АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2:004.65

И.В. Селиванова, Д.В. Косяков, А.Е. Гуськов

Классификация научных текстов на основе компрессии аннотаций публикаций

Исследуется возможность установления смысловой близости научных текстов методом их автоматической классификации, основанным на сжатии аннотаций. Идея метода состоит в том, что алгоритмы компрессии типа PPM (prediction by partial matching) сжимают терминологически близкие тексты существенно лучше, чем далекие. Если для каждой классифицируемой тематики будет сформировано ядро публикаций (аналог обучающей выборки), то наилучшая доля сжатия будет указывать на принадлежность классифицируемого текста к соответствующей тематике. Было определено 30 тематических категорий, каждой из них в базе данных Scopus получены аннотации около 500 публикаций, из которых разными способами выбирались 100 аннотаций для ядра и 20 аннотаций для тестирования. Установлено, что построение ядра на основе высокоцитируемых публикаций выявляет до 12% ошибок против 32% при случайной выборке. На качество классификации влияет и изначальное количество категорий: чем меньше категорий участвует в классификации и чем больше терминологические различия между ними, тем выше её качество.

Ключевые слова: *классификация научных текстов, сжатие текстов, библиографические базы данных, Scopus*

ВВЕДЕНИЕ

В последние десятилетия в связи с ростом количества научных публикаций проблема классификации текстов становится особенно актуальной. При этом классифицируются как целые художественные или поэтические тексты [1, 2], так и короткие текстовые сообщения, например, СМС или твиты [3-5]. Но единого метода классификации, который бы сразу показал и высокую эффективность, и низкие трудозатраты, до настоящего момента найдено не было.

Одной из областей, где эта задача имеет особую важность, является классификация научных документов. Неверно классифицированные публикации затрудняют поиск необходимых ученому статей, что является причиной потери актуальных исследований в интересующей его области наук.

В работах [6, 7] был предложен метод классификации научных текстов, основанный на сжатии информации. Он показал более чем 90%-эффективность и низкие трудозатраты при классификации англоязычных текстов Архива научных публикаций *arXiv.org*. Одним из недостатков этих работ являлось применение метода только к полнотекстовым документам. Полные тексты статей, из-за их объема, часто содержат много лишних фраз, что может привести к некоторым ошибкам при классификации. Аннотации публикаций, наоборот, должны содержать только ключевые моменты, используемые в статье [8, с. 35], что может облегчить автоматическую классификацию научных работ. Более того, во многих научных библиографических базах данных (ББД) не всегда удается получить полный текст статьи и часто доступна только ее аннотация.

Цель нашего исследования – применение метода, основанного на алгоритмах сжатия, к классификации аннотаций публикаций, индексируемых в международных ББД. Первоначальные результаты этой работы были представлены на XVII Российской конференции «Распределенные информационно-вычислительные ресурсы» *DICR-2019*.

Наиболее авторитетными ББД для ученых являются *Web of Science* (*WoS*) и *Scopus*. Они служат источником данных для множества наукометрических исследований, различных рейтингов университетов, а также используются при оценке публикационной активности исследователей, организаций и стран.

Тем не менее, классификация публикаций в этих библиографических базах данных вызывала большое количество вопросов. Одним из главных ее недостатков являлось то, что как в *WoS*, так и в *Scopus* опре-

деление тематики публикации происходило только на уровне журнала, в котором она издана [9, 10]. Особенно негативно это сказывалось на статьях из мультидисциплинарных журналов. Публикация, которая могла иметь только одно направление, была отнесена ко всем тем же предметным рубрикам, что и этот журнал. Это создавало «замусоренность» научных направлений публикациями, которые, возможно, не имели к ним никакого отношения.

