УДК 81'322
DOI 10.25205/1818-7900-2020-18-3-57-68

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

И. В. Селиванова

ГПНТБ СО РАН
Новосибирск, Россия

Аннотация

Приводятся ограничения применения метода классификации научных текстов, основанного на сжатии данных, ко всем категориям из классификации ASJC, используемой в библиографической базе данных Scopus. Показано, что автоматическое создание обучающих выборок для каждой категории является достаточно трудоемким процессом, а в ряде случаев невозможно из-за ограничения на выгрузку данных, установленного в Scopus, и отсутствия названий категорий в Scopus Search API. Другим фактором является то, что во многих областях наук полностью отсутствуют журналы и, соответственно, публикации, у которых указана только одна категория. Применение метода ко всем 26 областям наук невозможна в виду их обширности, а также изначальной классификации Scopus. Часто в разных областях наук находятся терминологически близкие категории, что затрудняет отнесение публикации к верной области. Проведенная работа также указывает на то, что многие исследования, основанные на использовании проклассифицированных по ASJC публикаций, могут иметь некоторые неточности.

Ключевые слова

классификация научных текстов, сжатие текстов, Scopus, SciVal, ASJC

Благодарности

Автор выражает благодарность канд. техн. наук А. Е. Гуськову и Д. В. Косякову за ценные советы при подготовке и оформлении статьи.

Для цитирования

Селиванова И. В. Ограничения применения метода на основе сжатия данных к классификации аннотаций публикаций, индексируемых в Scopus // Вестник НГУ. Серия: Информационные технологии. 2020. Т. 18, № 3. С. 57–68. DOI 10.25205/1818-7900-2020-18-3-57-68

Limitations of Applying the Data Compression Method to the Classification of Abstracts of Publications Indexed in Scopus

I. V. Selivanova

SPSTL SB RAS
Novosibirsk, Russian Federation

Annotation

The paper describes the limitations of applying the method of classification of scientific texts based on data compression to all categories indicated in the ASJC classification used in the Scopus bibliographic database. It is shown that the automatic generation of learning samples for each category is a rather time-consuming process, and in some cases is impossible due to the restriction on data upload installed in Scopus and the lack of category names in the Scopus Search API. Another reason is that in many subjecжt areas there are completely no journals and, accordingly, publications that have only one category. Application of the method to all 26 subject areas is impossible due to their vastness,

as well as the initial classification of Scopus. Often in different subject areas there are terminologically close categories, which makes it difficult to classify a publication as a true area.

These findings also indicate that the classification currently used in Scopus and SciVal may not be completely reliable. For example, according to SciVal in terms of the number of publications, the category "Theoretical computer science" is in second place among all publications in the subject area "Mathematics". The study showed that this category is one of the smallest categories, both in terms of the presence of journals and publications with only this category. Thus, many studies based on the use of publications in ASJC may have some inaccuracies.

Keywords

text classification, text compression, Scopus, SciVal, ASJC

Acknowledgements

The author is thanks to PhD A. E. Guskov and D. V. Kosyakov for valuable advice in the preparation and design of the article.

For citation

Selivanova I. V. Limitations of Applying the Data Compression Method to the Classification of Abstracts of Publications Indexed in Scopus. *Vestnik NSU. Series: Information Technologies*, 2020, vol. 18, no. 3, p. 57–68. (in Russ.) DOI 10.25205/1818-7900-2020-18-3-57-68

Введение

В 2017 г. Б. Я. Рябко и соавторами был предложен метод классификации научных текстов, основанный на сжатии информации [1]. В основу этого метода легло предположение, что в научных текстах, относящихся к одному направлению, используется много общих терминов. В связи с этим текст из определенной области наук будет сжиматься с ней лучше всего. Метод был опробован на полных англоязычных и русскоязычных научных текстах [2], а также на аннотациях англоязычных публикаций [3].

Несмотря на то, что в большинстве случаев метод показал высокую точность классификации (85–92 %), были и ограничения применения этого метода. Так, на точность работы метода большое влияние оказывали размер обучающей выборки («ядра»), ее состав, изначальное количество ядер, их тематическая близость, архиватор, алгоритм и параметры сжатия.

