для решения многоклассовых задач для вычисления весов объектов предлагается использовать метод хи-квадратов. Благодаря новому методу взвешивания признаков для выборки подпространства и метода выбора дерева эффективно уменьшается размер подпространства и повышается производительность классификации. В зависимости от массива данных метод может проявлять точность классификации от 72 % до 92 %. В работе [69] приводится алгоритм случайного леса с учетом семантики. Этот алгоритм на деревьях разного размера показывает точность 73–78 %, тогда как точность классического алгоритма составляет 57–60 % [70].

В последнее время для решения задачи классификации все чаще применяются нейронные сети. В работе [71] Siwei Lai и соавторы для решения задачи классификации текстов предлагают использовать рекуррентные сверточные нейронные сети. Авторы приходят к выводу, что применение нейронных сетей при классификации текстовых документов поможет избежать проблемы разреженности данных, а также собрать больше контекстуальной информации о сущностях по сравнению с традиционными методами. Сверточные нейронные сети показали высокую точность (83,98%) и при классификации патентных документов [31].

Существует множество работ, направленных на сравнение точности классификации текстовых документов различными методами. Так, в работе [45] при сравнении трех методов: k-ближайших соседей, на основе нечеткого интеграла, метода опорных векторов и байесовской классификации – наилучшую точность, 90%, показывает метод опорных векторов. В работе [72] при классификации твитов на турецком языке методы показывали различные результаты классификации в зависимости от размера обучающей выборки. Наилучшие результаты, от 63% до 83%, во всех трех случаях демонстрировала байесовская классификация. Наилучшую точность, 83 %, на одной из выборок байесовская классификация показывает и в работе [73]. В работе [74] при классификации книг наилучшую точность, 81 %, также показывает байесовский классификатор. Но при классификации индийских и английских твитов в работе [75], несмотря на то что

байесовская классификация была самой эффективной, ее точность не превышала 63 %. В работе [76] для классификации данных с новостных веб-сайтов используются пять классификаторов: k-ближайших соседей, случайный лес, полиномиальный наивный байесовский классификатор, логистическая регрессия и метод опорных векторов. Самым эффективным алгоритмом оказался метод опорных векторов, который продемонстрировал не только высокую точность в 91 %, но и самое быстрое время работы: минимум в полтора раза ниже, чем у других исследуемых алгоритмов. В работе [77] сравнение трех методов: k- ближайших соседей, наивной байесовской классификации и метода опорных векторов – показало, что при их применении к классификации публикаций по окружающей среде, спорту, политике и искусству методы показывают точность от 73 до 97 %. Сравнение методов k-ближайших соседей, метода опорных векторов, циклической нейронной сети и рекуррентной нейронной сети на корпусе английских текстов в работе [78] показало, что самую высокую точность классификации, достигающую 96 %, имеет рекуррентная нейронная сеть. Но на этом корпусе документов и остальные методы показывают точность не ниже 88 %.

Для улучшения точности классификации используют и комбинации различных алгоритмов классификации. Например, в работе [79] комбинация алгоритмов k-ближайших соседей и метода опорных векторов делает точность классификации выше на 1–2 %, чем при применении этих классификаторов отдельно. В работе [80] комбинация методов k-ближайших соседей, алгоритма Роккио и метода наименьших квадратов уменьшило число ошибок классификации на 15 %.

Таким образом, в среднем точность различных алгоритмов классификации текстовой информации варьируется от 70 % до 90 %. При этом точность классификации зависит не только от выбранного алгоритма классификации, но и от исходных данных.

1.2. Методы классификации научных текстов

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.

Основным преимуществом классификации научных публикаций является общность терминов, понятий и оборотов, используемых в текстах одной и той же области наук. При этом, чем более узконаправленной является научная область, тем более специфичной является лексика статей, относящихся к ней.

Для классификации научных публикаций широко используются подходы на основе цитирования. Обычно эти методы строятся на прямом цитировании, ко-цитировании и библиографическом сочетании. В работе [81] проводится сравнение эффективности следующих методов на основе цитирования: спектральных методов, алгоритмов модулярной оптимизации, картографических методов, матричной факторизации, статистических методов, кластеризации по ссылкам и других. В этом исследовании наилучшие результаты показали картографические методы. В работе [82] для классификации предложены два подхода: на основе связанных графов и на основе интеграции графов. Предложенные методы работают на основе интеграции атрибутной текстовой информации и информации о цитировании, полученной из ББД WoS. В работе [83] выявлено, что методы, основанные на библиографическом сочетании, дают лучшие результаты, чем прямое цитирование и ко-цитирование. В работе [84] прямое цитирование используется для кластеризации астрофизических данных.

В других методах в качестве исходной информации берутся аннотации публикаций, их полные тексты, информация об авторах, названиях публикаций,

ключевых словах и т.д. Далее к ним применяются традиционные методы классификации.

Примерами работ, в которых используются подобные подходы, являются следующие. В работе [85] они применялись при кластеризации биомедицинских публикаций в базе данных MEDLINE. Классификацией аннотаций публикаций из научных направлений Materials science, Physics и Chemistry ББД Scopus занимались Vahe Tshitoyan с соавторами в работе [86]. Проверка качества классификации медицинских и биологических публикаций из базы данных PubMed путем сравнения результатов применения методов k-ближайших соседей, байесовской классификации и опорных векторов к аннотациям рассмотрена в работе [87]. R. Koopman и S. Wang в работе [88] строят семантическое представление статей на основе семантических векторов темы, названия, авторов, журнала и цитирования. Далее для каждой статьи происходит объединение этих векторов, после чего применяются два алгоритма кластеризации: k-means и метод Лувена для обнаружения сообществ. Также упоминается о том, что этот метод может служить первым этапом кластеризации. В работе [89] классические методы классификации применяются для определения важности формульного содержимого в научных текстах математической тематики.

В некоторых работах используются комбинации этих подходов. Так, в работе [90] была проведена гибридная кластеризация, основанная на комбинации библиографического сочетания и расчета текстовой близости с использованием метода Лувена для данных области «Астрономия и Астрофизика».

1.3. Системы классификации научной информации

Важную роль для улучшения точности классификации научных документов играет выбранная система классификации. Как правило, коды классификаторов научных работ либо определяются экспертами вручную [91], 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.

Эти системы могут быть разделены на несколько типов:

- библиотечные классификаторы [92]. К ним относится, например, УДК [93] – индекс Универсальной десятичной классификации. За основу этой системы взята десятичная классификация, разработанная в 1876 году американским библиографом Мелвилом Дьюи. Центральной частью УДК являются основные таблицы, охватывающие всю совокупность знаний и построенные по иерархическому принципу деления от общего к частному с использованием цифрового десятичного кода. Также к библиотечным классификаторам относится ББК – библиотечно-библиографический классификатор. ББК является национальной классификационной системой России. Его принцип формирования аналогичен УДК. Другим примером библиотечных классификаторов является Классификация библиотеки Конгресса США [94].
- национальные классификаторы. Классификация Fields of research (FOR) из Australian and New Zealand Standard Research Classification (ANZSRC) [95] представляет собой иерархическую классификацию с тремя уровнями. На первой ступени находятся обширные научные направления. На второй – связанные с ними группы. На третьем уровне расположены более узкие области. FOR применяется, например, в платформе Dimensions. Другим примером национальных классификаторов являются шифры научных специальностей Высшей аттестационной комиссии (ВАК) [96], Общероссийский классификатор специальностей по образованию (ОКСО) [97], который сопоставлен с Международной стандартной классификацией образования МСКО и Государственный рубрикатор научно-технической информации (ГРНТИ) [98].
- международные классификаторы. Прежде всего к ним относится Field of science and technology (FOS) – система классификации, опубликованная Организацией экономического сотрудничества и развития в 2002 году. В 2006 году

для полноты отражения изменений в области науки и техники в нее были внесены некоторые изменения [99]. В основе классификации лежат шесть научных направлений, разделенных на 42 области. Эти области в свою очередь разделены на подкатегории. Другим примером международного классификатора является номенклатура UNESCO – система, разработанная ЮНЕСКО для классификации научных работ и диссертаций [100].

- классификаторы в международных библиографических базах данных. Как и в предыдущих случаях, эти классификаторы построены по иерархическому принципу. В WoS классификатор состоит из трех уровней. На основном уровне расположены шесть областей наук, которые разделены на 39 подуровней. На третьем подуровне классификатора WoS содержится 253 категории [101]. В Scopus используется All Science Journals Classification (ASJC). Она включает в себя издания, распределенные по четырем общим научным направлениям: биологические науки, физические науки, медицина, социальные и гуманитарные науки. Они разделены на 27 крупных предметных областей и более 300 узких категорий [102]. Несмотря на широкий охват тематик, у ASJC есть существенные недостатки. Например, в двух разных научных областях встречаются две близкие категории: Language and Linguistics (код – 1203, область – Arts and Humanities) и Linguistics and Language (код – 3310, область – Social Sciences). Эта же проблема была отмечена еще в 2016 году Q. Wang и L. Waltman в работе [103], где авторы предлагали либо произвести слияние этих категорий, либо обозначить более четкие различия между ними.

Различные системы классификаций часто не согласуются между собой[104]. Особенно это выражено в международных библиографических базах данных, что может негативно сказываться на результатах, получаемых при оценке научных направлений. Издания, включенные в одну категорию WoS, могут не совпадать с изданиями из аналогичной категории Scopus, поэтому даже проводя оценку по одной стране, но по данным, полученным из разных систем, можно получить противоположные результаты. Сравнением представленности журналов направления «Библиотечные и информационные науки» в категориях ББД Scopus

и Web of Science проводились в работе [105]. Авторы показывают, что 23 журнала, относящихся в WoS и библиотечно-информационным наукам, в Scopus были отнесены к категориям: «Закон», «Информационные системы», «Медицина», «Компьютерные науки», «Образование» и другим. И наоборот, 11 журналов из Scopus были отнесены к другим категориям WoS. В работе [106] при сравнении этих двух ББД на примере Словении авторы пришли к выводу, что наиболее часто различия происходят в социальных, гуманитарных, технических областях, причем в Scopus выдается значительно больше результатов по количеству общих документов и цитирований, чем в WoS.

Стоит также отметить, что существует множество работ, результаты которых основываются на классификаторах этих ББД. Так, в работе [107] тематические рубрики ББД Scopus используются при выявлении вклада публикаций, находящихся в открытом доступе. В работе [108] на основе тематических рубрик проводится сравнение цитирования, полученного из Google Scholar, WoS и Scopus. Классификатор ББД Scopus используется и для классификации малазийских публикаций в работе [109]. Результаты работы [110] также получены с помощью классификатора ББД Scopus. Авторы рассматривают вопрос «трансатлантического разрыва», то есть расстояния между Европой и США с точки зрения исследовательского процесса. Исследование строится на оценке количества публикаций и цитирований в каждой дисциплине, обозначенной в классификаторе Scopus. В работе [111] представляется количественный показатель международного научного влияния стран на основе данных о публикациях и цитировании из классификатора ББД Scopus по тематике «Energy (all)». В работе [112] на основе классификатора Scopus были посчитаны показатели «Средняя нормализованная оценка цитирования» и «Средняя нормализованная логарифмическая оценка цитирования». В статье также упоминается о том, что выбранные схемы классификации будут оказывать значительное влияние при использовании этих показателей при составлении различных политических решений. В работе [113] классификатор ББД Scopus используется в качестве

показателя «надежности» двухуровневой классификации журналов в контексте национальной программы содействия развитию Италии.

Таким образом, уже упоминаемые неточности классификаторов в ББД могут вносить значительные погрешности в результаты исследований и оценок, а выбор другой системы классификации привести к другим результатам.

Важность выбора системы классификации отмечается, например, в работе [114]. При рассмотрении двух португальских систем классификации: FCT и DeGóis platform, было обнаружено, что область науки Nursing полностью отсутствует в первой системе, а во второй она нечетко определена. В работе [115] показаны две основные проблемы номенклатуры UNESCO. Первая из них связана с тем, что в этой классификации в крупных научных областях могут быть потеряны более мелкие категории, вследствие чего классификация будет недостаточно полной. Вторая проблема заключается в том, что в этой классификации отсутствуют области наук, которые появились недавно, например, астробиология, что также может существенно ухудшить качество классификации.

1.4. Подходы к классификации, основанные на методах теории информации

Подходы, основанные на сжатии данных, стали применяться к классификации различного рода информации начиная с 2001 года. Так, в работе [116] были рассмотрены три различные техники построения метода при решении различных задач классификации:

1. Алгоритмом LZ78;
2. Алгоритмом PPM порядка 3 слов;
3. Алгоритмом PPM порядка 5 символов.

В результате было показано, что наиболее точным методом, определяющим авторство, принадлежность классам технических документов, идентифицирующим язык текста, является метод, основанный на алгоритме PPM порядка 5 символов.

Точность его применения составила 85 %, 75 % и 100 % соответственно. Также в работе был сделан вывод о том, что при размере обучающей выборки меньше, чем $0{,}5\times10^5$ байт точность работы метода составляет менее, чем 70 %, при этом чем меньше размер выборки, тем скорость падения точности быстрее.

В работе [117] показано, что метод, основанный на сжатии данных, является эффективным при определении авторства текста и простым в использовании ввиду того, что большинство архиваторов широко распространено и имеет реализацию на многих платформах. Применение этого метода к задаче определения авторства текстов также исследовалось Б. Я. Рябко и др. в работе [118].

В исследованиях [119–121] R. Cilibrasi, M. B. Vitanyi и др. применяли методы сжатия данных для классификации и кластеризации литературных произведений; музыкальных файлов; вирусов и других групп живых существ, где в качестве текста был использован их геном; языков человека; компьютерных вирусов и многих других объектов. Здесь на первом этапе была определена мера близости, рассчитанная по длинам сжатых файлов. Далее применялся метод иерархической кластеризации.

Таким образом, метод, основанный на сжатии данных, оказался довольно простым в использовании и показал высокие результаты при решении различного рода задач. Более того, в отличие от традиционных методов классификации, метод может быть применен к текстам без какой-либо дополнительной обработки.

Но для классификации научных текстов этот подход применен не был, хотя в научных текстах, относящихся к одной тематике, используются близкие термины и обороты. Поэтому применение к ним метода классификации, основанного на сжатии, представляется весьма перспективным.

1.5. Выводы к Главе 1

В Главе 1 представлен обзор наиболее известных методов классификации текстов. Большинство 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. При классификации различными методами между векторами рассчитывается мера близости, при этом ее выбор оказывает значительное влияние на качество классификации. Разные методы показывают разную точность классификации в зависимости от условий, в которых она проводится: массива данных, параметров алгоритмов. В основном точность классификации этими методами варьируется от 70 до 90 %. Чаще всего наилучшие результаты показывает наивная байесовская классификация и метод опорных векторов и их модификации.

Отдельно рассмотрены подходы, применимые 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, вирусов, литературных произведений, музыкальных

файлов, 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, т. к. в научных текстах, относящихся к одной тематике, используются близкие термины и обороты. Но до настоящего момента таких работ проведено не было.

ГЛАВА 2. МЕТОД КЛАССИФИКАЦИИ, ОСНОВАННЫЙ НА СЖАТИИ ДАННЫХ

2.1. Задача классификации текстовых документов

Классическая задача классификации текстовых документов формулируется следующим образом [122]. Пусть $D = \{d_1, \dots, d_{|d|}\}$ – это множество документов, $C = \{c_1, \dots, c_{|C|}\}$ – множество заранее заданных категорий (классов). Также имеется некоторая неизвестная целевая функция $\Phi: D \times C \to [0,1]$, задаваемая формулой

$$\Phi(d_j, c_i) = \begin{cases} 0, \text{если } d_j \notin c_i \\ 1, \text{если } d_j \in c_i \end{cases}$$

Задача текстовой классификации состоит в построении классификатора Φ', максимально близкого к Ф, при этом «близость» определяется одной из метрик, перечисленных в разделе 1.1.

2.2. Метод классификации научных текстов, основанный на сжатии данных

Рассмотрим метод классификации научных текстов, основанный на сжатии данных. Метод был предложен в работе [1]. Основная идея метода состоит в том, что в текстах, относящихся к одной области, используется много общих понятий, терминов и оборотов, причем чем у́же рассматриваемая область, тем «ближе» лексика текстов, относящихся к ней. Степень «лексической близости» текстов оценивается при помощи методов компрессии – сжатия текстов архиваторами: текст, который надо отнести к одной из областей наук, будет принадлежать к той из них, с которой он «лучше», или «сильнее», сжимается.

Рассмотрим формальное описание метода. Пусть есть n научных тематик: X_i, i=1, …, n. Для каждой X_i назовем тематическим *ядром* множество типичных для нее текстов $\{x_j^i;\ i = 1, \dots, n; j = 1, \dots, m_i\}$.

Пусть *y* – это классифицируемый научный текст, научную 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, – «архиватор». 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 %) текстов *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 $C(y/ x_1^j, x_2^j, \dots, x_{m_j}^j) = \min\limits_{i=1,\dots,n} C\left(y/ x_1^i, x_2^i, \dots, x_{m_i}^i\right)$.

Метод является состоятельным (вероятность ошибки классификации стремится к нулю), если код φ является универсальным, то есть по определению сжимает (в пределе) последовательности, порождаемые стационарным эргодическим источником, до минимально возможного значения – энтропии источника (на символ сообщения). Реализация метода представлена в приложениях А, Б.

2.3. Представление полученных результатов

Одним из удобных представлений результатов является матрица сжатия тестовых файлов с каждой из категорий (Таблица 1), нормированной на проценты сжатия, то есть из каждой строки был вычтен минимальный процент сжатия. В итоге, нулевой процент сжатия обозначает, к какой категории относится тест.

Таблица 1 – Представление полученных результатов (**жирным** обозначены минимальные проценты сжатия у междисциплинарных тестовых файлов)

Категория	**test1**	**test2**	**test3**	**test4**	**test5**	**test6**
Category 1	1,74 %	**0 %**	1,88%	0,95%	2,95%	0,53 %
Category 2	0 %	1,19 %	2,24 %	0,71 %	1,81 %	1,22 %
Category 3	1,05 %	1,76 %	2,55 %	0,79 %	**0 %**	0,85 %
Category 4	1,97 %	2,28 %	0,36 %	0 %	**0,05 %**	1,96 %
Category 5	1,71 %	**0 %**	0 %	1,70 %	1,24 %	0,73 %
Category 6	0,86 %	1,09 %	2,29 %	1,87 %	0,94 %	0 %

Данное представление также удобно, если статья является междисциплинарной. В таблице 1 к таким случаям относятся test 2 (у первой и пятой категорий были одинаковые минимальные проценты сжатия) и test 5 (в этом случае проценты сжатия у третьей и четвертой категорий отличается лишь на 0,05 %).

2.4. Тест на динамику сжатия

Для проверки работы метода проведем тест на динамику сжатия (ТДС), который состоит в следующем. Пусть X – научная тематика, $J \subseteq X$ – тематическое ядро этой тематики, $q = |X|$ – размер тематического ядра. Разобьем J на группы, состоящие из одного, двух и т.д. текстов, причем каждая предыдущая группа является подгруппой следующей: $J_1 = \{x_1\} \subseteq J_2 = \{x_1, x_2\} \subseteq J_3 = \{x_1, x_2, x_3\} \subseteq \ldots \subseteq J_q = \{x_1, x_2, \ldots, x_q\}$. Пусть x_s – произвольно выбранный научный текст из научной тематики X, $x_s \notin J$. Вычислим последовательно степени сжатия x_s с каждой из этих групп, то есть величины $C(x_s/ J_1)$, $C(x_s/ J_2)$, $C(x_s/ J_3)$, …, $C(x_s/ J_q)$, и рассмотрим

зависимость между этими величинами и размером группы J_i, i=1, …, q. В итоге эта зависимость должна представлять собой монотонно убывающую кривую. 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. При этом если рассматриваются ТДС текста x_s сразу с несколькими научными тематиками, то кривая, соответствующая научной тематике текста x_s, должна идти ниже всех остальных. На рисунке 1 представлен пример ТДС текста из научной тематики X_4 с группами ядер научных тематик X_1, X_2, X_3, X_4.