Для улучшения качества классификации в *Scopus* в 2017 г. введена новая система классификации *Topic Prominence in Science* [11], которая была построена подобно методике, описанной в работе [12]. Авторы предложили способ присвоения публикациям отдельных категорий, состоящий из трех этапов: на первом этапе происходило определение связанности публикаций путем использования прямого цитирования; на втором – происходила иерархическая кластеризация публикаций; третий этап был посвящен названию получившихся кластеров метками на основе терминов заголовков и аннотаций публикаций, находящихся в этой группе.

Важную роль для улучшения точности классификации научных документов играет выбор системы классификации. Эти системы могут быть разделены на несколько типов:

1) Библиотечные классификаторы (например, УДК). За основу этой системы взята десятичная классификация, разработанная в 1876 г. американским библиографом М. Дьюи. Центральная часть УДК – основные таблицы, охватывающие всю совокупность знаний и построенные по иерархическому принципу деления от общего к частному с использованием цифрового десятичного кода [13]. Также к библиотечным классификаторам относится ББК – национальная классификационная система России. Принцип ее формирования аналогичен УДК.

2) Национальные классификаторы, например, классификация Fields of research (*FOR*) из *Australian and New Zealand Standard Research Classification* (*ANZSRC*), представляющая собой иерархическую классификацию с тремя уровнями: на первом – находятся обширные научные направления, на втором – связанные с ними группы, – на третьем расположены более узкие области. *FOR* применяется, например, в платформе *Dimensions* [14]. Другими примерами национальных классификаторов является Номенклатура научных специальностей России, используемая Высшей аттестационной комиссией Минобрнауки (ВАК) [15], Общероссийский классификатор специальностей по образованию (ОКСО), который сопоставлен с Международной стандартной классификацией образования (МСКО) [16] и Государственный рубрикатор научно-технической информации (ГРНТИ) [17].

3) Международные классификаторы, к которым относится *Field of science and technology* (*FOS*) – система классификации, опубликованная Организацией экономического сотрудничества и развития в 2002 г. и измененная в 2007 г. [18]. В основе этой классификации лежат 6 научных направлений, разделенных на 42 области. Другим примером международного классификатора является номенклатура *UNESCO* – система, разработанная ЮНЕСКО для классификации научных работ и диссертаций [19].

4) Классификаторы в международных ББД. Эти классификаторы, как и предыдущие, 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 [20]. В *Scopus* используется *All Science Journals Classification* (*ASJC*). Она включает издания, распределенные по четырем общим научным направлениям: биологические науки, физические науки, медицина, социальные и гуманитарные науки. Эти направления разделены на 27 крупных предметных областей и более 300 узких категорий [21]. 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 проблема была отмечена в 2016 г. в работе [22], где авторы предлагали либо слить эти категории, либо обозначить более четкие различия между ними.

Важность выбора системы классификации отмечается в работе [23]. 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 [24] показаны две основные проблемы номенклатуры *UNESCO*: первая связана с тем, что в этой классификации в крупных научных областях могут быть потеряны более мелкие категории, вследствие чего классификация будет недостаточно полной; вторая – заключается в том, что в этой классификации отсутствуют области наук, которые появились недавно, что также существенно ухудшает качество классификации.

Для решения задачи классификации текстовых документов применяется множество различных методов. Одним из наиболее часто используемых алгоритмов является метод *k*-ближайших соседей и его модификации, где классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки. На этом методе базируется, например, работа [25]. Другим алгоритмом является байесовская классификация, которая работает на вычислении апостериорных вероятностей классов. Этот алгоритм применен в работе [26]. Представителем линейных классификаторов служит метод опорных векторов, который заключается в построении гиперплоскости, разделяющей объекты выборки наиболее оптимальным способом. Сравнение алгоритмов байесовской классификации и метода опорных векторов при классификации заголовков статей приводится в работе [27].

В последние время для решения задачи классификации все чаще применяются нейронные сети. В статье [28] для решения задачи классификации текстов предлагается использовать рекуррентные свёрточные нейронные сети. Ее 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 [29].