Ранее при применении подобного подхода Rudi Cilibrasi и Paul M. B. Vitanyi [4] также выявляли некоторые ограничения: например, при определении авторства текстов метод показывал неверные результаты на файлах с разной кодировкой.

Заметим также, что в работах [1–3] метод классификации научных текстов, основанный на сжатии данных, применялся только к определенному числу категорий. Но из-за выводов о том, что точность работы метода зависит как от изначального количества категорий, так и их тематической близости, возникает вопрос, насколько корректно применение метода ко всем тематическим рубрикам, используемым в выбранной системе классификации.

Так, для 30 ядер, применяемых в работе [3], использовалась классификации All Science Journals Classification (ASJC), применяемая в ББД Scopus и состоящая из трех уровней. На первом из них расположены 4 общих научных направлений: биологические науки, физические науки, медицина, социальные и гуманитарные науки. На втором – 27 областей наук, которые, в свою очередь, разделены на более чем 300 узких категорий [1]. Однако сама ASJC имеет ряд недостатков, которые рассматривались, например, в работе [5]. Во-первых, классификация в Scopus происходит на уровне изданий, а не для каждой отдельной публикации. Во-вторых, 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.

[1] Scopus. Руководство по охвату контента. URL: http://elsevierscience.ru/files/Scopus_Content_Guide_Rus_2017.pdf

Также вызывает вопросы и присутствие в Scopus категорий, названия которых содержат в себе «general» или «(all)» и (miscellaneous). Более того, иногда журналам присваиваются сразу две эти категории [6].

Множество существующих в настоящий момент методов классификации текстов базируются на терминологической близости. Текст представляется в виде вектора в евклидовом пространстве, где оси координат – это термы, n-граммы [7] или лексемы, выделяемые из текста, а координатой по оси является статистическая информация о них [8]. Таким образом, текст может быть представлен в виде частотных векторов встречаемости слов [9; 10] на основе схем tf, tf*idf, tf*CHI и других [11]. При классификации различными методами (например, kNN [12], основанными на теории графов [13] или SVM [14; 15]) между векторами рассчитывается мера близости, при этом ее выбор оказывает значительное влияние на качество классификации [16].

Таким образом, для 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.

В работе [3] было предложено включать в состав ядра самые высокоцитируемые публикации (предположительно, в них используется более характерная для области наук лексика, которая наследуется и остальными публикациями), имеющие только одну категорию. Однако из-за того, что классификация публикаций в Scopus происходит на уровне журнала, и многие журналы являются политематическими, для 333 категорий формирование ядер из публикаций, имеющих только одну категорию, может стать не только трудоемкой задачей, но и полностью невыполнимой. Другой недостаток работы [3] заключается в формировании ядер через аналитический инструмент компании Elsevier SciVal, который основан на данных ББД Scopus. SciVal допускает выгрузку только первых 20 000 результатов в формате .csv или .xls, высылаемых на указанный при регистрации e-mail. Поэтому SciVal может быть применен только для ограниченного числа категорий, а формирование ядер для большего количества категорий требует автоматического подхода. Более того, в выгруженных через SciVal данных отсутствует текст аннотаций, что требует применение дополнительных этапов формирования ядра.

Таким образом, целью работы является анализ возможности применения метода классификации, основанного на сжатии данных, ко второму (для 26 областей наук, исключая область «Multidisciplinary») и третьему (333 категории, включая «general» и «miscellaneous») уровню классификации ASJC.

Работа предполагает в себе решение следующих задач:

1) автоматическое формирование ядер для каждой из 26 областей наук;
2) применение к ним метода классификации, основанного на сжатии данных;
3) анализ возможности применения метода классификации ко всем 333 категориям, указанных в классификации ASJC, путем оценки представленности каждой из этих категорий в области наук, для которой она определена.

Результаты, полученные в работе, могут быть полезны при оценке погрешности, вносимой в различные наукометрические и библиометрические оценки, основывающиеся на классификации ASJC. Так, в работе [17] предметные области, указанные в Scopus, применяются для исследования литературы, находящейся в открытом доступе. В работе [18] ASJC используется для сравнения цитирований, получаемых из систем Web of Science, Scopus и Google Scholar. Рейтинг областей наук, публикуемых совместно с областью «Математика» по количеству публикаций, составляется в исследовании [19]. В работе [20] приведено распределение по категориям Scopus публикаций из Малайзии.