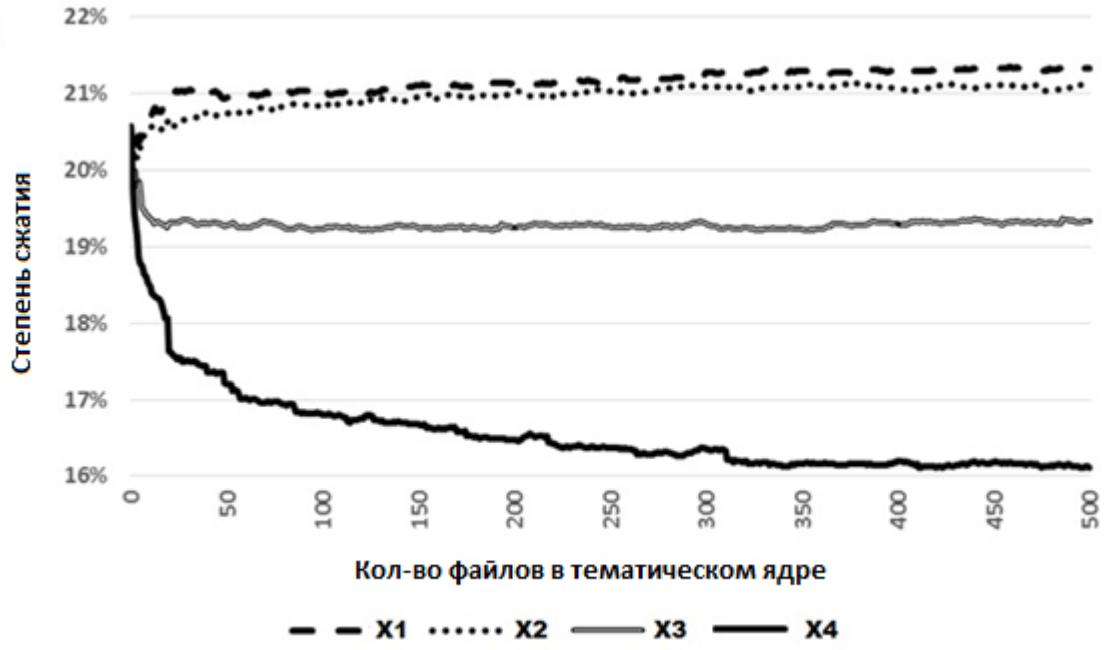


Рисунок 1 – ТДС тестового файла из научной тематики X_4

2.5. Выбор архиватора и оптимального размера ядра

ТДС позволяет провести выбор одного из важнейших параметров работы метода классификации, основанного на сжатии данных, – архиватора φ. Для этого возьмем одну область наук и тестовый файл из нее и проведем ТДС

с находящимися в открытом доступе файловыми архиваторами WinRAR, PeaZIP, 7z и различными реализованными в них алгоритмами: PPMd, BWT, PPMd, LZMA и BWT соответственно (Рисунок 2). В каждом алгоритме использовался максимальный размер памяти для словаря (память – параметр архиватора), а также максимальная степень сжатия.

Убывающая кривая получается при работе с алгоритмами PPMd и LZMA в архиваторах WinRAR и 7z. Но минимальную степень сжатия (наилучшее сжатие) дает архиватор WinRAR (алгоритм PPMd) при максимальном значении памяти 128 Мбайт. Таким образом, для дальнейших экспериментов будет использован архиватор WinRAR. К аналогичному выводу при сжатии текстовых файлов привели эксперименты, проведенные в работе [123].

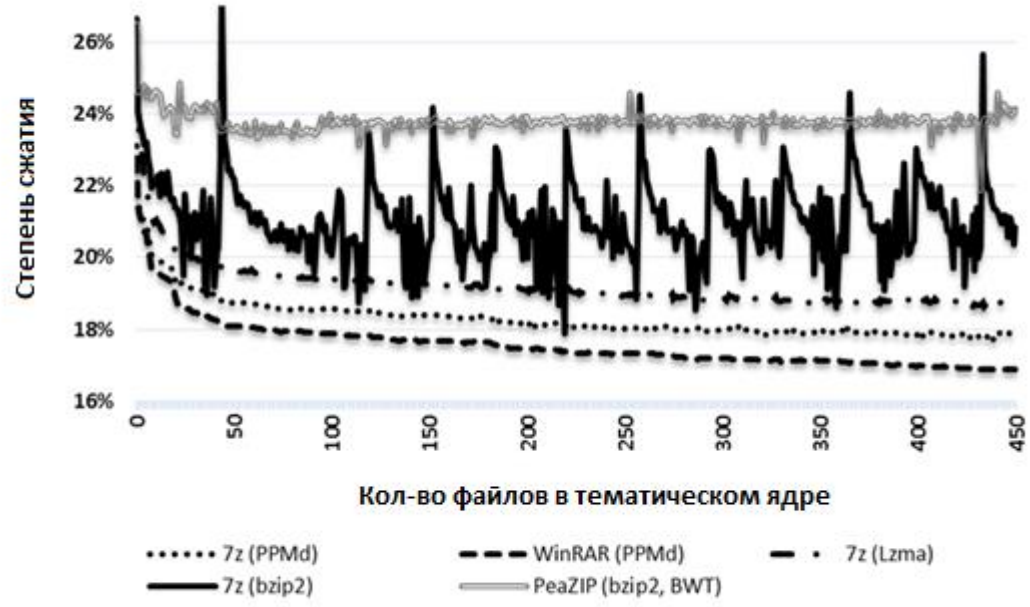


Рисунок 2 – Выбор архиватора

Другой важный вопрос – выбор количества текстов, входящих в ядро, то есть выбор параметра m_i в $J = \{x_j^i;\ i = 1, \dots, n; j = 1, \dots, m_i\}$.

На рисунках 3, 4 показано, за какое время[1] и с каким количеством ошибок обрабатывается каждое ядро. Всего проверка осуществлялась на 450 тестах, то есть по 15 файлов из каждой категории, не входящих в ядра.

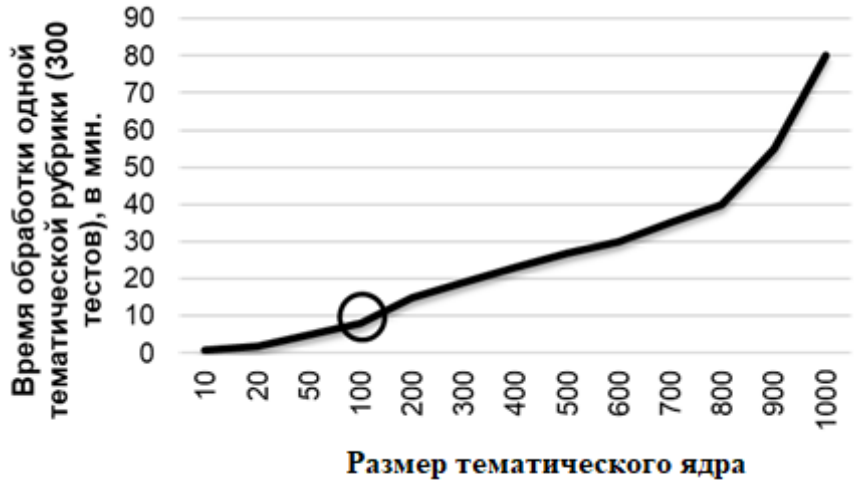


Рисунок 3 – Зависимость времени обработки одной категории (Intel® Core™ i5, RAM 8гб) от размера тематического ядра

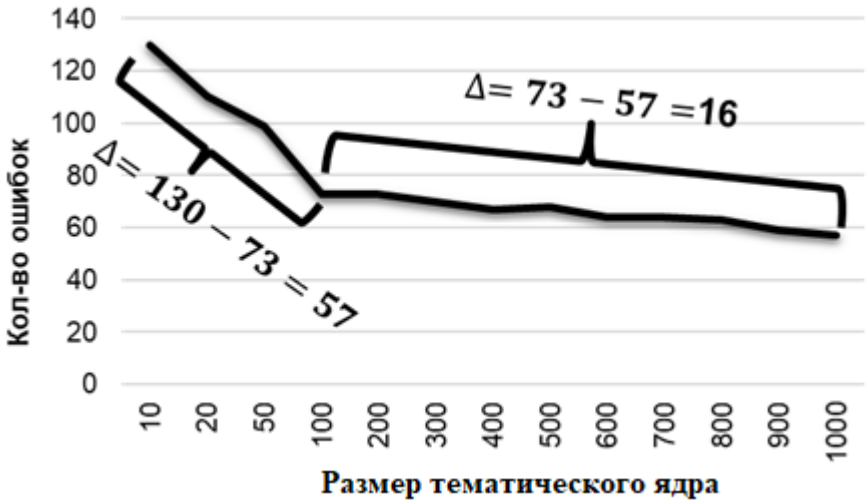


Рисунок 4 – Зависимость числа ошибок от размера тематического ядра

Чем больше файлов в ядре, тем дольше работает метод, но при этом уменьшается число ошибок. Таким образом, нужно подобрать такое ядро, при котором требуется небольшое время обработки одной категории, при этом количество ошибок минимально. При 100 файлах в ядре количество ошибок лишь

[1] Данные обрабатывались на машине Intel® Core™ i5, RAM 8гб

на 15 меньше, чем в ядре с 1000 файлами. Если же рассмотреть разницу в количестве ошибок при 100 файлах в ядре и 10, то она будет составлять 57, тогда как разница в количестве ошибок между 1000 и 100 файлами составляет только 16. При этом время обработки категории при 100 файлах уменьшается примерно в 8 раз по сравнению с 1000 файлами.

Исходя из вышеописанных результатов для классификации текстов в дальнейшем будет использоваться ядро, состоящее из 100 файлов, т.е. $J = \{x_j^i;\ i = 1, \dots, n; j = 1, \dots, 100\}$.

2.6. Методы 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, и отсортируем столбцы по возрастанию S_i. Это связано с тем фактом, что чем меньше степень сжатия, тем сжатие «лучше». Теперь удалим из строк и столбцов тот текст, где величина S_i будет минимальной (этот текст уже точно войдет в ядро как текст, с которым все остальные тексты в ядре имеют наилучшее сжатие), после чего повторим процедуру до тех пор, пока в тематическое ядро не войдет необходимое количество файлов. 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, чтобы включать в тематическое ядро самые высокоцитируемые публикации из рассматриваемой научной тематики. Такой подход улучшает качество 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.

2.7. Выводы к Главе 2

В Главе 2 рассматривается теория и обоснование работы метода классификации, основанного на сжатии данных. Производится выбор архиватора и параметров, обеспечивающих наилучшую работу метода. Показано, что архиватор WinRAR при максимальном значении памяти 128 Мбайт и действующий в нем алгоритм PPMd дает наилучшие результаты.

Экспериментально доказано, что при уменьшении размера тематического ядра количество ошибок классификации возрастает, а время обработки сокращается. Показано, что степень сжатия текста зависит от объема тематического ядра и определен его оптимальный объем в 100 файлов.

Для улучшения качества классификации предложены три метода формирования ядер:

- Случайный выбор,
- Матрица сжатия,
- Рейтинг цитирования.

Эти методы могут быть применены в зависимости от данных, для которых применяется метод классификации на основе сжатия.

ГЛАВА 3. РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ МЕТОДА КЛАССИФИКАЦИИ, ОСНОВАННОГО НА СЖАТИИ ДАННЫХ

3.1. Результаты классификации полных англоязычных научных текстов

Источником англоязычных полнотекстовых документов является архив научных текстов arXiv.org. Изначальная классификация на этом сайте построена следующим образом. На верхнем уровне расположено 8 научных областей: Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science и Economics. Для каждой из этой области определены научные категории. Например, область Mathematics разбивается на Symplectic Geometry, Metric Geometry, Number Theory, Numerical Analysis, Operator Algebras и др.

При размещении текста на сайте автор указывает одну или несколько категорий, к которым принадлежит его работа. Будем называть первую категорию, указываемую автором, главной, остальные – второстепенными.

Для проведения эксперимента были выбраны 30 категорий из четырех областей наук: Physics, Mathematics, Quantitative Biology и Computer Science. Отметим, что одна категория может относиться к нескольким научным направлениям: например, Mathematical physics относится как к Mathematics, так и к Physics, а Information theory относится к Computer Science и Mathematics.

Процесс извлечения данных и их последующая обработка состояли в следующем:

1. Получение полных текстов в формате .pdf с веб-сайта arXiv.org (таблица 2);
2. Извлечение текстового слоя из каждого .pdf-файла (при помощи PDF2Text Pilot – бесплатного программного обеспечения с функцией пакетной обработки);
3. Выбор только англоязычных текстов;

4. Удаление символов, получившихся при преобразовании математических формул, цифр, знаков препинания при помощи регулярного выражения;
5. Удаление стоп-слов, т.е. слов, не несущих научного смысла: предлогов, наречий, частиц и т.д. (рисунок 5);
6. Удаление из массивов текстов размером менее 10 кб.

Таблица 2 – Количество файлов, полученных с веб-сайта arXiv.org, по категориям

Код категории	**Название категории**	**Количество файлов**
astro-ph.CO	Cosmology and Nongalactic Astrophysics	1347
astro-ph.GA	Astrophysics of Galaxies	1327
astro-ph.HE	High Energy Astrophysical Phenomena	1133
cond-mat.dis-nn	Disordered Systems and Neural Networks	1300
cond-mat.stat-mech	Statistical Mechanics	1252
gr-qc	General Relativity and Quantum Cosmology	1357
hep-ex	High Energy Physics - Experiment	1310
hep-th	High Energy Physics - Theory	1306
nucl-ex	Nuclear Experiment	1206
nucl-th	Nuclear Theory	1314
physics.acc-ph	Accelerator Physics	1138
physics.atom-ph	Atomic Physics	1318
physics.ins-det	Instrumentation and Detectors	1056
physics.optics	Optics	1583
physics.soc-ph	Physics and Society	1084
quant-ph	Quantum Physics	1262
math-ph	Mathematical Physics	1796
math.AG	Algebraic Geometry	1222
math.CO	Combinatorics	1066
math.DG	Differential Geometry	1237
math.FA	Functional Analysis	1397
math.GR	Group Theory	1350
math.PR	Probability	1527
math.ST	Statistics Theory	1131
cs.IT	Information Theory	1198
cs.AI	Artificial Intelligence	1372

Продолжение таблицы 2

Код категории	Название категории	Количество файлов
cs.CR	Cryptography and Security	1338
cs.LO	Logic in Computer Science	1064
cs.SE	Software Engineering	1309
q-bio.BM	Biomolecules	863
Всего		**38 163**

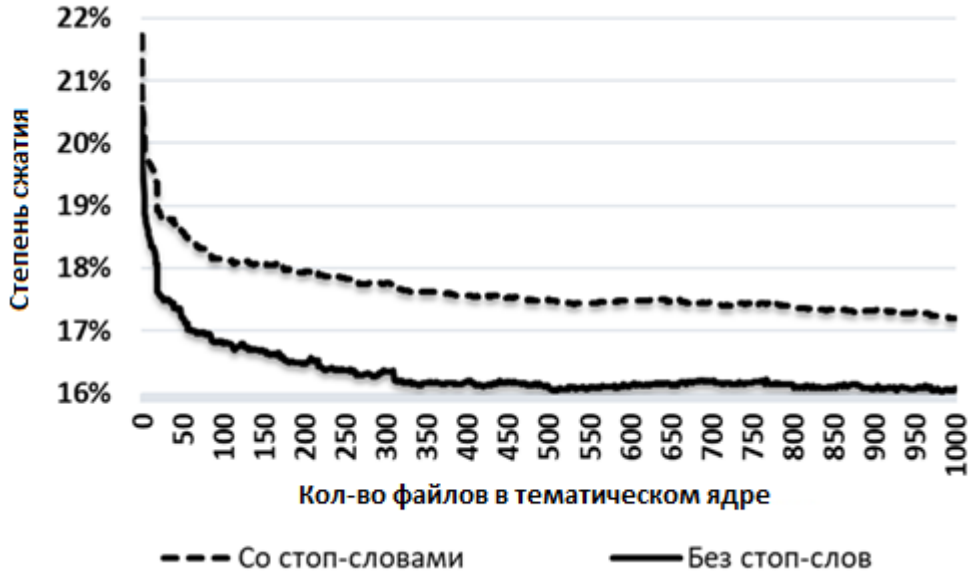


Рисунок 5 – Влияние стоп-слов на качество классификации

Для проведения экспериментов было выбрано 600 тестов (по 20 из каждой категории). В таблице 3 приведены результаты классификации статей с arXiv.org при ядрах, подобранными способами «Случайный выбор» и «Матрица сжатия». При ядрах, подобранных вторым способом, доля ошибок сократилась с 11 % до 8 %.

Таблица 3 – Количество ошибок при классификации текстов с arXiv.org

Код категории	Количество ошибок при ядрах, подобранных способом «Случайный выбор»	Количество ошибок при ядрах, подобранных способом «Матрица сжатия»
astro-ph.CO	2	2

Продолжение таблицы 3

Код категории	**Количество ошибок при ядрах, подобранных способом «Случайный выбор»**	**Количество ошибок при ядрах, подобранных способом «Матрица сжатия»**
astro-ph.GA	3	3
astro-ph.HE	2	2
cond-mat.dis-nn	2	4
cond-mat.stat-mech	3	1
cs.AI	3	6
cs.cr	0	2
cs.IT	2	2
cs.LO	2	0
cs.SE	2	0
gr-qc	2	0
hep-ex	1	4
hep-th	1	1
math.AG	0	0
math.CO	1	1
math.DG	0	2
math.FA	0	0
math.GR	0	0
math.PR	1	2
math.ST	0	0
math-ph	18	4
nucl-ex	4	0
nucl-th	1	2
physics.acc-ph	0	1
physics.atom-ph	3	1
physics.ins-det	7	3
physics.optics	0	1
physics.soc-ph	0	0
q-bio.BM	2	2
quant-ph	1	1
Всего	63	47
Проценты	11 %	8 %

Более того, ошибки с arXiv.org можно разделить на следующие категории:

1. Ложный выбор главной категории. К этому случаю относятся те ошибки, когда вместо главной категории определилась второстепенная.

2. Ложный выбор категории внутри области наук. К этому типу отнесены те ошибочно определенные тестовые файлы, у которых определилась другая категория из научной области;

3. Ложный выбор области наук.

В таблице 4 представлены результаты классификации тестов с arXiv.org с подобранными ядрами. Видно, что число ошибок по сравнению с произвольными ядрами сократилось примерно на 3 %. Второстепенная категория вместо главной определяется в 3 %; другая категория – в 4 %. Другая научная область определяется лишь в 1 %, причем, например, в cs.CR определяется math.PR, но в качестве второстепенной категории у этого теста указана другая математическая категория, которой нет в исследуемом списке.

Таблица 4 – Ошибки при классификации статей с arXiv.org при ядрах, сформированных способом «Матрица сжатия»

Код категории	**Всего ошибок**	**Ложный выбор главной категории**	**Ложный выбор категории внутри области наук**	**Ложный выбор области наук**
astro-ph.CO	2	2	0	0
astro-ph.GA	3	1	2	0
astro-ph.HE	2	1	1	0
cond-mat.dis-nn	4	3	1	0
cond-mat.stat-mech	1	0	1	0
cs.AI	6	1	4	1
cs.CR	2	0	1	1
cs.IT	2	1	0	1
cs.LO	0	0	0	0
cs.SE	0	0	0	0
gr-qc	0	0	0	0
hep-ex	4	0	4	0
hep-th	1	1	0	0
math.AG	0	0	0	0
math.CO	1	0	1	0
math.DG	2	0	2	0
math.FA	0	0	0	0
math.GR	0	0	0	0
math.PR	2	1	1	0

Продолжение таблицы 4

Код категории	Всего ошибок	Ложный выбор главной категории	Ложный выбор категории внутри области наук	Ложный выбор области наук
math.ST	0	0	0	0
math-ph	4	1	3	0
nucl-ex	0	0	0	0
nucl-th	2	2	0	0
physics.acc-ph	1	0	1	0
physics.atom-ph	1	0	1	0
physics.ins-det	3	1	2	0
physics.optics	1	1	0	0
physics.soc-ph	0	0	0	0
q-bio.BM	2	0	0	2
quant-ph	1	0	1	0
Итого	**47**	**16**	**26**	**5**
Доля от общего числа тестов	**8 %**	**3 %**	**4 %**	**1 %**

3.2. Результаты классификации полных русскоязычных научных текстов

В качестве массива русскоязычных полнотекстовых научных документов был выбран массив публикаций с научной электронной библиотеки «КиберЛенинка». Классификация статей в «КиберЛенинке» основана на Государственном рубрикаторе научно-технической информации (ГРНТИ). Научные публикации распределялись по более чем 80 категориям. В работе для проверки метода было выбрано 20 категорий («Астрономия»; «Автоматика»; «Биология»; «Экология»; «Экономика»; «Философия»; «Физика»; «Геофизика»; «География»; «Химия»; «История»; «Кибернетика»; «Литература»; «Математика»; «Медицина»; «Механика»; «Политика»; «Психология»; «Социология»; «Юридические науки»). Это обусловливалось тем, что в указанных категориях

было более чем 300 файлов, что позволило провести эксперименты с полученными данными.

Процесс подготовки данных состоял из следующих этапов:

1. Получение полных текстов в формате .pdf с веб-сайта cyberleninka.ru (таблица 5);
2. Извлечение текстового слоя из каждого .pdf-файла (PDF2Text Pilot);
3. Выбор только русскоязычных статей;
4. Удаление символов, получившихся при преобразовании математических формул, цифр, знаков препинания при помощи регулярного выражения;
5. Проведение стемминга – процесса приведения слов к начальной форме – при помощи консольной программы «Яндекс MyStem» [124]. Также из каждой публикации был удален список литературы. Таким образом, запрос выглядел следующим образом: «-e "/------/,+1d" -e "/^\s*Список литературы\s*$/,$d" <%%F > %%F.txt»;
6. Удаление стоп-слов;
7. Удаление из массивов текстов размером менее 10 кб.