В среднем точность различных алгоритмов классификации текстовой информации варьируется от 70% до 86% [30, 31].

Основное преимущество классификации научных публикаций – 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 относящихся к ней статей.

Для классификации научных публикаций могут быть использованы методы, базирующиеся на:

- цитировании, которое включает прямое цитирование, ко-цитирование и библиографическое сочетание [32-34]. Чаще всего источником данных для таких исследований является ББД *WoS*;
- метаданных публикаций (списков соавторов, названий публикаций, ключевых слов и др.). Такой метод применяется в работе [35] при кластеризации биомедицинских публикаций в базе данных *MEDLINE*;
- комбинации использования цитирования и метаданных. Например, в работе [36] авторы применяют этот подход для улучшения качества классификации документов в ACM *Digital Library*;
- аннотациях и полных текстах публикаций. Классификацией аннотаций публикаций из научных направлений *Materials science, Physics и Chemistry* ББД *Scopus* занимались авторы статьи [37]. 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 [38]. Полнотекстовые научные документы классифицировались в работах [7, 39].

1. МЕТОДЫ И ДАННЫЕ КЛАССИФИКАЦИИ

1.1. Метод

Рассмотрим подробнее метод, основанный на алгоритмах сжатия. Пусть есть n научных областей $X_1, \ldots, X_n$, для каждой из которых определен характерный для нее набор текстов (файлы, с которыми составляют «ядро» это области). Также есть некоторый тестовый файл y, тематику которого нужно определить, и архиватор φ, который может быть применен для сжатия любого множества текстов.

Работа метода состоит в том, что тестовый файл y начинает последовательно сжиматься с каждым из n ядер при помощи архиватора φ. В итоге, область тестового файла y определяется тем ядром, с которым он имеет наилучшее сжатие.

Для работы метода были выбраны следующие параметры:

- Архиватор – WinRAR при максимальном значении памяти 128 Мбайт, алгоритм сжатия – PPMd [6, 7];
- Объем ядра – 100 файлов. Этот объем является рекомендованным в работе [7], так как при большем количестве файлов в ядре качество классификации практически не улучшается, а время работы алгоритма увеличивается.

1.2. Данные

Источником информации для настоящего исследования служит ББД Scopus. Процесс извлечения данных состоит из трех этапов.

1. Извлечение информации о названии журнала, eid публикации, цитировании, категории через *Scival* – аналитический инструмент компании *Elsevier*, основанный на данных *Scopus*. Данные по 30 категориям были выбраны за период 2009-2018 гг. приведены в табл. 1.
2. Формирование файлов аннотаций путем получения их текстов через *Scopus Abstract Retrieval API*
3. Удаление файлов аннотаций с отсутствующим текстом

Всего было выгружено 15 тыс. файлов (по 500 для каждой категории).

1.3. Процесс классификации

Для проверки того, насколько эффективно метод, основанный на алгоритмах сжатия, работает на аннотациях публикаций, индексируемых в ББД *Scopus*, классификация проводилась для:

1) тестовых файлов с одной категорией,
2) тестовых файлов с несколькими категориями.

В первом случае осуществлялась проверка работы метода путем классификации файлов с одной категорией, а также подбор оптимальных по составу ядер.

На втором этапе классификация проводилась для тестов с несколькими категориями, ядра для которой были подобраны на первом этапе. Этот случай полезен при проверке качества классификации публикаций из мультидисциплинарных изданий.

Рассмотрим эти два этапа подробнее.

1.3.1. Классификация тестовых файлов с одной категорией

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

- Ошибки I типа связаны с неправильным определением категории внутри научного направления, например, вместо категории *Aquatic Science* определилась *Plant Science* из той же области *Agricultural and Biological Sciences*.
- К ошибкам II типа отнесены те тестовые файлы, у которых определилась другая область при одном научном направлении. 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 ошибок – самый серьезный, к нему отнесены те тестовые файлы, у которых ошибка в классификации произошла на самом верхнем уровне. Например, вместо научного направления *Physical Sciences* определилось *Social Sciences*.