Методология исследования

Метод классификации

Метод классификации, основанный на алгоритмах сжатия, состоит в следующем. Пусть есть n научных областей $X_1, \ldots, X_n$, для каждой из которых определен характерный для нее набор текстов – обучающая выборка или «ядро» этой 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 научной областью того ядра, с которым он имеет наилучшее сжатие.

Для классификации был использован архиватор WinRAR при максимальном значении памяти 128 Мбайт.

Классификация по областям наук

Для каждой из 26 научных областей процесс выгрузки данных происходил в 2 этапа.

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.

Оценка представленности категорий в исследуемой области наук

Как упоминалось во «Введении», классификация публикаций в Scopus происходит на журнальном уровне. Поэтому 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 г. Всего 39 743 журнала.
2. Выделение журналы 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 по 7 917 журналах. Из 9 133 ненайденных журналов статус у 8 875 журнала 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 период 2009–2018 гг. В дальнейшем для краткости будем обозначать этот рейтинг «рейтинг SciVal».

Результаты

Классификация по областям наук

Результаты классификации показали, что 57 % тестовых файлов было определено ошибочно (см. таблицу). Возможно, это связано с тем, что в отличие от узких категорий в научных областях встречается более разнообразная терминология, что затрудняет применение к ним метода.

Результаты классификации по областям наук
Results of classification by subject area

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

[*] Только по данным публикаций с одной категорией.
Only according to publications with one category.

Одним из предполагаемых путей решения этой проблемы является увеличения состава ядер за счет присутствия в них определенного числа публикаций из каждой категории области наук. Но стоит отметить, что в некоторых случаях 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.

Анализ журналов

Для оценки трудоемкости задачи формирования ядер по областям наук проведем исследование представленности всех категорий в каждой из этих областей. В связи с тем, что в классификации ASJC каждой публикации присваиваются те же категории, что были присвоены журналу, а также с тем, что выделение тематики публикации из мультидисциплинарных журналов является нерешенной задачей, будут исследоваться только публикации с одной категорией.

Рассмотрим, насколько полно представлены категории в базе данных Scopus на примере некоторых крупных областей наук (рис. 1–4).