Таблица 5 – Количество файлов, полученных с веб-сайта cyberleninka.ru, по категориям

Категория	**Количество файлов**
Автоматика	539
Астрономия	417
Биология	513
География	483
Геофизика	463
История	540
Кибернетика	568
Литература	545
Математика	586
Медицина	595
Механика	442
Политика	564
Психология	572
Социология	544

Продолжение таблицы 5

Категория	Количество файлов
Физика	476
Философия	540
Химия	496
Экология	523
Экономика	547
Юридические науки	547
Всего	**10 500**

Для проведения экспериментов с текстами из архива «КиберЛенинки» произвольным образом были отобраны 400 тестовых файлов, по 20 для каждой категории. Предварительный детальный анализ показал, что изначальная классификация на сайте содержала высокое число ошибок. В связи с этим при ядрах, сформированных методом «Случайный выбор», 47 % тестов было определено неправильно. При ядрах, сформированных методом «Матрица сжатия», количество ошибок сократилось более чем в 1,5 раза (таблица 6). В двух категориях («История» и «Литература») все тесты были определены полностью правильно. В остальных же случаях чаще всего определялась близкая категория: например, тесты «Математики» были отнесены к «Автоматике» или «Кибернетике», «Геофизики» – к «Географии», «Истории» – к «Политике» или «Юридическим наукам».

Таблица 6 – Количество ошибок при классификации текстов «КиберЛенинки»

Категория	Количество ошибок при ядрах, подобранных способом «Случайный выбор»	Количество ошибок при ядрах, подобранных способом «Матрица сжатия»
Автоматика	7	7
Астрономия	4	1
Биология	12	13
Экономика	17	5
Экология	16	6
Философия	8	4
Физика	8	9
Геофизика	7	4

Продолжение таблицы 6

Категория	**Количество ошибок при ядрах, подобранных способом «Случайный выбор»**	**Количество ошибок при ядрах, подобранных способом «Матрица сжатия»**
География	8	10
История	2	0
Кибернетика	13	14
Литература	1	0
Математика	13	10
Медицина	7	2
Механика	15	5
Химия	7	2
Политика	5	8
Психология	8	1
Социология	17	9
Юрид. науки	12	2
Всего	187	114
Проценты	47 %	28 %

3.3. Результаты классификации аннотаций публикаций

Зачастую при формировании тематических ядер не удается получить полный текст статьи, а доступна лишь ее аннотация. Более того, полные тексты статей, ввиду их объема, содержат много лишних фраз, что может привести к ошибкам при классификации, а также к увеличению времени работы алгоритма. Аннотации публикаций, наоборот, должны содержать только ключевые моменты, используемые в статье [125], что может облегчить автоматическую классификацию научных работ.

Источником аннотаций для исследования, проводимого в разделе 3.3., являлась ББД Scopus, где в качестве системы классификации используется ASJC. Классификация публикаций происходит только на уровне журнала, т. е. каждой публикации присваиваются все те же категории, что были у журнала. Это является

значительным недостатком, особенно для мультидисциплинарных журналов, так как создает «замусоренность» научных направлений публикациями, возможно, не имеющими к ним никакого отношения.

Процесс извлечения данных состоял из трех этапов:

1. Извлечение информации о названии журнала, eid публикации, цитировании, категории через Scival – аналитический инструмент компании Elsevier, основанный на данных Scopus;
2. Формирование файлов аннотаций путем получения их текстов через Scopus Abstract Retrieval API;
3. Удаление файлов аннотаций с отсутствующим текстом.

Данные были выбраны из 30 случайных категорий, у которых за 2009–2018 гг. было не менее 300 публикаций в одной категории с ненулевым числом цитирований (таблица 7).

Таблица 7 – Исследуемые области наук по уровням классификации

Направление	**Область**	**Категория**
Life Sciences	Agricultural and Biological Sciences	Animal Science and Zoology
Life Sciences	Agricultural and Biological Sciences	Aquatic Science
Life Sciences	Agricultural and Biological Sciences	Plant Science
Social Sciences	Arts and Humanities	History
Social Sciences	Arts and Humanities	Literature and Literary Theory
Life Sciences	Biochemistry, Genetics and Molecular Biology	Cell Biology
Life Sciences	Biochemistry, Genetics and Molecular Biology	Endocrinology
Social Sciences	Business, Management and Accounting	Marketing
Physical Sciences	Chemical Engineering	Catalysis
Physical Sciences	Chemistry	Inorganic Chemistry
Physical Sciences	Chemistry	Organic Chemistry
Physical Sciences	Computer Science	Artificial Intelligence

Продолжение таблицы 7

Направление	Область	Категория
Physical Sciences	Computer Science	Computer Vision and Pattern Recognition
Physical Sciences	Computer Science	Hardware and Architecture
Physical Sciences	Earth and Planetary Sciences	Geology
Physical Sciences	Earth and Planetary Sciences	Oceanography
Physical Sciences	Mathematics	Algebra and Number Theory
Physical Sciences	Mathematics	Geometry and Topology
Physical Sciences	Mathematics	Logic
\Physical Sciences	Mathematics	Numerical Analysis
Physical Sciences	Mathematics	Statistics and Probability
Health Sciences	Medicine	Ophthalmology
Health Sciences	Medicine	Surgery
Life Sciences	Pharmacology, Toxicology and Pharmaceutics	Pharmocology
Physical Sciences	Physics and Astronomy	Astronomy and Astrophysics
Physical Sciences	Physics and Astronomy	Condensed Matter Physics
Physical Sciences	Physics and Astronomy	Nuclear and High Energy Physics
Social Sciences	Psychology	Social Psychology
Social Sciences	Social Sciences	Library and Information Sciences
Social Sciences	Social Sciences	Sociology and Political Science

3.3.1. Классификация тестовых файлов с одной категорией

При классификации тестовых файлов с одной категорией введены три типа ошибок определения категории:

- I тип. Ложный выбор категории внутри области наук. Например, вместо категории Aquatic Science определилась Plant Science из той же области Agricultural and Biological Sciences;

- II тип. Ложный выбор области наук внутри научного направления. Например, вместо нужной категории Cell Biology из области Biochemistry, Genetics and Molecular Biology определилась категория Pharmocology из области Pharmacology, Toxicology and Pharmaceutics. При этом общее научное направление Life Sciences сохранилось;
- III тип. Ложный выбор научного направления. Например, вместо научного направления Physical Sciences определилось Social Sciences.

Классификация осуществлялась с использованием ядер двух типов:

- Случайный выбор,
- Рейтинг цитирования.

В обоих случаях использовался один и тот же набор произвольных тестовых файлов, для каждой из 30 категорий было отобрано по 20 тестовых файлов – суммарно 600 тестов.

Также было проверено влияние на качество классификации наличия стоп-слов и названий издательств, присутствующих в текстах аннотаций (например, «© 2009 The American Physical Society.»). Создание ядер и подготовка тестовых файлов были выполнены как с оригинальными текстами аннотаций, так и с удалением стоп-слов (например, «always», «every», «just» и т.п.), заменой заглавных букв на строчные и удалением всех символов, кроме цифр, букв, и следующих знаков препинания: «.», «!», «?», «:», «,», «-».

Результаты классификации тестовых файлов с одной категорией при произвольных ядрах по типам ошибок приведены в Таблица 8.

Таблица 8 – Результаты классификации тестовых файлов при произвольных ядрах с различным числом цитирования

Категория	**Общее количество тестов**	**Количество ошибок**	**Ошибки I типа**	**Ошибки II типа**	**Ошибки III типа**
Algebra and Number Theory	20	7	5	2	0

Продолжение таблицы 8

Категория	Общее количество тестов	Количество ошибок	Ошибки I типа	Ошибки II типа	Ошибки III типа
Animal Science and Zoology	20	1	0	1	0
Aquatic Science	20	12	11	0	1
Artificial Intelligence	20	8	6	2	0
Astronomy and Astrophysics	20	1	1	0	0
Catalysis	20	5	0	5	0
Cell Biology	20	1	0	0	1
Computer Vision and Pattern Recognition	20	3	3	0	0
Condensed Matter Physics	20	11	2	4	5
Endocrinology	20	3	1	2	0
Geology	20	0	0	0	0
Geometry and Topology	20	11	9	2	0
Hardware and Architecture	20	0	0	0	0
History	20	8	1	7	0
Inorganic Chemistry	20	13	4	1	8
Library and Information Sciences	20	12	2	8	2
Literature and Literary Theory	20	11	5	5	1
Logic	20	2	0	1	1
Marketing	20	2	0	2	0
Nuclear and High Energy Physics	20	4	4	0	0
Numerical Analysis	20	0	0	0	0
Oceanography	20	11	0	1	10
Ophthalmology	20	9	7	0	2
Organic Chemistry	20	3	0	2	1
Pharmocology	20	18	0	18	0
Plant Science	20	20	13	7	0

Окончание таблицы 8

Категория	**Общее количество тестов**	**Количество ошибок**	**Ошибки I типа**	**Ошибки II типа**	**Ошибки III типа**
Social Psychology	20	1	0	1	0
Sociology and Political Science	20	8	0	7	1
Statistics and Probability	20	6	0	2	4
Surgery	20	1	0	0	1
Общее количество	600	192	74	80	38
Доля от общего количества	100 %	32 %	12 %	13 %	6 %

Доля ошибочно определенных тестовых файлов составила 32 % от общего количества (192 из 600 тестов).

Чаще всего определение неверного научного направления происходит из-за категорий, близких по терминологии. Например, такими категориями являются Aquatic Science и Oceanography. Но иногда характер ошибки определить не удается, например, в случае с публикацией с eid=2-s2.0-67651018249 из категории Condensed Matter Physics, вместо которой определилась категория Marketing. Визуально определить причину по тексту аннотации не удалось (рисунок 6):

Two-particle dispersion is of central importance to a wide range of natural and industrial applications. It has been an active area of research since Richardson's (1926) seminal paper. This review emphasizes recent results from experiments, high-end direct numerical simulations, and modern theoretical discussions. Our approach is complementary to Sawford's (2001), whose review focused primarily on stochastic models of pair dispersion. We begin by reviewing the theoretical foundations of relative dispersion, followed by experimental and numerical findings for the dissipation subrange and inertial subrange. We discuss the findings in the context of the relevant theory for each regime. We conclude by providing a critical analysis of our current understanding and by suggesting paths toward further progress that take full advantage of exciting developments in modern experimental methods and peta-scale supercomputing. Copyright © 2009 by Annual Reviews. All right reserved. All rights reserved.

Рисунок 6 – Аннотация публикации с eid=2-s2.0-67651018249

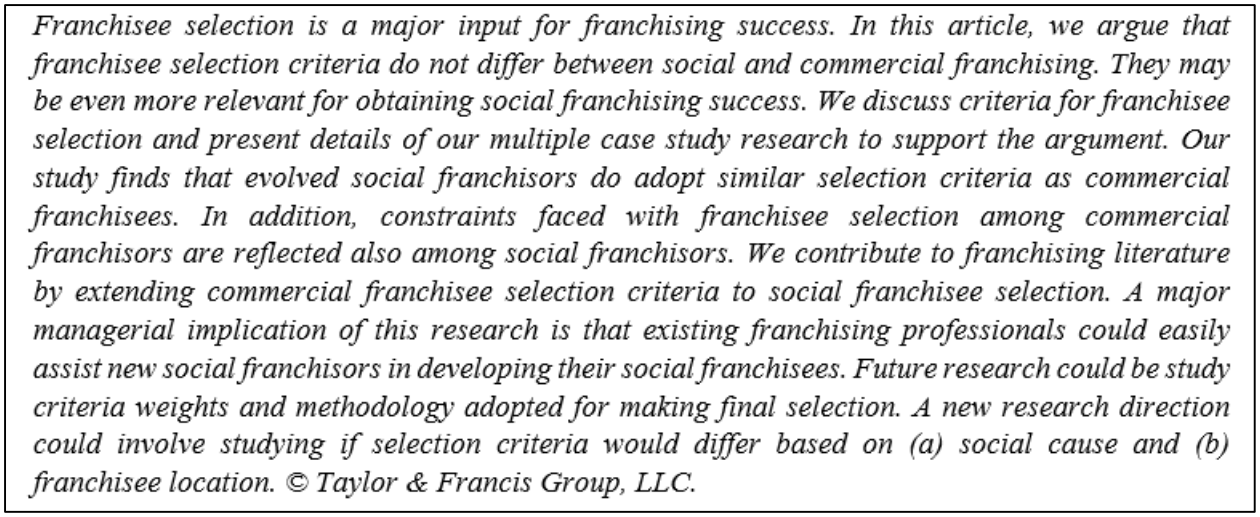
Нами было проведено попарное сжатие файлов из ядер этих двух категорий и тестового файла с eid=2-s2.0-67651018249. Почти по всем отдельным 100 файлам из ядра категории Marketing тест с eid=2-s2.0-67651018249 показывает лучшее сжатие. При этом средний нормированный коэффициент сжатия тестового файла с категорией Condensed Matter Physics составляет 9,76%, а с категорией Marketing – 9,13%.

В топ-10 файлов, с которыми произошло наилучшее сжатие этого теста, вошли 3 файла из категории Condensed Matter Physics и 7 из категории Marketing (таблица 9).

Таблица 9 – Топ-10 файлов, с которыми произошло наилучшее сжатие исследуемого теста с eid=2-s2.0-67651018249 категории Condensed Matter Physics

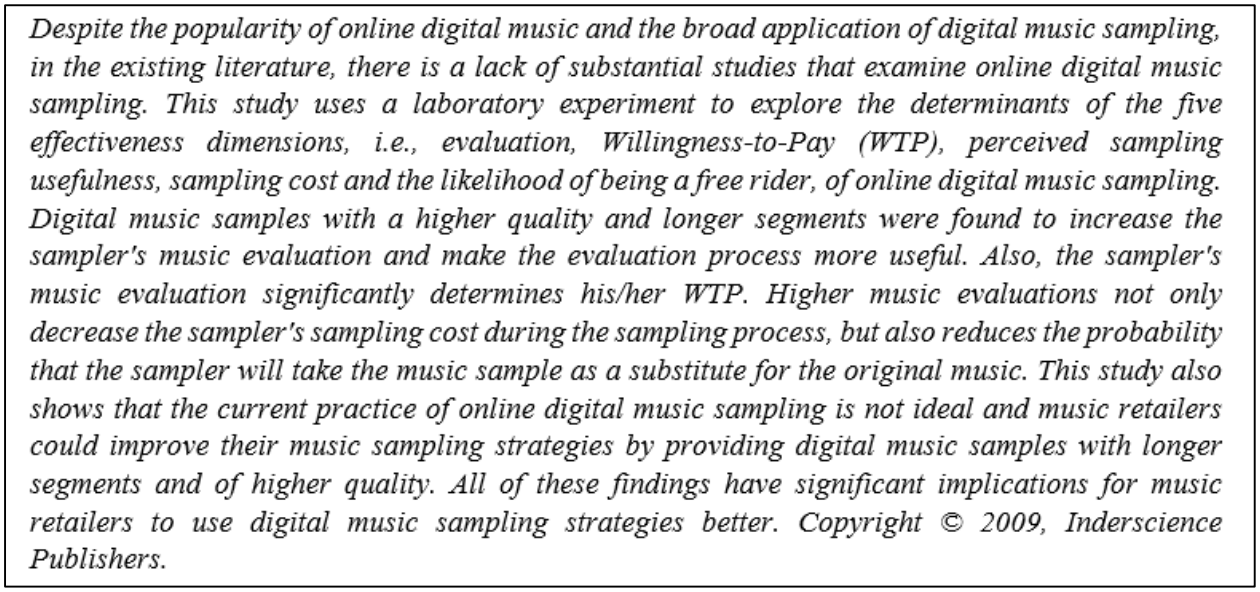
Неверно определившийся тест из «Condensed Matter Physics»	**Идентификатор файла**	**Категория файла**	**Нормированный процент сжатия**
2-s2.0-67651018249	2-s2.0-70350534620	Condensed Matter Physics	0,00%
2-s2.0-67651018249	2-s2.0-80052140988	Marketing	1,17%
2-s2.0-67651018249	2-s2.0-67149130202	Marketing	1,47%
2-s2.0-67651018249	2-s2.0-79960889541	Marketing	2,70%
2-s2.0-67651018249	2-s2.0-70449090433	Condensed Matter Physics	2,94%
2-s2.0-67651018249	2-s2.0-70449127336	Condensed Matter Physics	3,16%
2-s2.0-67651018249	2-s2.0-79959944133	Marketing	3,20%
2-s2.0-67651018249	2-s2.0-67149101079	Marketing	3,49%
2-s2.0-67651018249	2-s2.0-78650307261	Marketing	3,88%
2-s2.0-67651018249	2-s2.0-78751585438	Marketing	3,90%

Тексты двух аннотаций категории Marketing, с которыми у тестового файла произошло лучшее попарное сжатие, приведены на рисунках 7, 8. У этих файлов полностью различается терминология как между собой, так и с исследуемым тестовым файлом. Таким образом, результаты позволяют предположить, что при определении категории тестового файла с eid=2-s2.0-67651018249 категории Condensed Matter Physics ошибка может быть связана с работой метода.



Franchisee selection is a major input for franchising success. In this article, we argue that franchisee selection criteria do not differ between social and commercial franchising. They may be even more relevant for obtaining social franchising success. We discuss criteria for franchisee selection and present details of our multiple case study research to support the argument. Our study finds that evolved social franchisors do adopt similar selection criteria as commercial franchisees. In addition, constraints faced with franchisee selection among commercial franchisors are reflected also among social franchisors. We contribute to franchising literature by extending commercial franchisee selection criteria to social franchisee selection. A major managerial implication of this research is that existing franchising professionals could easily assist new social franchisors in developing their social franchisees. Future research could be study criteria weights and methodology adopted for making final selection. A new research direction could involve studying if selection criteria would differ based on (a) social cause and (b) franchisee location. © Taylor & Francis Group, LLC.

Рисунок 7 – Аннотация публикации категории Marketing с eid=2-s2.0-80052140988



Despite the popularity of online digital music and the broad application of digital music sampling, in the existing literature, there is a lack of substantial studies that examine online digital music sampling. This study uses a laboratory experiment to explore the determinants of the five effectiveness dimensions, i.e., evaluation, Willingness-to-Pay (WTP), perceived sampling usefulness, sampling cost and the likelihood of being a free rider, of online digital music sampling. Digital music samples with a higher quality and longer segments were found to increase the sampler's music evaluation and make the evaluation process more useful. Also, the sampler's music evaluation significantly determines his/her WTP. Higher music evaluations not only decrease the sampler's sampling cost during the sampling process, but also reduces the probability that the sampler will take the music sample as a substitute for the original music. This study also shows that the current practice of online digital music sampling is not ideal and music retailers could improve their music sampling strategies by providing digital music samples with longer segments and of higher quality. All of these findings have significant implications for music retailers to use digital music sampling strategies better. Copyright © 2009, Inderscience Publishers.

Рисунок 8 – Аннотация публикации категории Marketing
с eid=2-s2.0-67149130202

Рассмотрим результаты классификации тестовых файлов с одной категорией, ядра для которых были подобраны методом «Рейтинг цитирования» (таблица 10).

Таблица 10 – Результаты классификации тестовых файлов с одной категорией при ядрах, подобранных способом «Рейтинг цитирования»

Категория	**Общее количество тестов**	**Количество ошибок**	**Ошибки I типа**	**Ошибки II типа**	**Ошибки III типа**
Algebra and Number Theory	20	8	7	1	0
Animal Science and Zoology	20	7	6	1	0
Aquatic Science	20	2	2	0	0
Artificial Intelligence	20	5	2	2	1
Astronomy and Astrophysics	20	0	0	0	0
Catalysis	20	4	0	4	0
Cell Biology	20	4	1	3	0
Computer Vision and Pattern Recognition	20	3	3	0	0
Condensed Matter Physics	20	2	0	2	0
Endocrinology	20	1	0	1	0
Geology	20	0	0	0	0
Geometry and Topology	20	3	3	0	0
Hardware and Architecture	20	0	0	0	0
History	20	4	2	1	1
Inorganic Chemistry	20	3	0	3	0
Library and Information Sciences	20	3	1	1	1
Literature and Literary Theory	20	2	1	1	0
Logic	20	3	2	1	0

Продолжение таблицы 10

Категория	**Общее количество тестов**	**Количество ошибок**	**Ошибки I типа**	**Ошибки II типа**	**Ошибки III типа**
Marketing	20	1	0	1	0
Nuclear and High Energy Physics	20	3	3	0	0
Numerical Analysis	20	0	0	0	0
Oceanography	20	4	0	0	4
Ophthalmology	20	0	0	0	0
Organic Chemistry	20	1	0	0	1
Pharmocology	20	2	0	2	0
Plant Science	20	2	1	1	0
Social Psychology	20	2	0	2	0
Sociology and Political Science	20	2	1	0	1
Statistics and Probability	20	1	0	1	0
Surgery	20	0	0	0	0
Общее количество	600	72	35	28	9
Доля от общего количества		12 %	6 %	5 %	2 %

Использование ядер, подобранных способом «Рейтинг цитирования», улучшило результаты классификации на 20 %. Число ошибок III типа уменьшилось в три раза. В основном такие ошибки возникали из-за находящихся в разных научных направлениях категорий или файлов, в которых применяются схожие термины. Например, у теста из категории Sociology and Political Science неверно определилась категория Aquatic Science, но в тексте этой аннотации применяется много терминов, используемых в категории Aquatic Science (рисунок 9).