Таблица 1

Исследуемые области наук по уровням классификации

Направление	Область	Категория
Life Sciences	Agricultural and Biological Sciences	Animal Science, 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, 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
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

Таблица 2

Пример расчета нормированного коэффициента сжатия

Области теста	Algebra and Number Theory, %	Geometry and Topology, %	мин, %	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

Классификация осуществлялась с использованием ядер двух типов:

1) **произвольные ядра** были составлены из аннотаций публикаций, отобранных случайным образом.

2) **подобранные ядра** были составлены из 100 самых высокоцитируемых публикаций в каждой области наук. Такой подход, предположительно, увеличивает качество ядра в связи с тем, что публикации из той же области, цитирующие отобранные в ядро статьи, с высокой долей вероятности наследуют и характерную лексику.

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 [40] и названий издательств, присутствующих в текстах аннотаций (на-

пример, © 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 знаков препинания: ., !, ?, :, ,, -.

1.3.2. Классификация тестовых файлов с несколькими категориями

Для классификации тестовых файлов с несколькими категориями использовались только ядра с самыми высокоцитируемыми публикациями.

Отбор 20 тестов осуществлялся произвольным образом из публикаций, у которых по меньшей мере две категории совпадало с категориями из табл. 1. Суммарно, как и в предыдущих случаях, было отобрано 600 тестов.

Введем следующие группы результатов:

- у тестового файла верно определилось не менее 50% указанных категорий. Например, у теста было указано 4 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 два из трех 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* (табл. 2). 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*;
- 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.

1.4. Влияние количества категорий на качество классификации

В работе [30] 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 нами был проведен эксперимент с последовательным увеличением количества рассматриваемых категорий с 5 до 30 с шагом в 5 (5, 10, 15, 20, 25, 30). Для тестовых файлов были отобраны случайным образом по 20 публикаций из первых 5 категорий (всего 100 файлов).

2. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

2.1. Классификация тестовых файлов с одной категорией

2.1.1. Классификация при произвольных ядрах

В табл. 3 приведены результаты классификации тестовых файлов с одной категорией при произвольных ядрах по типам ошибок.

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 (рис. 1).

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* (табл. 4).

Тексты двух аннотаций категории *Marketing*, с которыми у тестового файла произошло лучшее попарное сжатие, приведены на рис. 2, 3. 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.

2.1.2. Классификация при подобранных ядрах

В табл. 5 приведены результаты классификации тестовых файлов с одной категорией при ядрах, в состав которых входят самые высокоцитируемые публикации.

Использование подобранных ядер улучшило результаты классификации на 20%. Число ошибок III типа уменьшилось в 3 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* (рис. 4).

Таблица 3

Результаты классификации тестовых файлов при произвольных ядрах с различным числом цитирования

Категория	Общее кол-во тестов	Кол-во ошибок	Ошибки		
			I типа	II типа	III типа
Algebra and Number Theory	20	7	5	2	0
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
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, %		32	12	13	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.

Рис. 1. Аннотация публикации с eid=2-s2.0-67651018249

Таблица 4

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

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.

Рис. 2. 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.

Рис. 3. Аннотация публикации категории Marketing с eid=2-s2.0-67149130202

Таблица 5

Результаты классификации тестовых файлов при ядрах с самыми высокоцитируемыми публикациями

Категория	Общее кол-во тестов	Кол-во ошибок	Ошибки		
			I типа	II типа	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
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

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* ***cost****s, 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.*

Рис. 4. Аннотация теста с eid=2-s2.0-70449527784 из категории *Sociology and Political Science* (**жирным** выделены термины, часто встречающиеся в категории *Aquatic Science*)

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.

Рис. 5. Текст аннотации публикации с eid=2-s2.0-77958562700 из категории Library and Information Sciences

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.