*This paper seeks to understand how the Brazilian Amazon, which many thought unsuitable for **agricultural development**, has yielded to a dynamic cattle **economy** in only a few decades. It does so by embedding the Thunian model of **location** rents within the regime of **capital accumulation** that has driven the Brazilian **economy** since the mid-20th century. The paper addresses policies that have created location rents in Amazônia, the effect of these rents on **land managers**, and the **spatial implications** of their **behavior** on **forests**. Thus, the paper connects macro-**processes** and **structures** to **agents** on the **ground**, in providing a political **ecological explanation** relevant to **land** change science. The policy discussion focuses on **reductions** in transportation **costs**, improvements in **animal health**, and monetary and **trade** reforms. To illustrate the impact of policy, the paper presents data on the **geography** of Amazonian herd **expansion**, on the growth of Amazonian exports, and on the profitability of the **region**'s cattle **economy**. It follows the empirical presentation with more abstract consideration of the spatial relations between cattle ranching and **soy farming**, and implications for **deforestation**. The paper concludes on a speculative note by considering the likelihood of **forest transition** in the **region**, given the transformation of Amazônia into a global **resource** frontier. © 2008 Elsevier Ltd. All rights reserved.*

Рисунок 9 – Аннотация теста с eid=2-s2.0-70449527784 из категории Sociology and Political Science. **Жирным** выделены термины, часто встречающиеся в категории Aquatic Science

Некоторые неверно определенные тесты, возможно, связаны с неверной изначальной классификацией. Так, в случае категории Library and Information Sciences тестовый файл отнесся к категории Artificial Intelligence. Текст аннотации содержит значительное количество терминов, характерных для области Computer Science (рисунок 10Рисунок *10*).

This chapter provides a tutorial overview of distributed optimization and game theory for decision-making in networked systems. We discuss properties of first-order methods for smooth and non-smooth convex optimization, and review mathematical decomposition techniques. A model of networked decision-making is introduced in which a communication structure is enforced that determines which nodes are allowed to coordinate with each other, and several recent techniques for solving such problems are reviewed. We then continue to study the impact of noncooperative games, in which no communication and coordination are enforced. Special attention is given to existence and uniqueness of Nash equilibria, as well as the efficiency loss in not coordinating nodes. Finally, we discuss methods for studying the dynamics of distributed optimization algorithms in continuous time. © 2010 Springer London.

Рисунок 10 – Текст аннотации публикации с eid=2-s2.0-77958562700 из категории Library and Information Science

В ряде случаев определения неверной категории закономерностей выявлено не было. Например, тестовый файл из History неверно отнесся к Condensed Matter Physics (рисунок 11).

The horse skeleton found in the autumn of 1958 at the fortress of Buhen in northern Sudan has become one of the most prominent, but also one of the most enigmatic equid remains from the second millennium BC: Firstly, because of its assumed early date of c. 1675 BC, deduced by W.B. Emery after analysing the stratigraphical data, This - according to our knowledge at the time - being several decades before the oldest known equid remains in Egypt. Secondly, because of wear on the lower left second premolar (LP2), which has led to the conclusion that it was most probably caused by bit-wear. Since the 1960s, both conclusions have been subject to criticism. The purpose of this study is to provide a review of the history of research and reception of the Buhen horse in its interdisciplinary context over the last fifty years with the result that only modern scientific techniques might be able to solve some of the outstanding questions. © 2009 Brill.

Рисунок 11 – Тестовый файл с eid=2-s2.0-77951062083 из категории History

Стоит отметить, что неверно определившийся файл с eid=2-s2.0-67651018249 (рисунок 6) при ядрах, подобранных способом «Рейтинг цитирования», определился верно. Таким образом, состав ядра оказывает большое влияние на качество классификации.

3.3.2. Влияние на классификацию аннотаций стоп-слов и названий издательств

Для изучения влияния присутствия названий издательств в аннотациях на качество классификации использовались ядра, подобранные способом «Рейтинг цитирования».

В таблице 11 приведено сравнение качества классификации аннотаций со стоп-словами и названиями издательств и без них. При удалении названий издательств количество ошибок возросло на 3 %. Почти в два раза увеличилось количество ошибок в категориях History, Geometry and Topology, Literature and Literary Theory, Sociology and Political Science. Возможно, это связано с тем, что

чаще всего высокоцитируемые публикации печатаются в одних издательствах, в названиях которых указаны важные термины для категории. Например, в одном из неверно определившихся после удаления издательства тесте раньше встречалась следующая строка: «© 2010 English Literary Renaissance Inc. Published by Blackwell Publishing Ltd.».

При удалении стоп-слов из аннотаций, где присутствовали названия издательств, количество ошибок уменьшилось до 11 %. Однако это уменьшение произошло неравномерно по всем категориям: если в категории Cell Biology удаление стоп-слов повлияло положительно на качество классификации, то в категории Literature and Literary Theory количество ошибок увеличилось в два раза.

В случае удаления как стоп-слов, так и названий издательств, количество ошибок увеличилось до 16 %. Аналогично предыдущему случаю на ряд категорий удаление стоп-слов и названий издательств повлияло положительно, тогда как другие стали определяться ошибочно. Так, например, в категории Geometry and Topology тест с eid= 2-s2.0-84055189802 при удалении стоп-слов определился верно, а тест с eid= 2-s2.0-77956268008, который раньше определялся верно, теперь отнесся к категории Numerical Analysis. Возможно, это связано с тем, что при удалении стоп-слов длина аннотации уменьшается.

Таблица 11 – Влияние стоп-слов и присутствия названий издательств на качество классификации

Категория	**Общее количество тестов**	**Коли-чество ошибок**	**Количество ошибок со стоп-словами, без названий издательств**	**Количество ошибок без стоп-слов, с названиями издательств**	**Количест-во ошибок без стоп-слов и названий изда-тельств**
Algebra and Number Theory	20	8	8	8	7

Продолжение таблицы 11

Категория	**Общее количество тестов**	**Количество ошибок**	**Количество ошибок со стоп-словами, без названий издательств**	**Количество ошибок без стоп-слов, с названиями издательств**	**Количество ошибок без стоп-слов и названий издательств**
Animal Science and Zoology	20	7	7	6	8
Aquatic Science	20	2	2	1	2
Artificial Intelligence	20	5	6	6	7
Astronomy and Astrophysics	20	0	0	0	0
Catalysis	20	4	3	5	5
Cell Biology	20	4	3	2	3
Computer Vision and Pattern Recognition	20	3	3	1	3
Condensed Matter Physics	20	2	2	1	1
Endocrinology	20	1	2	1	2
Geology	20	0	0	0	0
Geometry and Topology	20	3	6	2	6
Hardware and Architecture	20	0	0	0	0
History	20	4	7	4	8

Продолжение таблицы 11

Категория	**Общее количество тестов**	**Коли-чество ошибок**	**Количество ошибок со стоп-словами, без названий издательств**	**Количество ошибок без стоп-слов, с названиями издательств**	**Количест-во ошибок без стоп-слов и названий изда-тельств**
Inorganic Chemistry	20	3	5	3	5
Library and Information Sciences	20	3	4	3	5
Literature and Literary Theory	20	2	4	4	4
Logic	20	3	3	3	4
Marketing	20	1	1	1	1
Nuclear and High Energy Physics	20	3	3	2	5
Numerical Analysis	20	0	1	0	1
Oceanography	20	4	4	3	5
Ophthalmology	20	0	1	0	1
Organic Chemistry	20	1	2	1	2
Pharmocology	20	2	3	2	2
Plant Science	20	2	3	2	2
Social Psychology	20	2	1	1	3
Sociology and Political Science	20	2	3	2	4

Окончание таблицы 11

Категория	**Общее количество тестов**	**Коли-чество ошибок**	**Количество ошибок со стоп-словами, без названий издательств**	**Количество ошибок без стоп-слов, с названиями издательств**	**Количест-во ошибок без стоп-слов и названий изда-тельств**
Statistics and Probability	20	1	2	1	2
Surgery	20	0	0	0	0
Общее количество	600	72	89	65	98
Доля от общего количества		12 %	15 %	11 %	16 %

Таким образом, отсутствие названий издательств в текстах аннотаций негативно влияет на качество классификации. Про влияние стоп-слов однозначного вывода сделать нельзя.

3.3.3. Классификация тестовых файлов с несколькими категориями

Для классификации тестовых файлов с несколькими категориями использовались только ядра, подобранные методом «Рейтинг цитирования».
Отбор 20 тестов осуществлялся произвольным образом из публикаций, у которых по меньшей мере две категории совпадало с категориями из таблицы 7. Суммарно было отобрано 600 тестов.

Для анализа результатов классификации введем следующие группы:

- У тестового файла верно определилось не менее 50 % указанных категорий. Например, у теста было указано четыре категории: Algebra and Number Theory, Numerical Analysis, Geometry and Topology, Discrete Mathematics and Combinatorics. В число исследуемых нами категорий входят только первые

три. Соответственно, чтобы попасть в эту группу, у тестового файла должны определиться минимум две из трёх первых. Верно определенными считаются категории, у которых «нормированный коэффициент сжатия» (процент сжатия за вычетом минимального процента сжатия по всем категориям), меньше, чем минимальный процент сжатия со всеми категориями*0,50 % (при большем пороге количество ошибок почти не изменялось). В качестве примера такого расчета рассмотрим два тестовых файла с категориями Algebra and Number Theory и Geometry and Topology (таблица 12). Минимальный процент сжатия у первого теста определился с категорией Algebra and Number Theory. При этом если в качестве порогового значения выбирать не только минимальный процент сжатия, а минимальный процент сжатия со всеми категориями*0,50 %, то с этим тестом правильно будет определена и вторая указанная категория: Geometry and Topology. У второго же тестового файла была определена только категория Geometry and Topology.

Таблица 12 – Пример расчета нормированного коэффициента сжатия

Области теста	**Algebra and Number Theory**	**Geometry and Topology**	**Min значение**	**Algebra and Number Theory**	**Geometry and Topology**
Algebra and Number Theory, Geometry and Topology	***26,70 %***	26,80 %	26,70 %	OK	OK
Algebra and Number Theory, Geometry and Topology	33,25 %	***32,85 %***	32,85 %	missed	OK

- У тестового файла определилась хотя бы одна из указанных категорий

- Все категории тестового файла определились неверно. К этому случаю будут отнесены те тестовые файлы, у которых определилась какая-то другая категория

При этом один тестовый файл может относиться только к одной из этих групп.

Результаты классификации тестовых файлов с несколькими категориями приведены в таблице 13.

Таблица 13 – Результаты классификации файлов с несколькими категориями

Группа теста	**Количество тестов**	**Доля от 600 тестов**
Определилось не менее 50% категорий	413	69 %
Определилась хотя бы одна категория	47	8 %
Все категории определились неверно	140	23 %

23 % (140 из 600) тестовых файлов определились ошибочно. При этом неверно определилось научное направление у 6 % (37 из 600) тестов. Стоит отметить, что в некоторых случаях эта ошибка возникала из-за категорий, близких по терминологии, но принадлежащим разным научным направлениям. Например, категория Aquatic Science из направления Life Sciences и категория Oceanography из Physical Sciences. В качестве примера приведем тестовый файл с eid= 2-s2.0-57649228732, у которого указаны две категории: Aquatic Science и Plant Science. Метод определил категорию Oceanography. Текст аннотации приведен на рисунке 12.

В других случаях ошибки характер ошибки определить не удалось. Так, у тестового файла с eid= 2-s2.0-79451471007 вместо категорий Library and Information Sciences и History из направления Social Sciences определилась категория Aquatic Science научного направления Life Sciences (рисунок 13).

In California, the toxic algal species of primary concern are the dinoflagellate Alexandrium catenella and members of the pennate diatom genus Pseudo-nitzschia, both producers of potent neurotoxins that are capable of sickening and killing marine life and humans. During the summer of 2004 in Monterey Bay, we observed a change in the taxonomic structure of the phytoplankton community-the typically diatom-dominated community shifted to a red tide, dinoflagellate-dominated community. Here we use a 6-year time series (2000-2006) to show how the abundance of the dominant harmful algal bloom (HAB) species in the Bay up to that point, Pseudo-nitzschia, significantly declined during the dinoflagellate-dominated interval, while two genera of toxic dinoflagellates, Alexandrium and Dinophysis, became the predominant toxin producers. This change represents a shift from a genus of toxin producers that typically dominates the community during a toxic bloom, to HAB taxa that are generally only minor components of the community in a toxic event. This change in the local HAB species was also reflected in the toxins present in higher trophic levels. Despite the small contribution of A. catenella to the overall phytoplankton community, the increase in the presence of this species in Monterey Bay was associated with an increase in the presence of paralytic shellfish poisoning (PSP) toxins in sentinel shellfish and clupeoid fish. This report provides the first evidence that PSP toxins are present in California's pelagic food web, as PSP toxins were detected in both northern anchovies (Engraulis mordax) and Pacific sardines (Sardinops sagax). Another interesting observation from our data is the co-occurrence of DA and PSP toxins in both planktivorous fish and sentinel shellfish. We also provide evidence, based on the statewide biotoxin monitoring program, that this increase in the frequency and abundance of PSP events related to A. catenella occurred not just in Monterey Bay, but also in other coastal regions of California. Our results demonstrate that changes in the taxonomic structure of the phytoplankton community influences the nature of the algal toxins that move through local food webs and also emphasizes the importance of monitoring for the full suite of toxic algae, rather than just one genus or species. © 2008 Elsevier B.V.

Рисунок 12 – Аннотация публикации с eid= 2-s2.0-57649228732

Current records management methodologies and practices suffer from an inadequate understanding of the 'human activity systems' where records managers operate as 'mediators' between a number of complex and interacting factors. Although the records management and archival literature recognizes that managing the active life of the records is fundamental to their survival as meaningful evidence of activities, the context where the records are made, captured, used, and selectively retained is not explored in depth. In particular, the various standards, models, and functional requirement lists, which occupy a vast portion of that literature, especially in relation to electronic records, do not seem to be capable of framing records-related 'problems' in ways that account for their dynamic and multiform nature. This paper introduces the idea that alternative, 'softer' approaches to the analysis of organizational functions, structures, agents, and artifacts may usefully complement the 'hard', engineering-like approaches typically drawn on by information and records specialists. Three interrelated theoretical and methodological frameworks-namely, Soft Systems Methodology, Adaptive Structuration Theory, and Genre Theory-are discussed, with the purpose of highlighting their contributions to our understanding of the records context. © 2010 Springer Science+Business Media B.V.

Рисунок 13 – Аннотация публикации с eid= 2-s2.0-79451471007

3.3.4. Влияние количества категорий на качество классификации

В работе [71] авторы приходят к выводу, что количество категорий влияет на классификацию, и если объединить категории со схожими терминами в одну, то качество классификации улучшится.

Для оценки влияния количества категорий на качество классификации аннотаций был проведен эксперимент с последовательным увеличением количества рассматриваемых категорий с 5 до 30, с шагом в 5 (5, 10, 15, 20, 25, 30). При этом в первые пять категорий вошли категории с наибольшим терминологическим различием: Algebra and Number Theory, Computer Vision and Pattern Recognition, Condensed Matter Physics, Literature and Literary Theory и Surgery.

Для тестовых файлов были отобраны случайным образом по 20 публикаций из первых пяти категорий (всего 100 файлов).

На рисунке 14 изображена зависимость количества ошибок от количества ядер, участвующих в классификации. Результаты показывают, что при расширении числа категорий также увеличивается и число ошибок.

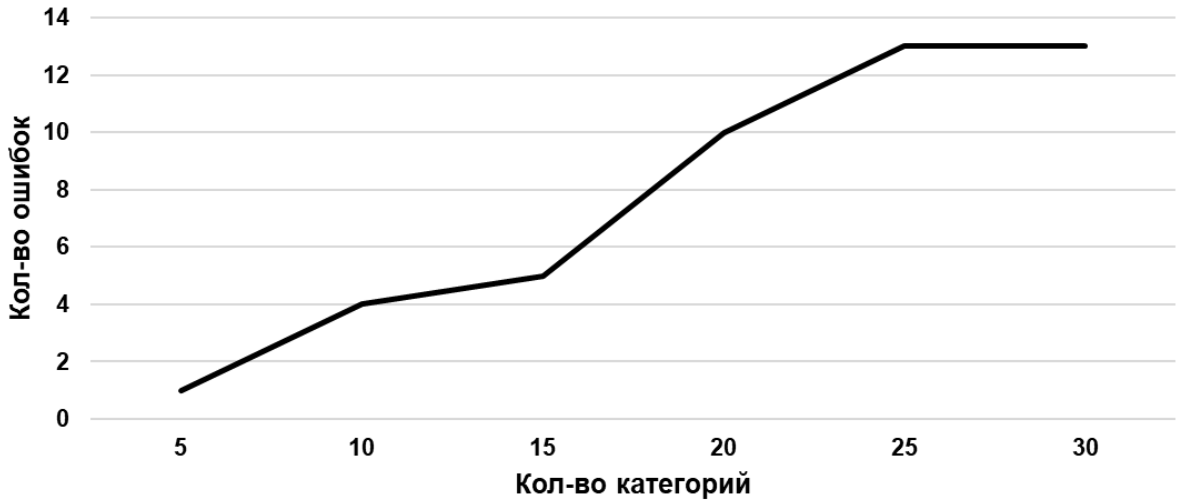


Рисунок 14 – Зависимость качества классификации от количества категорий

Если при пяти ядрах ошибочно определился только один тест из категории Algebra and Number Theory, то при 30 категориях количество ошибок возросло до 13. При этом как при пяти5, так и при 30 ядрах безошибочно определялись тесты из категории Surgery. Это связано с тем, что среди оставшихся 29 категорий не было терминологически близкой к этой категории.

Таким образом, изначальная выборка количества и состава категорий влияет на точность классификации. Чем меньше категорий участвуют при классификации и чем больше терминологическое различие между ними, тем выше ее качество.

3.3.5. Ограничения применения метода на основе сжатия данных к классификации аннотаций публикаций, индексируемых в Scopus

Рассмотрим подробнее недостатки, присутствующие в классификации ББД Scopus [103, 126]:

1. Классификация в Scopus происходит на уровне изданий, а не для каждой отдельной публикации;
2. В ASJC присутствуют близкие как по названию, так и по терминологии категории, в большинстве случаев находящиеся в разных научных областях: категории Language and Linguistics и Linguistics and Language областей Arts and Humanities и Social Sciences соответственно, две категории Archaeology в этих же самых областях, три категории Pharmacology в областях Nursing, Pharmacology, Toxicology and Pharmaceutics и Medicine и др.;
3. В ASJC присутствуют категории, названия которых содержат в себе «general» или «(all)» и «(miscellaneous)». Более того, иногда журналам присваиваются сразу две эти категории.

Как было показано выше, для метода классификации, основанного на сжатии данных (как и для других методов классификации, основанных на лексической

близости), важен набор терминов, используемых в текстах. При использовании публикаций системы классификации ASJC в качестве обучающей выборки неоднозначность категорий, присутствующих в них, может значительно ухудшить качество классификации.

Метод выбора ядер «Рейтинг цитирования», использование которого значительно улучшило качество классификации, может быть неприменим при создании ядер по всем 333 категориям ББД Scopus. Это связано с журнальной классификацией ББД Scopus, и из-за того, что многие журналы являются политематическими, для 333 категорий формирование ядер из публикаций, имеющих только одну категорию, может стать задачей не только трудоемкой, но и полностью невыполнимой. Формирование ядер через аналитический инструмент SciVal, как было проделано в предыдущих разделах, для всех 333 категорий также не представляется возможным: SciVal допускает выгрузку только первых 20 000 результатов в формате .csv или .xls, высылаемых на указанный при регистрации email. Более того, в выгруженных через SciVal данных отсутствует текст аннотаций, что требует применения дополнительных этапов формирования ядра.

Таким образом, метод извлечения данных, предложенный в разделе 3.3., может быть применен только для ограниченного числа категорий, а формирование ядер для большего количества категорий требует автоматического подхода.

Для анализа возможности применения метода классификации, основанного на сжатии данных, ко всем 333 категориям оценим представленность каждой из категорий.

Из-за того, что классификация публикаций в ББД Scopus происходит на журнальном уровне, оценку представленности каждой категории в исследуемой области наук можно проводить не по выгрузке каждой публикации отдельно, а по анализу тематик журналов из этой области.

Процесс получения данных происходил в три этапа:

1. Выгрузка списка журналов за 2019 года. Всего 39 743 журнала;

2. Выделение журналов с единственной категорией. Всего 17 050 журналов имели одну категорию;

3. Для каждого журнала при помощи Scopus Serial Title API по ISSN выгрузка суммарного количества его публикаций за все годы существования журнала.

Последний этап был проведен из-за отсутствия в списке журналов сведений о количестве публикаций. Информация была найдена в 7917 журналах. Из 9133 ненайденных журналов статус у 8875 журналов в списке был отмечен как Inactive.

Оценка проводилась только для источников типа Journal, Book Series и Trade Journal в связи с тем, что тематика сборников материалов конференций зачастую достаточно многообразна и не вносила значительной погрешности в проводимый анализ, однако увеличивала трудозатраты для проведения эксперимента.

Также для оценки представленности категорий при помощи SciVal были выгружены списки публикаций типа Article по рейтингу убывания числа цитирований за 2009–2018 гг. В дальнейшем для краткости будем обозначать этот рейтинг «рейтинг SciVal».

Рассмотрим, насколько полно представлены категории в базе данных Scopus на примере некоторых крупных областей наук.

Пример 1. Категория Mathematics

На рисунке 15 показано, что большинство журналов области Mathematics относятся к Mathematics (all). Наименьшее число журналов (по две на каждую категорию) относятся к областям – Theoretical Computer Science, Control and Optimization, Numerical Analysis. Полностью отсутствуют журналы, относящиеся только к категории Mathematical Physics.

Рассмотрим подробнее журналы категорий Numerical Analysis, Theoretical Computer Science и Control and Optimization, у которых была указана только одна из этих категорий.

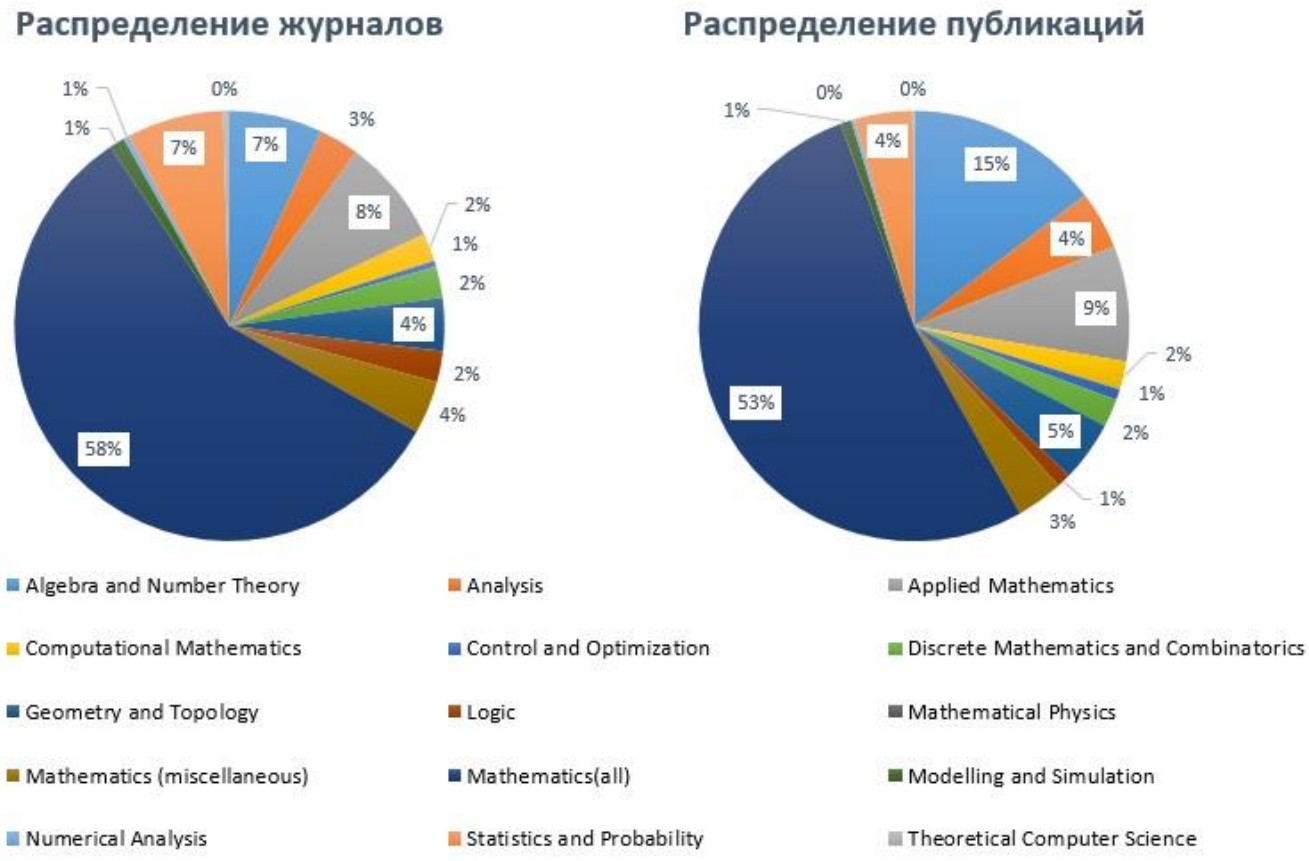


Рисунок 15 – Распределение журналов и публикаций по категориям области наук Mathematics

К категории Numerical Analysis относятся журналы International Journal Of Numerical Analysis и Numerical Analysis And Applications. Максимальное число цитирований – 111 – получила статья из первого журнала. В рейтинге SciVal для области наук Mathematics эта статья находится на 6752 месте. Следующие четыре статьи с общим числом цитирования от 86 до 62 находятся на 10 886, 10 898, 12 036 и 19 540 соответственно. Таким образом, только для того, чтобы в ядро автоматически попало четыре публикации категории Numerical Analysis, потребуется перебрать 19 540 публикаций.

К категории Theoretical Computer Science относятся «Journal of Experimental Algorithmics» и «Foundations and Trends in Theoretical Computer Science». Максимальное число цитирований из этих журналов составляет 774, что занимает 281 место в рейтинге SciVal для Mathematics. Наиболее цитируемые публикации занимают 2614, 5313, 6796, 8974, 10 216, 14 938 и 14 995 соответственно.

Остальные статьи в первые 20 000 результатов не попали так же, как и в случае с Numerical Analysis.

К Control and Optimization тоже относятся два журнала: «Optimization Letters» и «Springer Optimization and Its Applications». Наиболее цитируемые публикации здесь расположены на 6894, 10 443, 10 647, 11 703, 15 167, 15 690, 17 441, 18 525 соответственно.

Этот пример показывает, что автоматическое формирование ядер по категориям Numerical Analysis, Theoretical Computer Science и Control and Optimization является трудоемким процессом. Это связано с тем, что в Scopus Search API возможна выгрузка только по области наук. Для того чтобы получить категорию третьего уровня, необходимо дополнительно использовать Scopus Abstract Retrieval API, где для одного ключа допустима недельная выгрузка только 20 000 записей.

По данным SciVal (рисунок 16), категориям Control and Optimization и

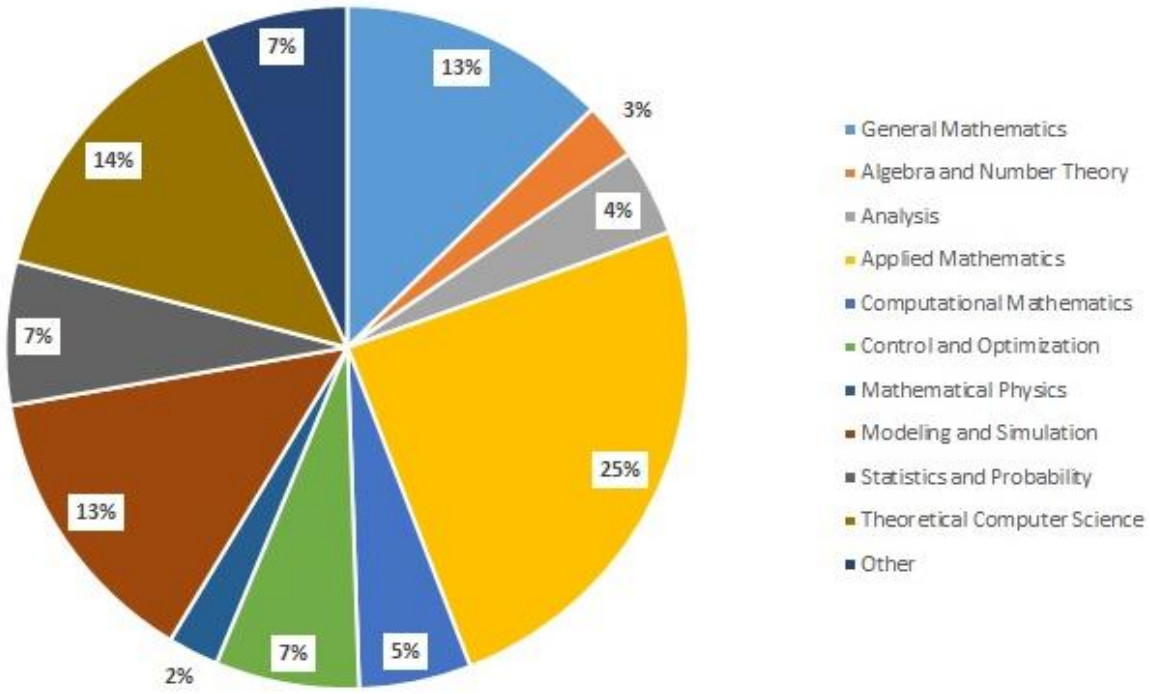


Рисунок 16 – Доли публикаций по категориям области Mathematics по данным SciVal

Theoretical Computer Science соответствует 6,8 % и 14,0 % от общемирового количества публикаций за 2009–2018 гг., что равносильно шестому и второму месту рейтинга по числу публикаций области Mathematics. Однако в силу мультидисциплинарности многих журналов и проведенным оценкам по монодисциплинарным журналам мы не можем однозначно утверждать, что эти доли действительно являются корректными.

На рисунке 17 приведены категории, которые чаще всего указаны в мультидисциплинарных журналах совместно с Theoretical Computer Science. Чаще всего это категории из области наук Computer Science. Таким образом, из-за таких журналов возможна потеря публикаций внутри не только области наук, но и целых научных направлений.

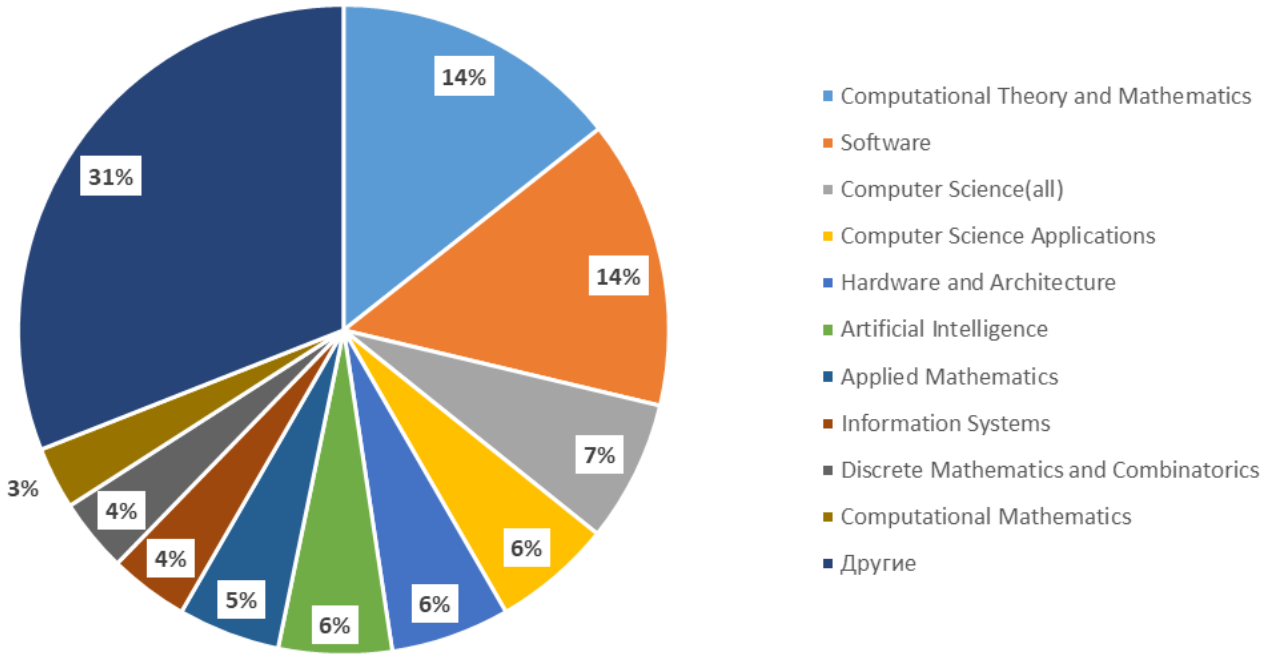


Рисунок 17 – Категории журналов, указываемые совместно с Theoretical Computer Science

Пример 2. Категория Medicine

Область наук Medicine представлена в классификации ASJC 49 различными категориями. Всего моножурналов в этой области 7301, при этом максимальное

число журналов относится к категории Medicine (all). Полностью отсутствуют журналы категорий Drug guides, Embryology, Reviews and References, Medical.

В основном в мультидисциплинарных журналах эти категории сочетаются с другими категориями областей Medicine(all), Health Professions, Pharmacology, Toxicology and Pharmaceutics и других.

Таким образом, собрать ядра по всем категориям области наук Medicine также не представляется возможным.

Пример 3. Категория Earth and Planetary Sciences

В области наук Earth and Planetary Sciences несмотря на то, что наибольшая доля журналов относится к категории Earth and Planetary Sciences, лидирующей категорией по числу публикаций является Atmospheric Science (рисунок 18).

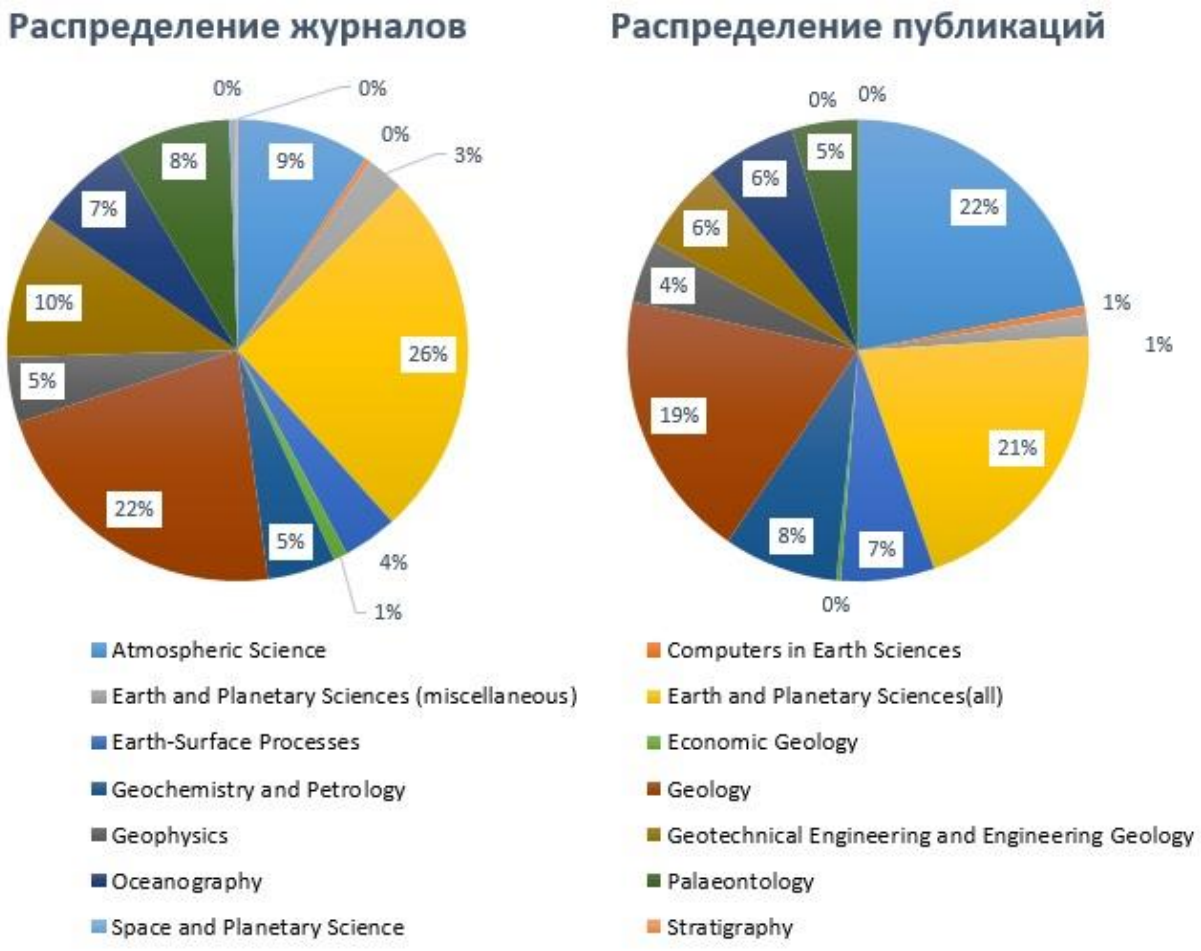


Рисунок 18 – Распределение журналов и публикаций по категориям области наук Earth and Planetary Sciences

Тем не менее, как и во многих областях наук, также встречаются и категории, публикации в которых отсутствуют полностью: Space and Planetary Science, Stratigraphy.

Таким образом, автоматическое формирование всех 333 ядер методом «Рейтинг цитирования» является невыполнимой задачей.

Рассмотрим возможность применения метода классификации на основе сжатия ко второму уровню классификации Scopus – областям наук – с автоматическим формированием ядер методом «Рейтинг цитирования».

Для каждой из 26 научных областей (кроме области Multidisciplinary) процесс выгрузки данных происходил в два этапа (приложение В):

1. При помощи Scopus Search API был произведен поиск публикаций, удовлетворяющий следующим условиям:
   - Период публикаций: 2009–2018 гг.;
   - Сортировка: по убыванию числа цитирований.
2. При помощи Scopus Abstract Retrieval API были выгружены аннотации публикаций и их категории.

Далее для каждой из 26 научных областей были автоматически сформированы ядра из 100 самых высокоцитируемых публикаций, у которых все категории принадлежали области наук, для которой создавалось ядро.

Тестовые файлы общим количеством 1040 (по 40 для каждой категории) были отобраны произвольным образом.

Результаты классификации показали, что 57 % тестовых файлов было определено ошибочно (таблица 14). Возможно, это связано с тем, что в отличие от узких категорий в научных областях встречается более разнообразная терминология, что затрудняет применение к ним метода. Более того, в некоторых областях наук в 100 самых высокоцитируемых публикаций попало большое количество публикаций из категории (all). Например, в области Arts and Humanities (все 100 публикаций в ядре), Chemical Engineering (39 из 56), Chemistry (96 из 97), Dentistry (55 из 98), Mathematics (48 из 56) и т.д.

Таблица 14 – Результаты классификации по областям наук (* – только по данным публикаций с одной категорией)

Область	**Количество ошибок**	**Общее количество категорий**	**Количество категорий в ядре***	**Количество публикаций с одной категорией в ядре**
Agricultural and Biological Sciences	31	12	7	77
Arts and Humanities	12	14	1	100
Biochemistry, Genetics and Molecular Biology	39	16	5	81
Business, Management and Accounting	24	11	3	25
Chemical Engineering	35	9	2	56
Chemistry	33	8	2	97
Computer Science	20	13	7	20
Decision Sciences	17	5	3	64
Dentistry	13	7	4	98
Earth and Planetary Sciences	20	14	7	76
Economics, Econometrics and Finance	30	4	3	86
Energy	22	6	2	90
Engineering	34	17	4	77
Environmental Science	17	13	7	27
Health Professions	16	17	6	100
Immunology and Microbiology	19	7	6	45
Materials Science	19	9	3	82
Mathematics	12	15	8	65
Medicine	37	49	9	66
Neuroscience	30	10	4	68
Nursing	15	24	12	93
Pharmacology, Toxicology and Pharmaceutics	32	6	4	45
Physics and Astronomy	19	11	5	91
Psychology	30	8	7	94

Продолжение таблицы 14

Область	**Количество ошибок**	**Общее количество категорий**	**Количество категорий в ядре***	**Количество публикаций с одной категорией в ядре**
Social Sciences	13	23	8	89
Veterinary	5	5	4	100
Всего	594			
Доля от общего количества	57 %			

Автоматическое формирование ядер по области наук с условием, что в него будет входить определенная доля статей по каждой из категорий, приписанной к этой области, является невозможным из-за оценки представленности категорий в области наук.