Рис. 6. Тестовый файл с eid=2-s2.0-77951062083 из категории *History*

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 значительно количество терминов, характерных для области *Computer Science* (рис. 5).

В некоторых закономерностях не было выявлено определения неверной категории. Например, тестовый файл из *History* неверно отнесся к *Condensed Matter Physics* (рис. 6).

Стоит отметить, что неверно определившийся файл из пункта 2.1.1. с eid=2-s2.0-67651018249 при подобранных ядрах определился верно. Таким образом, состав ядра оказывает большое влияние на качество классификации.

2.1.3. Влияние на классификацию стоп-слов и названий издательств

Для изучения влияния присутствия названий издательств в аннотациях на качество классификации использовались подобранные ядра (см. п. 2.1.2).

В табл. 6 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 2 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 теста 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 не равномерно 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 2 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.

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

Таблица 6

Влияние стоп-слов и присутствия названий издательств на качество классификации

Категория	**Общее кол-во тестов**	**Кол-во ошибок**	**Количество ошибок**		
			со стоп-словами, без названий издательств	**без стоп слов, с названиями издательств**	**без стоп-словам и названий издательств**
Algebra and Number Theory	20	8	8	8	7
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
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
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

2.2. Классификация тестовых файлов с несколькими категориями

Результаты классификации тестовых файлов с несколькими категориями приведены в табл. 7.

Ошибочно определились 23% (140 из 600) тестовых файлов. 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 находящихся в разных научных направлениях. 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 рис. 7.

В других случаях ошибки её характер 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* (рис. 8).

2.3. Влияния количества ядер на качество классификации

На рис. 9 изображена зависимость количества ошибок от количества ядер, участвующих в классификации. Результаты показывают, что при расширении числа категорий также увеличивается и число ошибок.

Если при 5 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 5, так и при 30 ядрах безошибочно определялись тесты из категории *Surgery*.

Таким образом, изначальная выборка количества и состава категорий влияет на точность классификации.

Таблица 7

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

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.

Рис. 7. 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.