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

1. ограничения 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.

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.

3.4. Классификация публикаций из журнала «Геология и геофизика»

Журнал «Геология и геофизика» издается Сибирским отделением РАН с 1960 г. Его англоязычная версия индексируется в ББД Scopus. Среди тематических рубрик русскоязычной версии этого журнала на сайте издательства [127] обозначены:

- палеонтология и региональная геология;
- минералогия и петрология;
- проблемы геотектоники и геоморфологии полезных ископаемых

и другие.

Однако в ББД Scopus у журнала указаны только две общие тематики: «Геология» и «Геофизика». В связи с тем, что классификация в Scopus происходит только на журнальном уровне, у всех публикаций этого журнала также проставлены обе эти тематики.

Такая классификация может как затруднять поиск статей из этого журнала, так и понижать рейтинг этого журнала. Например, если ученый выполнит запрос в Scopus по публикациям области «Геохимия и петрология», то ему не попадутся статьи из этого журнала. И наоборот, категории «Геология» и «Геофизика» будут содержать лишние публикации.

Для классификации были отобраны следующие категории области «Earth and Planetary Sciences»:

- Atmospheric Science
- Computers in Earth Sciences
- Earth-Surface Processes
- Economic Geology
- Geochemistry and Petrology
- Geology
- Geophysics
- Geotechnical Engineering and Engineering Geology

- Oceanography
- Paleontology
- Space and Planetary Science
- Stratigraphy

В ядра вошли 100 самых высокоцитируемых публикаций каждой категории только с одной указанной категорией. В связи с отсутствием в категории Stratigraphy необходимого количества публикаций только с одной категорией, в ядро этой категории вошли 100 самых высокоцитируемых публикаций, у которых в названии и аннотации встречается термин «Stratigraphy».

Метод классификации, основанный на сжатии данных, был применен к 651 аннотации статей, опубликованных в журнале в 2014–2019 гг.

На рисунке 19 представлены самые многочисленные по числу публикаций категории. Лидирующей категорией является категория Geology, однако категория Geophysics, указанная также в ББД Scopus, занимает только четвертое место.

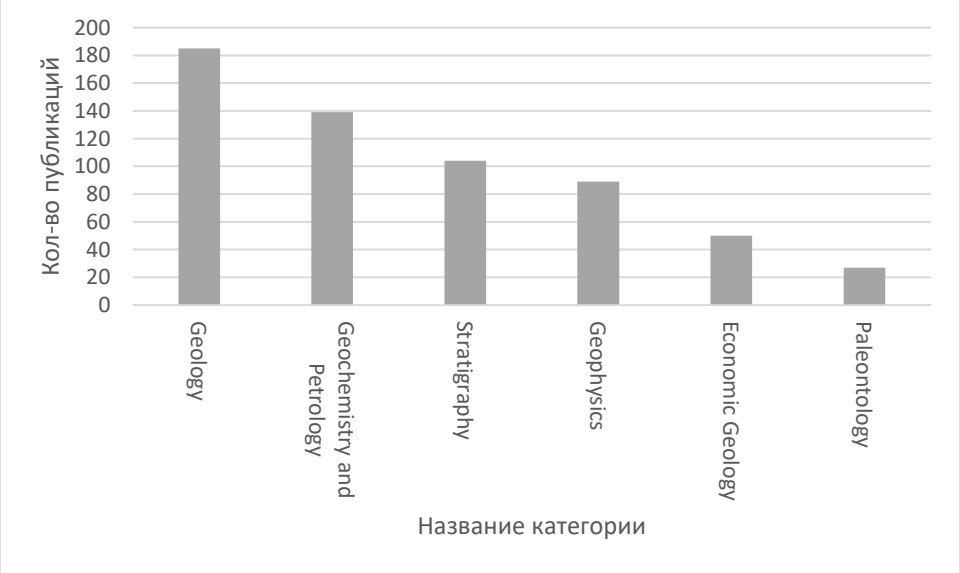


Рисунок 19 – Результаты классификации аннотаций публикаций журнала «Геология и геофизика»

Для проверки качества классификации была проведена экспертная оценка. Для этого случайным образом были отобраны 250 публикаций (по 50 на каждую

категорию, присутствующую на рисунке, Stratigraphy и Paleontology совмещены) и отправлены ведущим специалистам в этих областях. Данные были представлены в виде таблицы со столбцами:

- eid
- Авторы
- Аннотация
- Ссылка на полный текст (если есть)
- Категория, определенная методом, основанном на сжатии

Эксперт проставлял отметки, верно ли определена категория, и если нет, то к какой категории, по его мнению, должна быть отнесена классифицируемая публикация. Результаты экспертной оценки были сгруппированы следующим образом. Такие комментарии эксперта, как: «Метод классификации, основанный на сжатии данных, верно определяет категорию» и «Определенная категория может быть указана как второстепенная» были отнесены к случаям, что метод верно определил категорию. Отрицательные результаты определения были разбиты на три группы:

- Другая категория Scopus
- Другая категория не из Scopus
- Экспертом не указана верная категория

Результаты экспертной оценки представлены на рисунке 20.

Больше всего метод на основе сжатия ошибся при отнесении публикаций к области Stratigraphy/Paleontology. Возможно, это связано с изначальным выбором ядра этой категории.

Таким образом, экспертная оценка и метод на основе сжатия показали, что в журнале «Геология и геофизика» действительно встречаются публикации не только из категорий Geology и Geophysics, указанных в ББД Scopus. Отметим, что в рейтинге Scimago Journal & Country Rank журнал «Геология и геофизика» также присутствует только в этих двух направлениях. SJR журнала равен 0.877. Если бы «Геология и геофизика» вошел в этих рейтингах дополнительно в категории

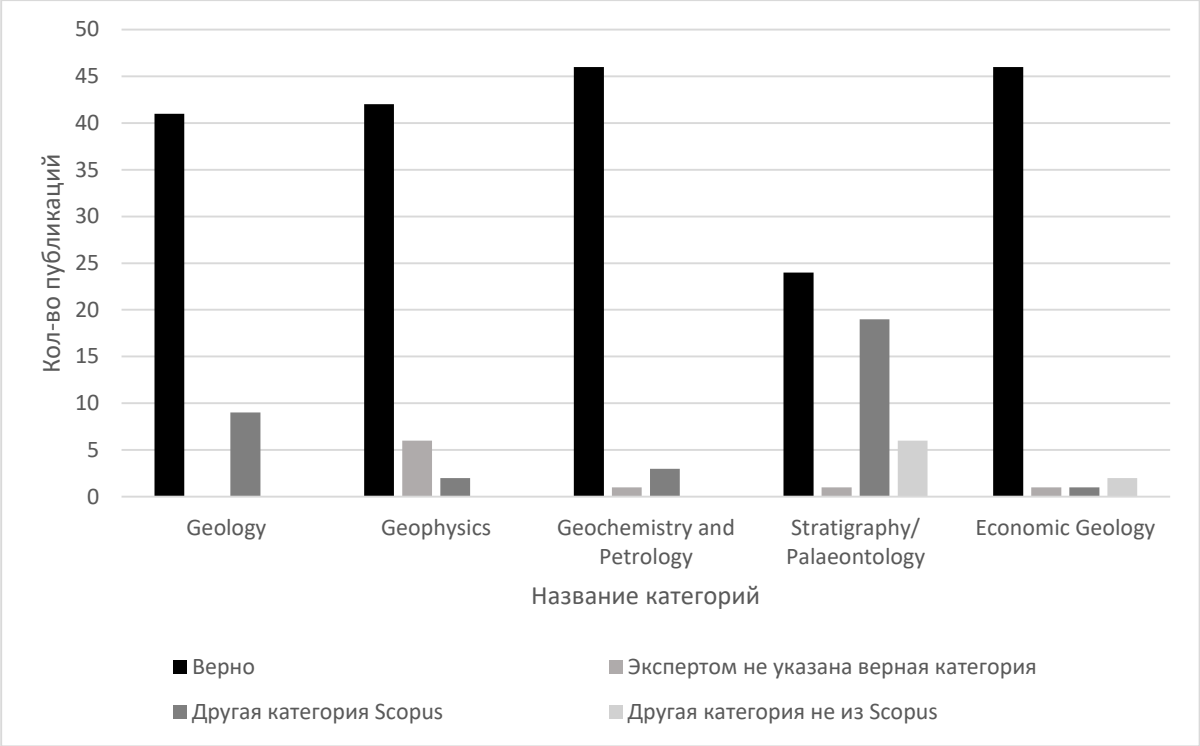


Рисунок 20 – Результаты экспертной оценки работы метода, основанного на сжатии данных

Stratigraphy и Geochemistry and Petrology, то он бы мог попасть на 10 место в Q1 и 39 место в Q2 в этих категориях соответственно.

3.5. Выводы к Главе 3

В Главе 3 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 с научной электронной библиотеки «КиберЛенинка»);

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

Результаты классификации полных англоязычных текстов 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 классификация 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 %).

В результате классификации русскоязычных полных текстов 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.

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 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:

- ограничений на выгрузку данных, установленного в Scopus, и отсутствием названий категорий в Scopus Search API.

- отсутствия в некоторых категориях журналов и, соответственно, публикаций, у которых указана единственная категория.

Показано, что на результаты классификации, проводимой с использованием метода, основанного на сжатии, влияет также изначальное количество ядер и их тематическая близость.

Также рассматривается применение метода классификации, основанного на сжатии данных, к классификации публикаций из журнала «Геология и геофизика». В качестве ядер используются все категории 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, т. е. в направления, указанные на сайте издательства, то он бы мог попасть на 16 место в Q2 и 59 место в Q2 соответственно.

ГЛАВА 4. СРАВНЕНИЕ МЕТОДА КЛАССИФИКАЦИИ, ОСНОВАННОГО НА СЖАТИИ ДАННЫХ, С ДРУГИМИ МЕТОДАМИ

Одним из важных этапов проверки эффективности алгоритма классификации является его сравнение с другими методами. Следующие методы были реализованы при помощи библиотеки scikit-learn для языка программирования Python [128]:

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

Для этих методов при помощи функции TfidfVectorizer были построены частотные вектора встречаемости слов на основе схемы tf*idf.

Отметим, что в исследовании не проводились подробные эксперименты с целью подбора наилучших параметров для методов из библиотеки scikit-learn, в ом числе и по оптимальному размеру обучающей выборки.

Эти методы и метод классификации, основанный на сжатии данных, (DCM – Data Compression Method) были применены к двум типам данных:

- Аннотации публикаций из ББД Scopus (Обучающая выборка – ядра, подобранные методом «Рейтинг цитирования» из п. 3.3.1; тесты – п.3.3.1);
- Полные тексты с веб-сайта arXiv.org (Обучающая выборка – 100 научных текстов с единственной категорией (табл. 15); тесты – 20 тестовых файлов, не входящих в обучающие выборки также с единственной категорией).

Таблица 15 – Категории arXiv.org, используемые при сравнении методов классификации

Сокращенное название категории	**Полное название**	**Область**
astro-ph.EP	Earth and Planetary Astrophysics	Physics
astro-ph.SR	Solar and Stellar Astrophysics	Physics
cond-mat.quant-gas	Quantum Gases	Physics
cond-mat.stat-mech	Statistical Mechanics	Physics
cs.AI	Artificial Intelligence	Computer Science
cs.DC	Distributed, Parallel, and Cluster Computing	Computer Science
cs.IT	Information Theory	Mathematics, Computer Science
econ.EM	Econometrics	Economics
hep-ex	High Energy Physics - Experiment	Physics
hep-th	High Energy Physics - Theory	Physics
math.AG	Algebraic Geometry	Mathematics
math.AP	Analysis of PDEs	Mathematics
math.CO	Combinatorics	Mathematics
math.DG	Differential Geometry	Mathematics
math-ph	Mathematical Physics	Physics, Mathematics
nucl-ex	Nuclear Experiment	Physics
nucl-th	Nuclear Theory	Physics
physics.optics	Optics	Physics
q-bio.BM	Biomolecules	Quantitative Biology
quant-ph	Quantum Physics	Physics

На рисунке 21 приведено сравнение точности классификации различными методами в зависимости от вида источника данных.

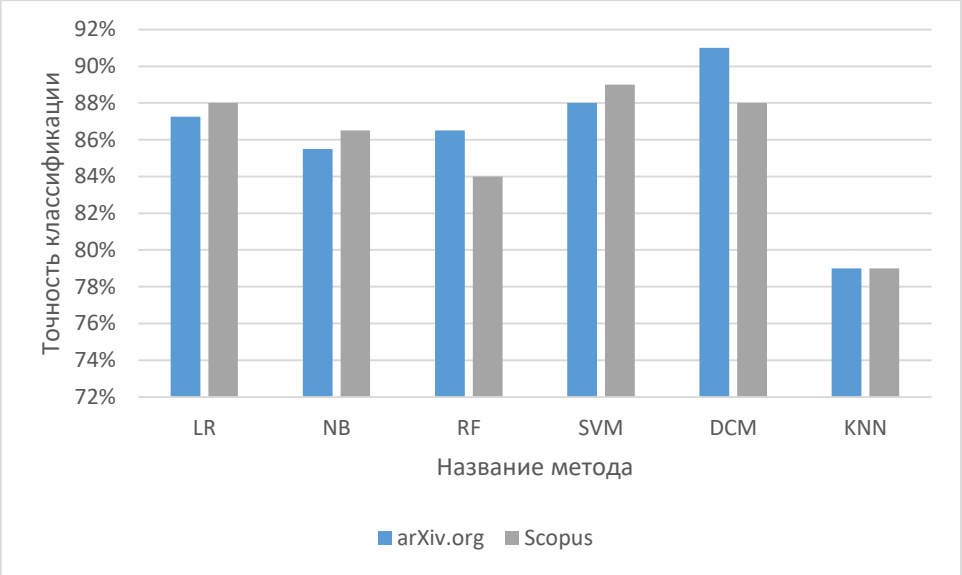


Рисунок 21 – Сравнение точности классификации различными методами в зависимости от источника данных

Результаты показывают, что на аннотациях и полнотекстовых документах все методы, кроме KNN, показывают точность от 84 %. На полнотекстовых документах наилучшую точность показал DCM (91 %), следующий за ним – SVM (88 %). На аннотациях – SVM (89 %), DCM (88%) и LR (88 %).

В таблице 16 приведены основные метрики оценки качества классификации для полных текстов и аннотаций:

- Accuracy
- Precision
- Recall

Таблица 16 – Сравнение основных характеристик классификации в разных методах для полных текстов и аннотаций

	accuracy		precision		recall	
Метод	**Полные тексты**	**Аннотации**	**Полные тексты**	**Аннотации**	**Полные тексты**	**Аннотации**
LR	87%	88%	87%	88%	88%	88%

Продолжение таблицы 16

Метод	Полные тексты	Аннотации	Полные тексты	Аннотации	Полные тексты	Аннотации
NB	86%	87%	85%	87%	87%	87%
RF	87%	84%	86%	84%	87%	85%
SVM	88%	**89%**	88%	**89%**	89%	**89%**
DCM	**91%**	88%	**91%**	87%	**92%**	88%
KNN	79%	79%	79%	79%	81%	80%

Рассмотрим подробнее каждый из полученных результатов.

4.1. Результаты классификации полных текстов

В этом случае ошибки были разделены на следующие типы:

- I тип. Ложный выбор категории внутри области наук. К этому типу отнесены те ошибочно определенные тестовые файлы, у которых определилась другая категория из научной области;
- II тип. Ложный выбор области наук.

Распределение ошибок, полученных при классификации текстов с arXiv.org разными методами, по типам приведено на рисунке 22.

Наименьшее число ошибок второго типа получается при классификации методом на основе сжатия данных. Таких ошибок всего лишь одна: в тестовом файле вместо категории «q-bio.BM» определилась «astro-ph.EP», но эта ошибка встречается только в этом методе.

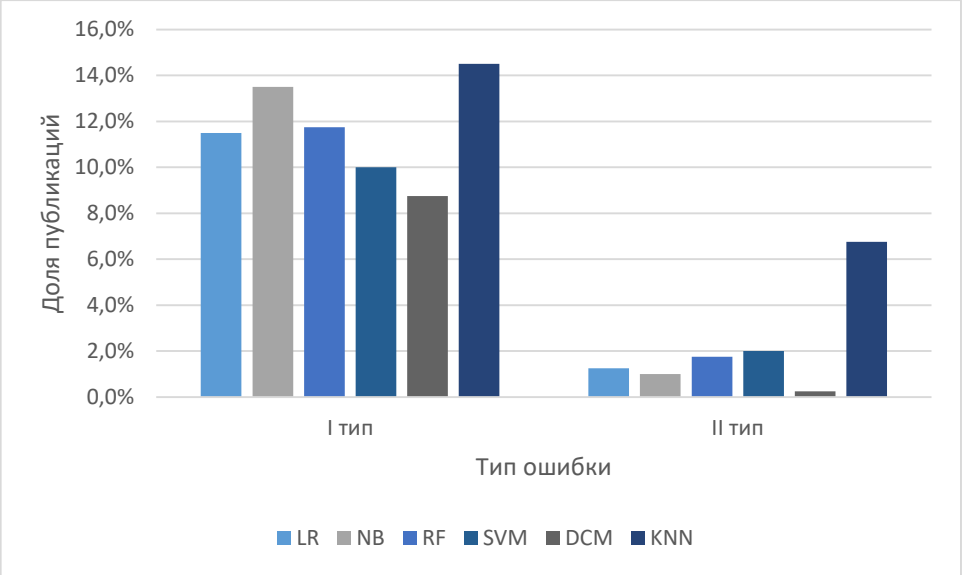


Рисунок 22 – Сравнение типов ошибок, получаемых при классификации полных текстов различными методами

В шести случаях все методы определили одну и ту же неверную категорию. Общее количество случаев, когда ошибается только один метод, приведено в таблице 17.

Таблица 17 – Количество случаев, когда ошибся только один метод

Метод	**Количество случаев, когда ошибся только один метод**	**Всего ошибок**	**Доля от общего числа ошибок этого метода**
DCM	4	36	11%
LR	2	51	4%
NB	2	58	3%
RF	9	54	17%
SVM	0	48	0%
KNN	30	85	35%

Большая доля ошибок классификации, полученных только методом DCM, при низкой общей доле ошибок обуславливается, вероятно, тем, что только этот метод, в отличие от остальных, работает на другом представлении данных.

4.2. Результаты классификации аннотаций публикаций

Разобьем ошибки определения категории различными методами на следующие типы:

- I тип. Ложный выбор категории внутри области наук;
- II тип. Ложный выбор области наук внутри научного направления;
- III тип. Ложный выбор научного направления.

Сравнение типов ошибок по различным методам приведено на рисунке 23. DCM и NB в основном неверно определяют категории научных публикаций. В остальных же методах лидирующее число ошибок приходится на II тип.

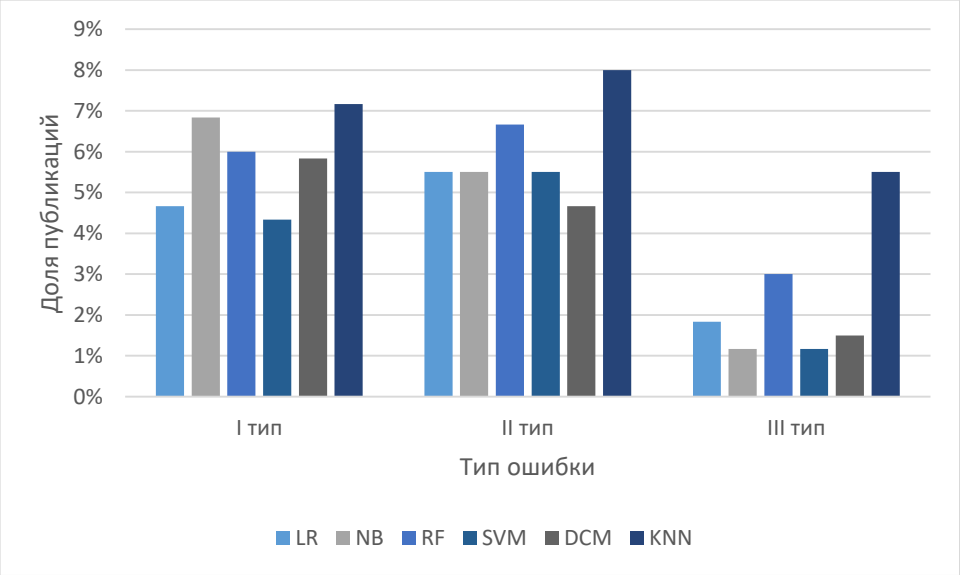


Рисунок 23 – Сравнение типов ошибок, получаемых при классификации аннотаций различными методами

У 12 тестов из 600 все методы определили одну и ту же отличную от исходной категорию. На рисунке 24 приведен пример аннотации одного из таких тестовых файлов. В качестве категории в Scopus у этой публикации была указана «Literature». Но все методы отнесли ее к категории «History». Подробное изучение

аннотации показало, что терминологически эта аннотация больше близка к категории «History».