Рис. 8. Аннотация публикации с eid= 2-s2.0-79451471007

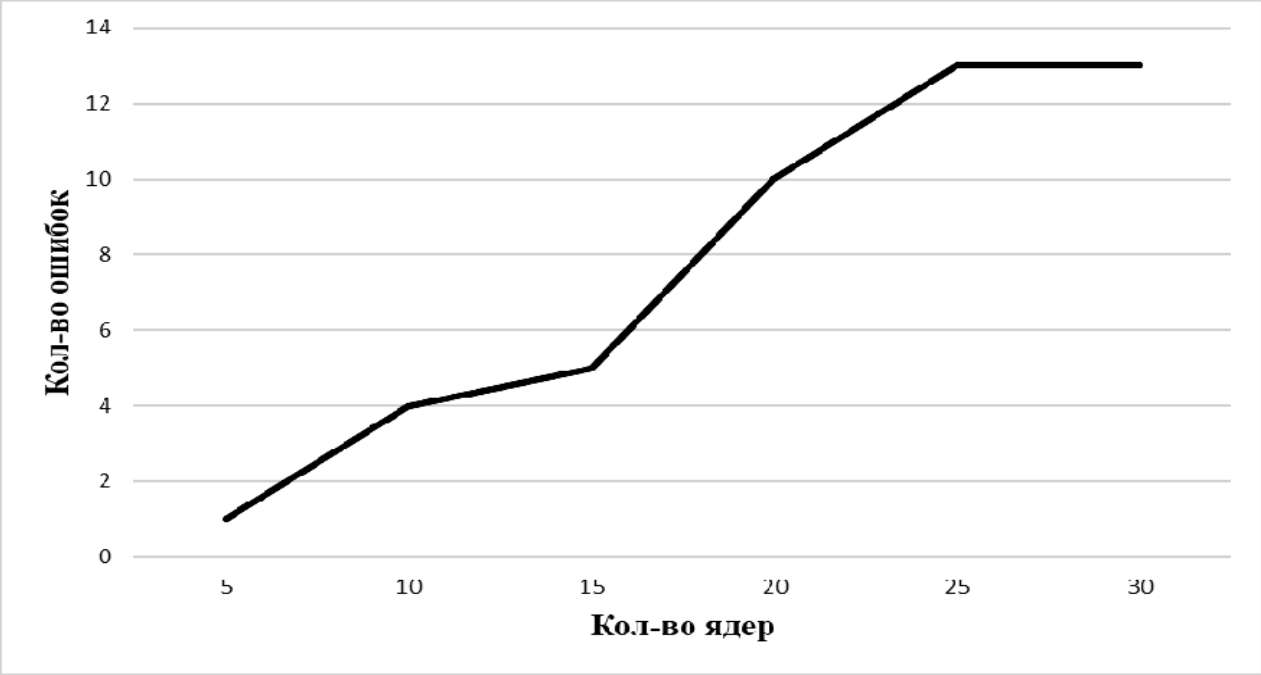

Рис. 9. Зависимость качества классификации от количества ядер

ЗАКЛЮЧЕНИЕ

Исследование показало, что метод классификации научных текстов, основанный на алгоритмах сжатия информации, показывает разную эффективность в зависимости от условий, в которых проводится классификация. 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%. Это, предположительно, связано с тем, что в таких аннотациях используется более характерная для области наук лексика, которая наследуется и остальными публикациями.

Удаление стоп-слов и названий издательств в большинстве случаев негативно сказывается на качестве классификации. Возможно, это связано с тем, что в названии издательств используются специальные термины, а также с тем, что длина аннотации уменьшается.

Помимо состава ядер на качество классификации влияет и изначальное количество категорий: чем меньше категорий участвует при классификации и чем больше терминологическое различие между ними, тем выше качество этой классификации.

Тем не менее, исходная классификация статей была сделана на основе журнальной классификации *Scopus*, которая также имеет ошибки, и их количество крайне трудно измерить. Это вносит свою объективную погрешность в наши исследования.