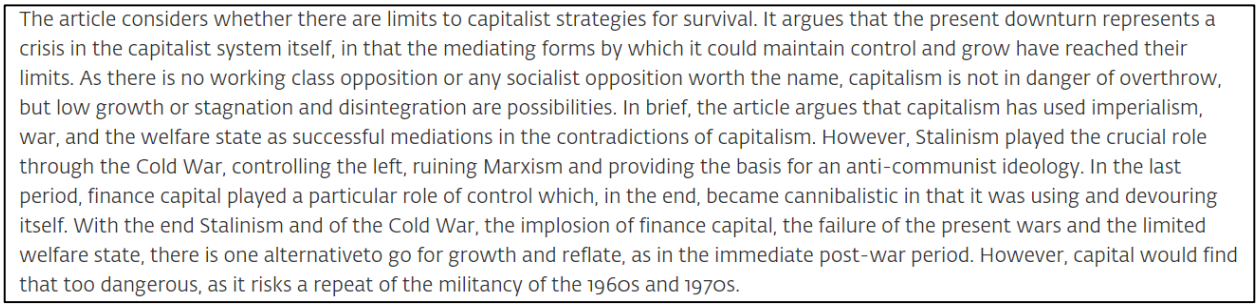

The article considers whether there are limits to capitalist strategies for survival. It argues that the present downturn represents a crisis in the capitalist system itself, in that the mediating forms by which it could maintain control and grow have reached their limits. As there is no working class opposition or any socialist opposition worth the name, capitalism is not in danger of overthrow, but low growth or stagnation and disintegration are possibilities. In brief, the article argues that capitalism has used imperialism, war, and the welfare state as successful mediations in the contradictions of capitalism. However, Stalinism played the crucial role through the Cold War, controlling the left, ruining Marxism and providing the basis for an anti-communist ideology. In the last period, finance capital played a particular role of control which, in the end, became cannibalistic in that it was using and devouring itself. With the end Stalinism and of the Cold War, the implosion of finance capital, the failure of the present wars and the limited welfare state, there is one alternativeto go for growth and reflate, as in the immediate post-war period. However, capital would find that too dangerous, as it risks a repeat of the militancy of the 1960s and 1970s.

Рисунок 24. Пример публикации категории Literature, у которой все методы определили категорию History

На рисунке 25 приведен пример публикации категории Library and Information Sciences, у которой все методы определили категорию Marketing. Терминологически она также является близкой к категории Marketing.

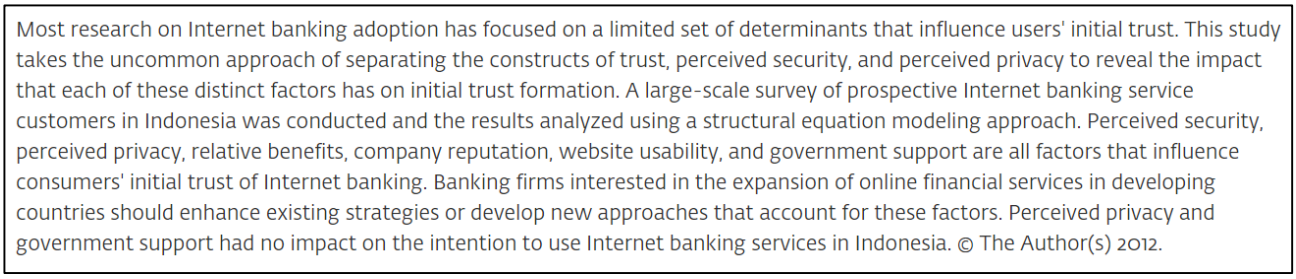

Most research on Internet banking adoption has focused on a limited set of determinants that influence users' initial trust. This study takes the uncommon approach of separating the constructs of trust, perceived security, and perceived privacy to reveal the impact that each of these distinct factors has on initial trust formation. A large-scale survey of prospective Internet banking service customers in Indonesia was conducted and the results analyzed using a structural equation modeling approach. Perceived security, perceived privacy, relative benefits, company reputation, website usability, and government support are all factors that influence consumers' initial trust of Internet banking. Banking firms interested in the expansion of online financial services in developing countries should enhance existing strategies or develop new approaches that account for these factors. Perceived privacy and government support had no impact on the intention to use Internet banking services in Indonesia. © The Author(s) 2012.

Рисунок 25 – Пример публикации категории Library and Information Sciences, у которой все методы определили категорию Marketing

В таблице 18 показано, как часто ошибался только один из представленных методов.

Таблица 18 – Количество случаев, когда ошибся только один метод

Метод	**Количество случаев, когда ошибся только один метод**	**Всего ошибок**	**Доля от общего числа ошибок этого метода**
DCM	17	72	24 %

Продолжение таблицы 18

Метод	Количество случаев, когда ошибся только один метод	Всего ошибок	Доля от общего числа ошибок этого метода
LR	2	72	3 %
NB	3	81	4 %
RF	29	94	31 %
SVM	1	66	2 %
KNN	37	124	30 %

Таким образом, 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 %.

4.3. Выводы к Главе 4

В Главе 4 рассматривается сравнение метода классификации, основанного на сжатии данных, с традиционными методами классификации:

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

Сравнение происходит на полных текстах и аннотациях публикаций. Результаты исследования показали, что 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 классификации научных текстов, основанный на сжатии данных, показывает высокую эффективность, сравнимую с традиционными методами классификации, при этом является более точным при классификации полнотекстовых документов.

ЗАКЛЮЧЕНИЕ

В рамках диссертационного исследования был разработан метод автоматической классификации научных текстов, показывающий высокую эффективность на различных типах данных. Предложенный метод основан на применении алгоритмов сжатия данных для сравнительного анализа «близости» текстов.

В результате выполнения работы были получены следующие 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 точность «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 из 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, возможно, могли бы значительно улучшить качество классификаций.

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

СПИСОК СОКРАЩЕНИЙ

DCM Data Compression Method

KNN k-nearest Neighbors

LR Logistic regression

NB Naive Bayes classifier

RF Random forest

SVM Support vector machine

WoS Web of Science

ББД Библиографическая база данных

Текст Текстовый документ

СПИСОК ИЛЛЮСТРАЦИЙ

СПИСОК ТАБЛИЦ

СПИСОК ЛИТЕРАТУРЫ

1. Ryabko B. Y. Information-Theoretic method for classification of texts / B. Y. Ryabko, A. E. Gus'kov, I. V. Selivanova // Problems of Information Transmission – 2017. – V. 53, № 3. – P. 294–304.

2. Selivanova I. V. Classification by compression: Application of information-theory methods for the identification of themes of scientific texts / I. V. Selivanova, B. Y. Ryabko, A. E. Guskov // Automatic Documentation and Mathematical Linguistics – 2017. – V. 51, № 3. – P. 120–126.

3. Selivanova I. V. Classification of Scientific Texts Based on the Compression of Annotations to Publications / I. V. Selivanova, D. V. Kosyakov, A. E. Guskov // Automatic Documentation and Mathematical Linguistics – 2019. – V. 53, № 6. – P. 329–342.

4. 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.

5. Ryabko B. Using data-compressors for statistical analysis of problems on homogeneity testing and classification / B. Ryabko, A. Guskov, I. Selivanova // IEEE International Symposium on Information Theory – Proceedings – 2017. – P. 121–125.

6. Sebastiani F. Machine learning in automated text categorization / F. Sebastiani // ACM Computing Surveys (CSUR) – 2002. – V. 34, № 1. – P. 1–47.

7. Yu B. An evaluation of text classification methods for literary study / B. Yu // Literary and Linguistic Computing – 2008. – V. 23, № 3 – P. 327–343.

8. Barakhnin, V. B. Computer Classification of Russian Poetic Texts by Genres and Styles / V. B. Barakhnin, O. Yu. Kozhemyakina, I. S. Pastushkov, E. V. Rychkova // Vestnik NSU. Series: Linguistics and Intercultural Communication – 2017. – V. 15, № 3. – P.13–23.

9. Can E. F. Automatic Categorization of Ottoman Literary Texts by Poet and Time Period / E. F. Can, F. Can, P. Duygulu, M. Kalpakli // Computer and Information Sciences. – 2011. – P. 51–57.

10. Батура Т. В. Формальные методы определения авторства текстов / Т. В. Батура // Вестник НГУ. Серия: Информационные технологии – 2012. – Т. 10, № 4 – С. 81–94.

11. Singh S. A Comparative Analysis of Text Classification Algorithms for Ambiguity Detection in Requirement Engineering Document Using WEKA / S. Singh, L. P. Saikia // Lecture Notes in Networks and Systems: ICT Analysis and Applications. – 2019. – P. 345–354.

12. Oliveira E. Automatic classification of journalistic documents on the Internet / E. Oliveira, D. B. Filho // Transinformacao – 2017. – V. 29, № 3 – P. 245–255.

13. Liu W. Index-based online text classification for SMS spam filtering / W. Liu, T. Wang // Journal of Computers – 2010. – V. 5, № 6. – P.844–851.

14. Kiritchenko S. Sentiment analysis of short informal texts / S. Kiritchenko, X. Zhu, S. M. Mohammad // Journal of Artificial Intelligence Research – 2014. – V. 50. – P. 723–762.

15. Bahgat E. M. Efficient email classification approach based on semantic methods / E. M. Bahgat, S. Rady, W. Gad, I. F. Moawad // Ain Shams Engineering Journal – 2018. – V. 9, № 4. – P.3259–3269.

16. Shi T. Research on the Application of E-Mail Classification Based on Support Vector Machine / T. Shi // Frontiers in Computer Education, – 2012. – P. 987–994.

17. Islam M. R. An innovative analyser for multi-classifier e-mail classification based on grey list analysis / M. R. Islam, W. Zhou, M. Guo, Y. Xiang // Journal of Network and Computer Applications – 2009. – V. 32, № 2. – P. 357–366.

18. Hasan M. EMOTEX: Detecting Emotions in Twitter Messages / M. Hasan, E. Rundensteiner, E. Agu // SocialCom Conference – 2014. – P. 27–31.

19. Rubtsova Y. V. Research and Development of Domain Independent Sentiment Classifier / Y. V. Rubtsova // SPIIRAS Proceedings – 2014. – V. 5, № 36. – P. 59.

20. Shtovba S. Detection of social network toxic comments with usage of syntactic dependencies in the sentences / S. Shtovba, O. Shtovba, M. Petrychko // CEUR Workshop Proceedings – 2019. – V. 2353. – P. 313–323.

21. Ma J. The classification of rumour standpoints in online social network based on combinatorial classifiers / J. Ma, Y. Luo // Journal of Information Science – 2020. – V. 46, № 2. – P. 191–204.

22. Santos C. N. Dos Deep convolutional neural networks for sentiment analysis of short texts / C. N. Dos Santos, M. Gatti // COLING 2014 – 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers – 2014. – P. 69–78.

23. Sriram B. Short text classification in twitter to improve information filtering / B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, M. Demirbas // SIGIR 2010 Proceedings – 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – 2010. – P. 841–842.

24. Zantout R. A universal method for author identification using statistical properties of text / R. Zantout, Z. Osman, L. Hamandi // ACM International Conference Proceeding Series – 2018.

25. Tang X. Authorship attribution of the golden lotus based on text classification methods / X. Tang, S. Liang, Z. Liu // ACM International Conference Proceeding Series – 2019. – T. Part F1481 – P. 69–72.

26. Stańczyk U. Recognition of author gender for literary texts / U. Stańczyk // Advances in Intelligent and Soft Computing – 2011. – V. 103 – P. 229–238.

27. Walkowiak T. Feature Extraction in Subject Classification of Text Documents in Polish / T. Walkowiak, S. Datko, H. Maciejewski // ICAISC 2018: Artificial Intelligence and Soft Computing – 2018. – P. 445–452.

28. Miao Y. Document clustering using character N-grams: A comparative evaluation with term-based and word-based clustering / Y. Miao, V. Kešelj, E. Milios // International Conference on Information and Knowledge Management, Proceedings – 2005. – P. 357–358.

29. Волкова Л. Об ассоциативных бинарных мерах близости документов: классификация и приложение к кластеризации / Л. Волкова, Ю. Строганов // Новые Информационные Технологии В Автоматизированных Системах – 2014. – Т. 17. – С. 421–432.

30. Baghel R. A Frequent Concepts Based Document Clustering Algorithm / R. Baghel, D. R. Dhir // International Journal of Computer Applications – 2010. – V. 4, № 5. – P. 6–12.

31. Beil F. Frequent term-based text clustering / F. Beil, M. Ester, X. Xu // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – 2002. – P. 436–442.

32. Deng Z.H. A comparative study on feature weight in text categorization / Z. H. Deng, S.W. Tang, D. Q. Yang, M. Zhang, L. Y. Li, K. Q. Xie // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) – 2004. – V. 3007. – P. 588–597.

33. Lunh H. P. The Automatic Creation of Literature Abstracts / H. P. Lunh // IBM Journal of Research Development – 1958. – V. 2, № 2. – P.159–165.

34. Riloff E. Little words can make a big difference for text classification / E. Riloff // SIGIR Forum (ACM Special Interest Group on Information Retrieval) – 1995. – P. 130–136.

35. Lebanon G. Metric learning for text documents / G. Lebanon // IEEE Transactions on Pattern Analysis and Machine Intelligence – 2006. – V. 28, № 4. – P. 497–508.

36. Hu L. Y. The distance function effect on k-nearest neighbor classification for medical datasets / L. Y. Hu, M.W. Huang, S. W. Ke, Tsai C.F. // SpringerPlus – 2016. – V. 5, № 1. – P. 1–9.

37. Zhang S. A novel text classification based on Mahalanobis distance / S. Zhang, X. Pan // ICCRD2011 – 2011 3rd International Conference on Computer Research and Development – 2011. – V. 3. – P. 156–158.

38. Dhar A. Classification of text documents through distance measurement: An experiment with multi-domain Bangla text documents / A. Dhar, N. Dash, K. Roy // 3rd

International Conference on Advances in Computing, Communication & Automation (ICACCA) (Fall), Dehradun, 2017, – 2017. – P. 1–6.

39. Walkowiak T. Distance Metrics in Open-Set Classification of Text Documents by Local Outlier Factor and Doc2Vec / T. Walkowiak, S. Datko, H. Maciejewski // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) – 2019. – P. 102–109.

40. Zu G. Automatic text classification of English newswire articles based on statistical classification techniques / G. Zu, W. Ohyama, T. Wakabayashi, F. Kimura // Electrical Engineering in Japan (English translation of Denki Gakkai Ronbunshi) – 2005. – V. 152, № 1. – P. 50–60.

41. Chen Z. The Lao Text Classification Method Based on KNN / Chen Z., Zhou L.J., Li X. Da, Zhang J. N., Huo W. J. // Procedia Computer Science – 2020. – T. 166. – C. 523–528.

42. Метод ближайших соседей [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/index.php?title=Метод_ближайшего_соседа (дата обращения: 08.05.2020).

43. Wang X. A fuzzy KNN algorithm based on weighted chi-square distance / Wang X., Yao P. // ACM International Conference Proceeding Series – 2018. – P. 1–6.

44. Wang C.-Y. A K-Nearest Neighbor Algorithm based on cluster in text classification, 2010. – P. 225–228.

45. Zhang X. A k-nearest neighbor text classification algorithm based on fuzzy integral / X. Zhang, B. Li., X. Sun // Proceedings - 2010 6th International Conference on Natural Computation, ICNC 2010 – 2010. – V. 5. – P. 2228–2231.

46. Yang Y. A Scalability Analysis of Classifiers in Text Categorization / Y. Yang, J. Zhang, B. Kisiel // SIGIR Forum (ACM Special Interest Group on Information Retrieval) – 2003. – № SPEC. ISS. – P. 96–103.

47. Tan S. Neighbor-weighted K-nearest neighbor for unbalanced text corpus / S. Tan // Expert Systems with Applications – 2005. – V. 28, № 4 – P. 667–671.

48. Keller J. M. A Fuzzy K-Nearest Neighbor Algorithm / J. M. Keller, M. R. Gray // IEEE Transactions on Systems, Man and Cybernetics – 1985. – V. SMC-15, № 4 – P. 580–585.

49. Denœux T. A k-nearest neighbor classification rule based on Dempster-Shafer theory / T. Denœux // Classic Works of the Dempster-Shafer Theory of Belief Functions – 2008. – P. 737–760.

50. Garg A. Understanding probabilistic classifiers / A. Garg, D. Roth // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) – 2001. – V. 2167. – P. 179–191.

51. Jiang L. Deep feature weighting for naive Bayes and its application to text classification / L. Jiang, C. Li, S. Wang, L. Zhang // Engineering Applications of Artificial Intelligence – 2016. – V. 52. – P. 26–39.

52. Howedi F. Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data / F. Howedi, M. Mohd // Computer Engineering and Intelligent Systems – 2014. – V. 5, № 4. – P. 48–56.

53. Myaeng S. H. Some effective techniques for naive bayes text classification / S. H. Myaeng, K.S. Han, H.C. Rim // IEEE Transactions on Knowledge and Data Engineering – 2006. – V. 18, № 11. – P. 1457–1466.

54. Zhang W. An improvement to naive bayes for text classification / W. Zhang, F. Gao // Procedia Engineering – 2011. – V. 15. – P. 2160–2164.

55. Wang B. A novel text classification algorithm based on Naïve Bayes and KL-divergence / B. Wang, S. Zhang // Parallel and Distributed Computing, Applications and Technologies, PDCAT Proceedings – 2005. – V. 2005. – P. 913–915.

56. Wang S. Adapting naive Bayes tree for text classification / S. Wang, L. Jiang, C. Li // Knowledge and Information Systems – 2015. – V. 44, № 1. – P. 77–89.

57. Xu S. Bayesian Multinomial Naïve Bayes Classifier to Text Classification / S. Xu // Lecture Notes in Electrical Engineering – 2017. – P. 347–352.

58. Jiang L. Structure extended multinomial naive Bayes / L. Jiang, S. Wang, C. Li, L. Zhang // Information Sciences – 2016. – V. 329. – P. 346–356.

59. Tan Z. Research on the Text Emotion of Multinomial Naïve Bayes Integration Algorithm / Z. Tan, Y. Zhang, C. Zhang, R. Huang, P. Lei, X. Duan // Proceedings of 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference, IMCEC 2019 – 2019. – № Imcec – P. 107–111.

60. Narayanan V. Fast and accurate sentiment classification using an enhanced Naive Bayes model / V. Narayanan, I. Arora, A. Bhatia // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) – 2013. – V. 8206 LNCS – P. 194–201.

61. Bi Z. Gaussian Naive Bayesian Data Classification Model Based on Clustering Algorithm / Z. Bi, Y. Han, C. Huang, M. Wang. – 2019. – V. 168 – P. 396–400.

62. Singh G. Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification / Singh G., Kumar B., Gaur L., Tyagi A. // 2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019 – 2019. – P. 593–596.

63. Xu S. Bayesian Naïve Bayes classifiers to text classification / S. Xu // Journal of Information Science – 2018. – V. 44, № 1. – P. 48–59.

64. Метод опорных векторов (SVM) [Электронный ресурс]. URL: https://neerc.ifmo.ru/wiki/index.php?title=Метод_опорных_векторов_(SVM) (дата обращения: 21.05.2020).

65. Wang Z. Q. An optimal SVM-based text classification algorithm / Z. Q. Wang, X. Sun, D. X. Zhang, X. Li // Proceedings of the 2006 International Conference on Machine Learning and Cybernetics – 2006. – V. 2006. – P. 1378–1381.

66. Ji L. A SVM-based text classification system for knowledge organization method of crop cultivation/ L. Ji, X. Cheng, L. Kang, D. Li, D. Li, K. Wang, Y. Chen // International Conference on Computer and Computing Technologies in Agriculture – 2012. – P. 318–324.

67. Чистяков С. П. Случайные леса: обзор / С.П. Чистяков // Труды Карельского научного центра РАН – 2013. – № 1. – С. 117–136.

68. Xu B. An improved random forest classifier for text categorization / B. Xu, X. Guo, Y. Ye, J. Cheng // Journal of Computers (Finland) – 2012. – V. 7, № 12 – P. 2913–2920.

69. Islam M. Z. A semantics aware random forest for text classification / M. Z. Islam, J. Liu, J. Li, L. Liu, W. Kang // International Conference on Information and Knowledge Management, Proceedings – 2019. – P. 1061–1070.

70. Bouaziz A. Short text classification using semantic random forest / A. Bouaziz, C. Dartigues-Pallez, C. Costa Pereira, F. Precioso, P. Lloret // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) – 2014. – V. 8646 LNCS. – P. 288–299.

71. Lai S., Xu L., Liu K. Z.J. Recurrent convolutional neural networks for text classification / S. Lai, L. Xu, K. Liu // Twenty-Ninth AAAI Conference on Artificial Intelligence – 2015. – P. 2267–2273.

72. Alqaraleh S. Classification of Turkish text using machine learning: A case study using disasters tweets / S. Alqaraleh // International Journal of Scientific and Technology Research – 2020. – V. 9, № 3. – P. 4953–4956.

73. Li Y.H. Classification of text documents / Y.H. Li, A.K. Jain // Computer Journal – 1998. – V. 41, № 8. – P. 537–546.

74. Xia R. Ensemble of feature sets and classification algorithms for sentiment classification / R. Xia, C. Zong, S. Li // Information Sciences – 2011. – V. 181, № 6. – P. 1138–1152.

75. Pratama B. Y. Personality classification based on Twitter text using Naive Bayes, KNN and SVM / B. Y. Pratama, R. Sarno // Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015 – 2016. – P. 170–174.

76. Telnoni P. A. Comparison of Machine Learning Classification Method on Text-based Case in Twitter / P. A. Telnoni, R. Budiawan, M. Qana'a // Proceeding – 2019 International Conference on ICT for Smart Society: Innovation and Transformation Toward Smart Region, ICISS 2019 – 2019.

77. Liu Z. Study on SVM compared with the other text classification methods / Z. Liu, X. Lv, K. Liu, S. Shi // 2nd International Workshop on Education Technology and Computer Science, ETCS 2010 – 2010. – V. 1. – P. 219–222.

78. Liu C. Quality-related English Text Classification Based on Recurrent Neural Network / C. Liu, X. Wang // Journal of Visual Communication and Image Representation – 2019. – P. 1–7.

79. Lin Y. Research on text classification based on SVM-KNN / Lin Y., Wang J. // Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS – 2014. – P. 842–844.

80. Srinivas M. Efficient text classification using best feature selection and combination of methods , 2009. – P. 437–446.

81. Šubelj L. Clustering scientific publications based on citation relations: A systematic comparison of different methods / L. Šubelj, N.J. Eck, L. Waltman // PLoS ONE – 2016. – V. 11, № 4 – P. 1–23.

82. Liu X. Hybrid Clustering by Integrating Text and Citation Based Graphs in Journal Database Analysis/ X. Liu, S. Yu, Y. Moreau, F. Janssens, B. De Moor, W. Glänzel // IEEE International Conference on Data Mining Workshops (ICDM Workshops), 2009. – P. 521–526.

83. Waltman L. A principled methodology for comparing relatedness measures for clustering publications / L. Waltman, K. W. Boyack, G. Colavizza, N. J. Eck // Quantitative Science Studies – 2020. – P. 1–23.

84. Velden T. Mapping the cognitive structure of astrophysics by infomap clustering of the citation network and topic affinity analysis / T. Velden, S. Yan, C. Lagoze // Scientometrics – 2017. – V. 111, № 2. – P. 1033–1051.

85. Boyack K. W. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches / K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Börner // PLoS ONE – 2011. – V. 6, № 3.

86. Tshitoyan V. Unsupervised word embeddings capture latent knowledge from materials science literature / V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain // Nature – 2019. – V. 571, № 7763. – P. 95–98.

87. Borrajo L. Improving imbalanced scientific text classification using sampling strategies and dictionaries. / L. Borrajo, R. Romero, E. L. Iglesias, C. Redondo // Journal of integrative bioinformatics – 2011. – V. 8, № 3. – P.176.

88. Wang S. Clustering articles based on semantic similarity / S. Wang, R. Koopman // Scientometrics – 2017. – V. 111, № 2. – P. 1017–1031.

89. Лавренов А. О. Классификация текстов, содержащих формулы / А. О. Лавренов, Б. В. Олейников // Образовательные ресурсы и технологии – 2014. – Т. 1 – № 4 – С.108–114.

90. Glänzel W. Using hybrid methods and ‘core documents’ for the representation of clusters and topics: the astronomy dataset / W. Glänzel, B. Thijs B. // Scientometrics – 2017. – V. 111, № 2. – P. 1071–1087.

91. Пескова О. В. Автоматизация работы с классификаторами документов библиотеки МГТУ им. Н.Э. Баумана / О. В. Пескова // Культура народов Причерноморья – 2004. – Т. 2, № 48. – С. 38–41.

92. УДК, ББК, ISBN – обязательные элементы выходных сведений издания [Электронный ресурс]. URL: https://www.ipu.ru/structure/information-services/polygraphy/20804 (дата обращения: 07.05.2020).

93. Гиляревский Р.С. Тематическая систематизация книжной продукции на основе УДК / Р. С. Гиляревский, О. А. Антошкова, Т. С. Астахова, В. Н. Белоозеров // Книга. Культура. Образование. Инновации ("КРЫМ-2016"). Материалы Второго Международного профессионального форума, Судак, 4-12 июня, 2016. – 2016. – С. 161–163.

94. Сукиасян Э. Р. Классификация Библиотеки Конгресса США / Э. Р. Сукиасян // Научные и технические библиотеки – 1998. – № 5 – С. 59–69.

95. 1297.0 – Australian and New Zealand Standard Research Classification (ANZSRC), 2008 [Электронный ресурс].
URL: https://www.abs.gov.au/Ausstats/abs@.nsf/Latestproducts/1297.0Main Features32008?opendocument&tabname=Summary&prodno=1297.0&issue=2008 (дата обращения: 07.05.2020).

96. Паспорта научных специальностей [Электронный ресурс]. URL: https://teacode.com/online/vak/ (дата обращения: 07.05.2020).

97. ОКСО — Общероссийский классификатор специальностей по образованию [Электронный ресурс]. URL: https://classifikators.ru/okso (дата обращения: 07.05.2020).

98. Государственный рубрикатор научно-технической деятельности 2020 [Электронный ресурс]. URL: http://grnti.ru (дата обращения: 07.05.2020).

99. Revised field of science and technology (FOS) classification in the Frascati Manual [Электронный ресурс]. URL: http://www.oecd.org/science/inno/38235147.pdf (дата обращения: 07.05.2020).

100. Proposed international standard nomenclature for fields of science and technology [Электронный ресурс]. URL: https://unesdoc.unesco.org/ark:/48223/pf0000082946 (дата обращения: 07.05.2020).

101. Parfenova S. L. Methodical approach to the formation of rubricators-adapter for analysis of Web of science and Scopus area in terms of priorities the strategy of scientific and technological development of the Russian Federation / S. L. Parfenova, V. N. Dolgova, V. V. Bogatov, A. V. Khaltakshinova V. Y. Korobatov // The Economics of Science – 2018. – V. 4, № 2. – P. 143–153.

102. Scopus. Руководство по охвату контента. [Электронный ресурс]. URL: http://elsevierscience.ru/files/Scopus_Content_Guide_Rus_2017.pdf (дата обращения: 07.05.2020).

103. Wang Q. Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus / Q. Wang, L. Waltman // Journal of Informetrics – 2016. – V. 10, № 2. – P. 347–364.

104. Tsvetkova V. Combining classification systems and building the array of keyworks for defining the space of microbiological knowledge / V. Tsvetkova, T. Kharybina, Y. Mokhnacheva, E. Beskaravaynaya, I. Mitroshina // Scientific and Technical Libraries – 2019. – V. 1 – № 11 – P.25–43.

105. Abrizah A. LIS journals scientific impact and subject categorization: A comparison between Web of Science and Scopus / A. Abrizah, A. N. Zainab, K. Kiran, R. G. Raj // Scientometrics – 2013. – T. 94, № 2 – C. 721–740.

106. Bartol T. Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia / T. Bartol, G. Budimir, D. Dekleva-Smrekar, M. Pusnik, P. Juznic // Scientometrics – 2014. – V. 98, № 2. – P. 1491–1504.

107. Chung J. A Bibliometric Analysis of the Literature on Open Access in Scopus / J. Chung, M.-Y. Tsay // Qualitative and Quantitative Methods in Libraries – 2015. – V. 4, № 0. – P. 821–841.

108. Martín-Martín A. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories / A. Martín-Martín, E. Orduna-Malea, M. Thelwall, E. Delgado López-Cózar // Journal of Informetrics – 2018. – V. 12, № 4. – P. 1160–1177.

109. Bakri A. Publication Productivity Pattern of Malaysian Researchers in Scopus from 1995 to 2015 / A. Bakri, N. M. Azura, M. Nadzar, R. Ibrahim, M. Tahira // Journal of Scientometric Research – 2017. – V. 6, № 2 – P. 86–101.

110. Bonaccorsi A. Explaining the transatlantic gap in research excellence / A. Bonaccorsi , T. Cicero, P. Haddawy, S. U. L. Hassan // Scientometrics – 2017. – V. 110, № 1. – P. 217–241.

111. Hassan S. U. Measuring international knowledge flows and scholarly impact of scientific research / S. U. Hassan, P. Haddawy // Scientometrics – 2013. – V. 94, № 1. – P. 163–179.

112. Thelwall M. The influence of highly cited papers on field normalised indicators / M. Thelwall // Scientometrics – 2019. – V. 118, № 2. – P. 519–537.

113. Cicero T. On the use of journal classification in social sciences and humanities: evidence from an Italian database / T. Cicero, M. Malgarini // Scientometrics – 2020. – № 0123456789.

114. Mendes A. Science classification, visibility of the different scientific domains and impact on scientific development / A. Mendes // Revista de Enfermagem Referência – 2016. – V. IV Série, № 10. – P. 143–152.

115. Martínez-Frías J. Classifying science and technology: Two problems with the UNESCO system / J. Martínez-Frías, D. Hochberg // Interdisciplinary Science Reviews – 2007. – V. 32, № 4. – P. 315–319.

116. Thaper N. Using Compression For Source Based Classification Of Text / N. Thaper – 2001.

117. Кукушкина О. В. Определение авторства текста с использованием буквенной и грамматической информации / О. В. Кукушкина, А. А. Поликарпов, Д. В. Хмелёв // Пробл. передачи информ. – 2001. – Т. 37, № 2. – С. 96–109.

118. Ryabko B. Compression-based methods of statistical analysis and prediction of time series / B. Ryabko, J. Astola, M. Malyutov – Springer International Publishing, 2016. – 144 p.

119. Cilibrasi R. Clustering by compression / R. Cilibrasi, P.M.B. Vitányi // IEEE Transactions on Information Theory – 2005. –V. 51, № 4. – P. 1523–1545.

120. Cilibrasi R. Algorithmic clustering of music based on string compression / R. Cilibrasi, P. Vitányi, R. Wolf // Computer Music Journal – 2004. – Т. 28, № 4. – С. 49–67.

121. Li M. The similarity metric / M. Li, X. Chen, X. Li, B. Ma, P.M.B. Vitányi // IEEE Transactions on Information Theory – 2004. – V. 50, № 12. – P. 3250–3264.

122. Батура Т. В. Методы автоматической классификации текстовAutomatic text classification methods / Т. В. Батура // Международный журнал «Программные продукты и системы» – 2017. – Т. 23, № 30. – С. 85–99.

123. Marton Y. On Compression-Based Text Classification / Y. Marton, N. Wu, L. Hellerstein // European Conference on Information Retrieval. – P. 300-314

124. Яндекс MyStem. [Электронный ресурс]. URL: https://yandex.ru/dev/mystem/ (дата обращения: 10.07.2020).

125. Hall G. M. How to Write a Paper / G. M. Hall – A John Wiley & Sons, Ltd., 2013. – 35 p.

126. Bordignon F. Tracking content updates in Scopus (2011-2018): A quantitative analysis of journals per subject category and subject categories per journal / F. Bordignon

// 17th International Conference on Scientometrics and Informetrics, ISSI 2019 – Proceedings – 2019. – V. 2. – P. 1630–1640.

127. Журнал «Геология и геофизика» [Электронный ресурс]. URL: https://www.sibran.ru/journals/GiG/ (дата обращения: 30.07.2020).

128. scikit-learn: machine learning in Python [Электронный ресурс]. URL: https://scikit-learn.org/stable/ (дата обращения: 31.07.2020).

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. – № 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, № 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. – № 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, Gerrmany. – Aachen, 2017. – P. 121–125.

Прочие публикации

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.
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.

Приложение А. Метод классификации, основанный на сжатии данных (batch-скрипт)

```
set mask=*.txt
for /f "delims=" %%a in ('dir /b /ad /os') do (
	if %%a NEQ test (
		for /f %%b in ('dir /b /s /a-d "test\%mask%"') do (
			for %%q in (%%b) do set testsize=%%~zq
			copy "%%b" "%%a"
			rar a -s -m5 -mc0:0t+ %%a-%%~nb.rar %%a
			rar v %%a-%%~nb.rar
			for /f "tokens=3,8" %%x in ('rar v %%a-%%~nb.rar') do (
				if %%y equ %%a\%%~nb.txt (
					set pack=%%x
					@echo !pack!
				)
			)
			@echo %%a; %%~nb; !testsize!; !pack!   >>comment.txt
			del "%%a\%%~nb.txt"
			del "%%a-%%~nb.rar"
		)
	)
)
```

Приложение Б. Основные функции для обработки результатов классификации (на языке Python)

```
def get_txt(subjarea):
    collection_core = db[subjarea]
    for document in collection_core.find():
        sub=document['subjArea']
        path = '' + sub
        if not os.path.exists(path):
            os.makedirs(path)
        s = path + '\\'
        m = document['eid']
        filename = s + m + '.txt'
        f = open(filename, 'w', encoding='utf-8')
        t = document['descr']
        f.write(str(t))
        f.close()

def compress_files():
    import subprocess

    p = subprocess.Popen('1.bat', shell=True, stdin=subprocess.PIPE,
                         cwd='')
    stdout, stderr = p.communicate()

def get_classif_res(path_csv):
    import csv
    with open(path_csv, newline='') as csvfile:
        reader = csv.reader(csvfile, delimiter=';')
        for row in reader:
            area = row[0]
            test = row[1]
            test_area, test_name = test.split("+")

            publ = {
                "test": test,
                "area": area,
                "core_area": test_area,
                "test_name": test_name,
                "perc": int(row[3]) / int(row[2])
            }
            rec_publ = collection_temp.insert_one(publ)
```

```
result = collection_temp.aggregate([
    {"$group": {
        "_id": {
            "test_name": "$test_name"
        },
        "areas": {"$min": {"perc": "$perc", "area": "$area"}},
        "count": {"$sum": 1}
    }}, {"$out": "res"}])
```

Приложение В. Основные функции для извлечения данных через API Scopus (на языке Python)

```
from pymongo import MongoClient

def get_scopus_data(subjarea):
    doc_count=0
    n='*'
    while True:
        resp =
requests.get("http://api.elsevier.com/content/search/scopus?cursor/view=COMPLETE&
cursor="+n+"&count=25&query=subjarea("
                    + subjarea +
                    ")&field=eid,dc:title,subtypeDescription,citedby-
count,dc:description,prism:publicationName&date=2009-2018&sort=citedby-count",
                    headers={'Accept': 'application/json',
                        'X-ELS-APIKey': MY_API_KEY})

        results = resp.json()
        print(results)
        n = results['search-results']["cursor"].get('@next')

        for r in results['search-results']["entry"]:

            sp=r['eid']
            print(sp)
            if ((int('citedby-count' in r) > 0)  and collection_core.find({'eid':
sp}).limit(1).count() < 1):

                resp1 = requests.get("http://api.elsevier.com/content/abstract/eid/" +
str(r['eid']),
                                headers={'Accept': 'application/json',
                                    'X-ELS-APIKey': MY_API_KEY})

                res = resp1.json(object_hook=remove_dot_key)
                print(res)
                leng = len(res['abstracts-retrieval-response']['subject-areas']['subject-area'])
                des=res['abstracts-retrieval-response']['coredata']

                sum=0
                for i in range (0,leng):

                    if (res['abstracts-retrieval-response']['subject-areas']['subject-
```

```
area'][i].get('@abbrev') == subjarea):

                    sum=sum+1

            if (leng == sum and 'dc:description' in des):

                    filter = {"subjArea": subjarea}
                    doc_count = collection_core.count_documents(filter)

                    publ = {
                            "eid": str(res['abstracts-retrieval-response']['coredata']['eid']),
                            "title": str(res['abstracts-retrieval-
response']['coredata']['dc:title']),
                            "subtypeDescription": str(
                               res['abstracts-retrieval-
response']['coredata']['subtypeDescription']),
                            "publicationName": str(
                               res['abstracts-retrieval-
response']['coredata']['prism:publicationName']),
                            "cited": str(res['abstracts-retrieval-
response']['coredata']['citedby-count']),
                            "descr": str(res['abstracts-retrieval-
response']['coredata']['dc:description']),
                            "subjcode":
                               res['abstracts-retrieval-response']['subject-areas']['subject-
area'],
                            "numberofCats": str(leng),
                            "subjArea": str(subjarea)

                    }
                    rec_publ = collection_core.insert_one(publ)

                else:

                    return
            else:
                pass
```

Приложение Г. Акты о внедрении

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ГОСУДАРСТВЕННАЯ
ПУБЛИЧНАЯ
НАУЧНО–ТЕХНИЧЕСКАЯ
БИБЛИОТЕКА
СИБИРСКОГО ОТДЕЛЕНИЯ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ГПНТБ СО РАН)**

Восход ул., д. 15, Новосибирск, 630200
Тел./факс (383) 266-18-60
e-mail: office@gpntbsib.ru; http://www.spsl.nsc.ru
ОКПО 03533820; ОГРН 1025401929981

АКТ
о внедрении системы тематической классификации научных текстов

Настоящий акт свидетельствует о том, что результаты диссертационного исследования Селивановой Ирины Вячеславовны «Методы тематической классификации научных текстов на основе теоретико-информационного подхода», специальность: 05.13.17 – «Теоретические основы информатики», внедрены в процессе реализации базового проекта научно-исследовательских работ ГПНТБ СО РАН № 0334-2019-0006 «Наукометрический анализ публикационного потока российских исследователей и факторов его трансформации, изучение способов и методов повышения публикационной активности, развития российской научной периодики».

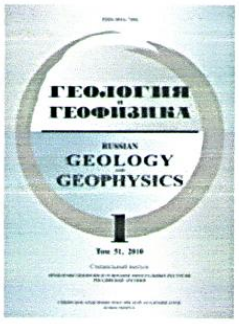
Метод на основе сжатия данных, описываемый в диссертационном исследовании, применялся для проверки и уточнения тематической классификации публикаций в библиографических базах данных (таких как Web of Science или Scopus). Полученные Селивановой И.В. экспериментальные данные показали недостатки существующих подходов к тематической классификации научных текстов и характерные примеры ошибок в выдаче поисковых запросов.

Использование указанных результатов позволяет повысить точность наукометрических исследований, улучшить качество справочно-поискового аппарата и создает новые возможности для повышения эффективности процессов каталогизации.

Председатель комиссии:
Д.п.н., профессор, г.н.с. О.Л. Лаврик

Члены комиссии
Д.п.н., г.н.с. Е.Б. Артемьева
К.т.н., с.н.с. С.Р. Баженов
К.филол.н., зав отделом редких книг и рукописей А.Ю. Бородихин
Д.и.н., проф., г.н.с. С.Н. Лютов
Д.п.н., зам. директора по научной работе Н.С. Редькина

Секретарь
К.п.н. Н.В. Махотина



Сибирское отделение Российской академии наук
Новосибирский государственный университет
Институт геологии и минералогии им. В.С. Соболева СО РАН
Институт нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН

Редакционная коллегия журнала
ГЕОЛОГИЯ и ГЕОФИЗИКА
Russian Geology and Geophysics

630090, Новосибирск, просп. Морской, 2; тел. 8(383) 330-81-27; e-mail: geo@sibran.ru

АКТ
о внедрении результатов кандидатской диссертации И. В. Селивановой

Настоящим актом подтверждается, что результаты диссертационного исследования И.В. Селивановой по теме «Методы тематической классификации научных текстов на основе теоретико-информационного подхода», специальность: 05.13.17 – «Теоретические основы информатики», используются при тематическом анализе публикаций англоязычной версии журнала «Геология и геофизика».

Изначально в базе данных Scopus у журнала было проставлено только две тематики: «Geology» и «Geophysics». Анализ принадлежности публикаций другим категорий проводился при помощи алгоритма классификации научных текстов, основанного на сжатии данных, которому посвящено диссертационное исследование И.В. Селивановой. Для лучшей точности классификации использовался метод подбора «ядра», описанный в диссертационной работе, где в качестве обучающих выборок были использованы аннотации самых высокоцитируемых публикаций из категорий, принадлежащих области «Earth and Planetary Science».

Внедренный алгоритм классификации позволил улучшить точность тематического распределения публикаций англоязычной версии журнала «Геология и геофизика».

Главный редактор
Академик РАН Н.В. Соболев

ПОДПИСЬ УДОСТОВЕРЯЮ
ЗАВ. КАНЦЕЛЯРИЕЙ
ШИПОВА Е.Е.
13.10.2020г.