СПИСОК ЛИТЕРАТУРЫ

1. Барахнин В.Б., Кожемякина О.Ю., Пастушков И.С., Рычкова Е.В. Автоматизированная классификация русских поэтических текстов по жанрам и стилям // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. – 2017. – Т.15, №3. – С. 13-23.
2. Батура Т.В. Формальные методы определения авторства текстов // Вестник НГУ. Серия: Информационные технологии. – 2012. – Т.10, №4. – С. 81-94.
3. Dos Santos C.N., Gatti M. Deep convolutional neural networks for sentiment analysis of short texts // COLING 2014 – 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers. – 2014. – P. 69-78.
4. Sriram B., Fuhry D., Demir E., Ferhatosmanoglu H., Demirbas M. Short text classification in twitter to improve information filtering // SIGIR 2010 Proceedings – 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2010. – P. 841-842.
5. Kiritchenko S., Zhu X., Mohammad S.M. Sentiment analysis of short informal texts // Journal of Artificial Intelligence Research. – 2014. – Vol.50. – P. 723-762.
6. Рябко Б.Я., Гуськов А.Е., Селиванова И.В. Теоретико-информационный метод классификации текстов // Пробл. передачи информ. – 2017. – Т. 53, №3. – С. 100–111; Ryabko B.Y., Gus'kov A.E., Selivanova I.V. Information-Theoretic Method for Classification of Texts // Problems of Information Transmission. – 2017. – Vol.53, Iss. 3. – P.294-304. – URL: https://link.springer.com/article/10.1134/S0032946017030115.
7. Селиванова И.В., Рябко Б.Я., Гуськов А.Е. Классификация посредством компрессии: применение методов теории информации для определения тематики научных текстов // Научно-техническая информация. Сер. 2. – 2017. – № 6. – С.8-15; Selivanova I.V., Ryabko B.Ya., Guskov A.E. Classification by Compression: Application of Information-Theory Methods for the Identification of Themes of Scientific Texts // Automatic Documentation and Mathematical Linguistics. – 2017. – Vol. 51, № 3. – P.120-126.
8. Hall G.M. How to write a paper. – A John Wiley & Sons, Ltd., Publication, 2013. – 170 c.
9. Perianes-Rodriguez A., Ruiz-Castillo J. A comparison of the Web of Science and publication-level classification systems of science // Journal of Informetrics. – 2017. – Vol. 11, Iss.1. – P.32-45.
10. Shu F., Julien C.A., Zhang L., Qiu J., Zhang J., Lariviere V. Comparing journal and paper level classifications of science // Journal of Informetrics. – 2019. – Vol.13, Iss.1. – P.202-209.
11. Topic Prominence in Science стал доступен пользователям SciVal. – URL: http://elsevierscience.ru/news/428/topic-prominence-in-science-stali-dostupny-polzovatelyam-scival (дата обращения: 14.10.2019).
12. Waltman L., van Eck N.J. A new methodology for constructing a publication-level classification system of science // Journal of the American Society for Information Science and Technology. – 2012. – Vol.63, Iss.12. – P.2378-2392.
13. УДК, ББК, ISBN – обязательные элементы выходных сведений издания. – URL: https://www.ipu.ru/structure/information-services/polygraphy/20804 (дата обращения: 14.10.2019).
14. 1297.0 – Australian and New Zealand Standard Research Classification (ANZSRC), 2008. – URL: https://www.abs.gov.au/Ausstats/abs@.nsf/Latestproducts/1297.0Main%20Features32008?opendocument&tabname=Summary&prodno=1297.0&issue=2008 (дата обращения: 14.10.2019).
15. Паспорта научных специальностей. – URL: http://arhvak.minobrnauki.gov.ru/316 (дата обращения: 14.10.2019).
16. ОКСО – Общероссийский классификатор специальностей по образованию. – URL: https://classifikators.ru/okso (дата обращения: 14.10.2019).
17. ГРНТИ – Государственный рубрикатор научно-технической деятельности 2019. – URL: http://grnti.ru/ (дата обращения: 14.10.2019).
18. Revised field of science and technology (FOS) classification in the Frascati Manual. – URL: http://www.oecd.org/science/inno/38235147.pdf (дата обращения: 14.10.2019).
19. Proposed international standard nomenclature for fields of science and technology. – URL: https://unesdoc.unesco.org/ark:/48223/pf0000082946 (дата обращения: 14.10.2019).
20. Парфенова С.Л., Долгова В.Н., Богатов В.В., Халтакшинова Н.В., Коробатов В.Я. Методический подход к формированию рубрикаторов-переходников для анализа направлений *Web of Science* и *Scopus* в разрезе приоритетов Стратегии научно-технологического развития РФ // Экономика науки. – 2018. – Т.4, №2. – С.143-153.
21. Scopus. Руководство по охвату контента. – URL: http://elsevierscience.ru/files/ Scopus_Content_Guide_Rus_2017.pdf. – С. 21 (дата обращения: 14.10.2019).
22. Wang Q., Waltman L. Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus // Journal of Informetrics. – 2016. – Vol.10, Iss.2. – P.347-364.
23. Mendes A.C. Science classification, visibility of the different scientific domains and impact on scientific development Scopus // Revista de Enfermagem Referência. – 2016. – Vol.10, Iss.4. – P.143-149.
24. Martínez-Frías J., Hochberg D. Classifying science and technology: Two problems with the UNESCO system // Interdisciplinary Science Reviews. – 2007. – Vol.32, Iss.4. – P.315-319.
25. Tan S. Neighbor-weighted K-nearest neighbor for unbalanced text corpus // Expert Systems with Applications. – 2005. – Vol.28, Iss.4. – P.667-671.

26. Jiang L., Li C., Wanga S., Zhanga L. Deep feature weighting for naive Bayes and its application to text classification // Engineering Applications of Artificial Intelligence. – 2016. – Vol.52. – P.26-39.
27. Wang S., Manning C.D. Baselines and bigrams: Simple, good sentiment and topic classification // 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference. – 2012. – Vol.2. – P.90-94.
28. Lai S., Xu L., Liu K., Zhao J. Recurrent convolutional neural networks for text classification // In Twenty-Ninth AAAI Conference on Artificial Intelligence. – 2015. – P. 2267-2273.
29. Li S., Hu J., Cui Y., Hu J. DeepPatent: patent classification with convolutional neural networks and word embedding // Scientometrics. – 2018. – Vol.117, Iss.2. – P.721-744.
30. Li Y.H., Jain A.K. Classification of Text Documents // The Computer Journal. – 1998. – Vol.41, Iss.8. – P.537-546.
31. Xia R., Zong C., Li S. Ensemble of feature sets and classification algorithms for sentiment classification // Information Sciences. – 2011. – Vol.181, Iss.6. – P.1138-1152.
32. Šubelj L, van Eck N.J., Waltman L. Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods // PLoS ONE. – 2016. – Vol.11, Iss.4. – P.1-23.
33. Liu X., Yu S., Moreau Y., Janssens F., Moor B.D., Glänzel W. Hybrid Clustering by Integrating Text and Citation Based Graphs in Journal Database Analysis // IEEE International Conference on Data Mining Workshops, Miami. – 2009. – P.521-526.
34. Waltman L., Boyack K.W., Colavizza G., van Eck N.J. A principled methodology for comparing relatedness measures for clustering publications // arXiv:1901.06815. – URL: https://arxiv.org/ftp/arxiv/papers/1901/1901.06815.pdf (дата обращения: 14.10.2019).
35. Boyack K.W., Newman D., Duhon R.J., Klavans R., Patek M., Biberstine J.R., Schijvenaars B., Skupin A., Ma N., Börner K. Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches // PLoS ONE. –2011. – Vol.6, Iss.6. – P.1-11.
36. Zhang B., Chen Y., Fan W., Fox E.A., Gonçalves M.A., Cristo M., Calado P. Intelligent GP fusion from multiple sources for text classification // Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 – November 5. – 2005.
37. Tshitoyan V., Dagdelen J., Weston L., Dunn A., Rong Z., Kononova O., Persson K.A., Ceder G., Jain A. Unsupervised word embeddings capture latent knowledge from materials science literature // Nature. – 2019. – Vol.571. – P.95-98.
38. Borrajo L., Romero R., Iglesias E.L., Redondo Marey C.M. Improving imbalanced scientific text classification using sampling strategies and dictionaries // Journal of Integrative Bioinformatics. – 2011. – Vol.8, Iss.3. – P.1-15.
39. Sinclair G., Webber B. Classification from full text: A comparison of canonical sections of scientific papers // In Proc. of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland. – 2004. – P. 66-69.
40. Riloff E. Little words can make a big difference for text classification // Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA. – 1995. – P. 130-136.

Материал поступил в редакцию 15.10.19.

Сведения об авторах

СЕЛИВАНОВА Ирина Вячеславовна – младший научный сотрудник ГПНТБ СО РАН,
г. Новосибирск
e-mail: selivanova@spsl.nsc.ru

КОСЯКОВ Денис Викторович – заместитель директора по развитию ГПНТБ СО РАН
e-mail: kosyakov@spsl.nsc.ru

ГУСЬКОВ Андрей Евгеньевич – кандидат технических наук, директор ГПНТБ СО РАН
e-mail: guskov@spsl.nsc.ru