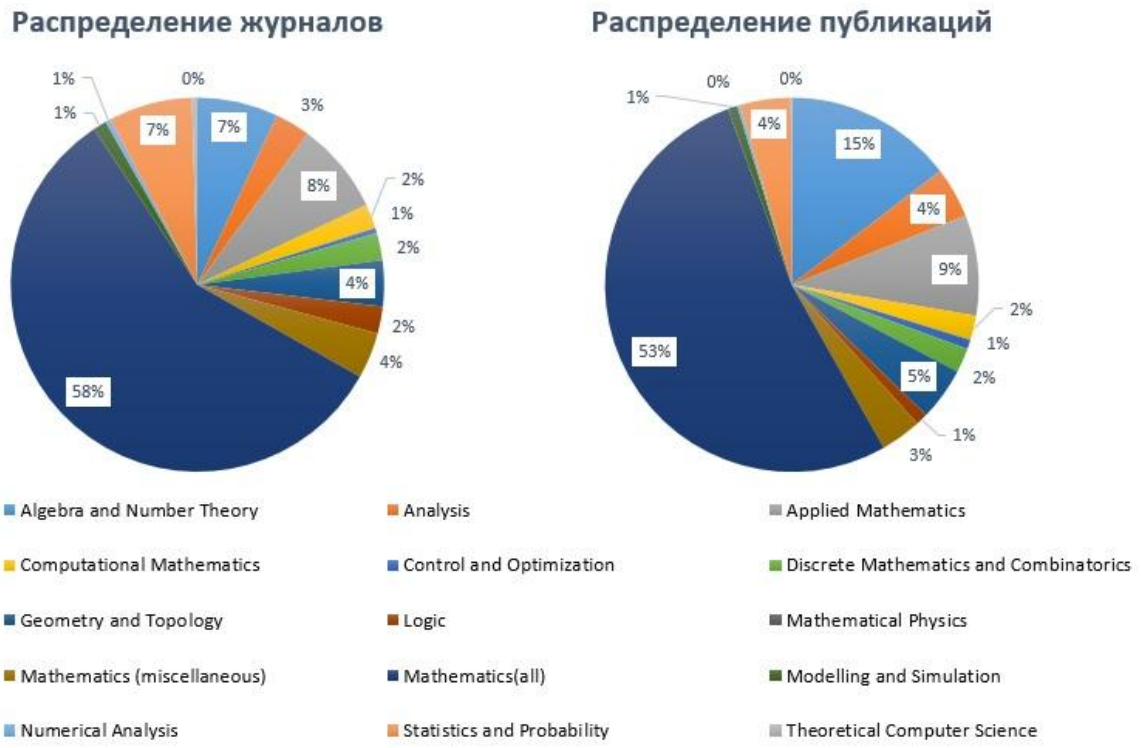

Рис 1. Распределение журналов и публикаций по категориям области наук «Mathematics»
Fig. 1. The distribution of journals and publications by category of subject area “Mathematics”

На рис. 1 показано, что большинство журналов области «Mathematics» относятся к Mathematics (all). Наименьшее число журналов (по 2 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.

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 6 752 месте. Следующие 4 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 2 614, 5 313, 6 796, 8 974, 10 216, 14 938 и 14 995 соответственно. Остальные статьи в первые 20 000 результатов не попали, 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 6 894, 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 (см. рис. 2), категориям «Control and Optimization» и «Theoretical Computer Science» соответствует 6,8 и 14,0 % от общемирового количества публикаций за период 2009–2018 гг., что равносильно 6-му и 2-му 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.

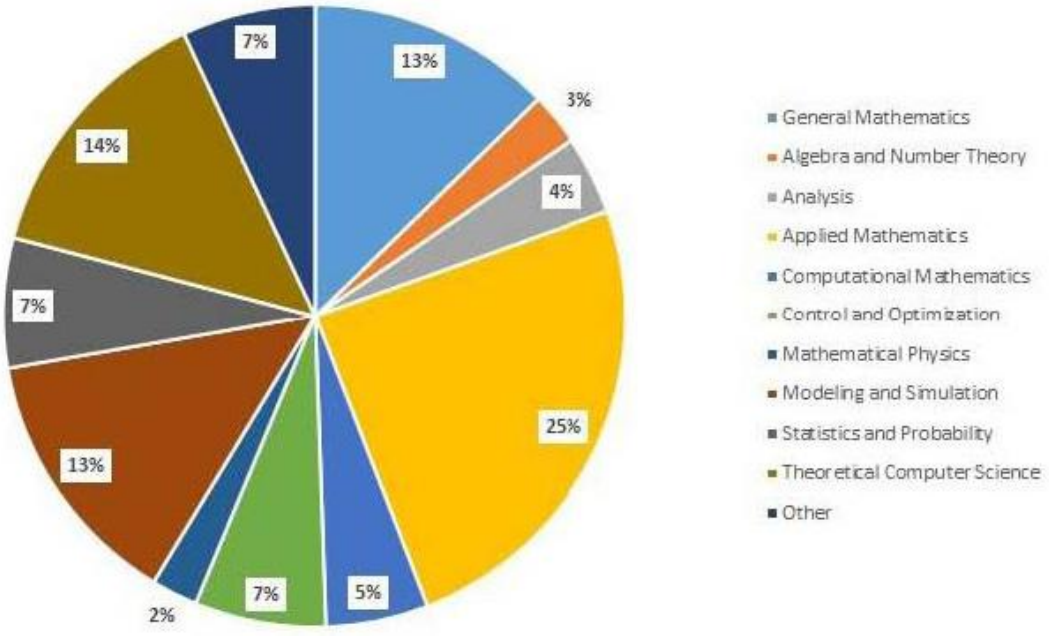

Рис. 2. Доли публикаций по категориям области «Mathematics» по данным SciVal
Fig. 2. Shares of publications by category of subject area “Mathematics” according to SciVal

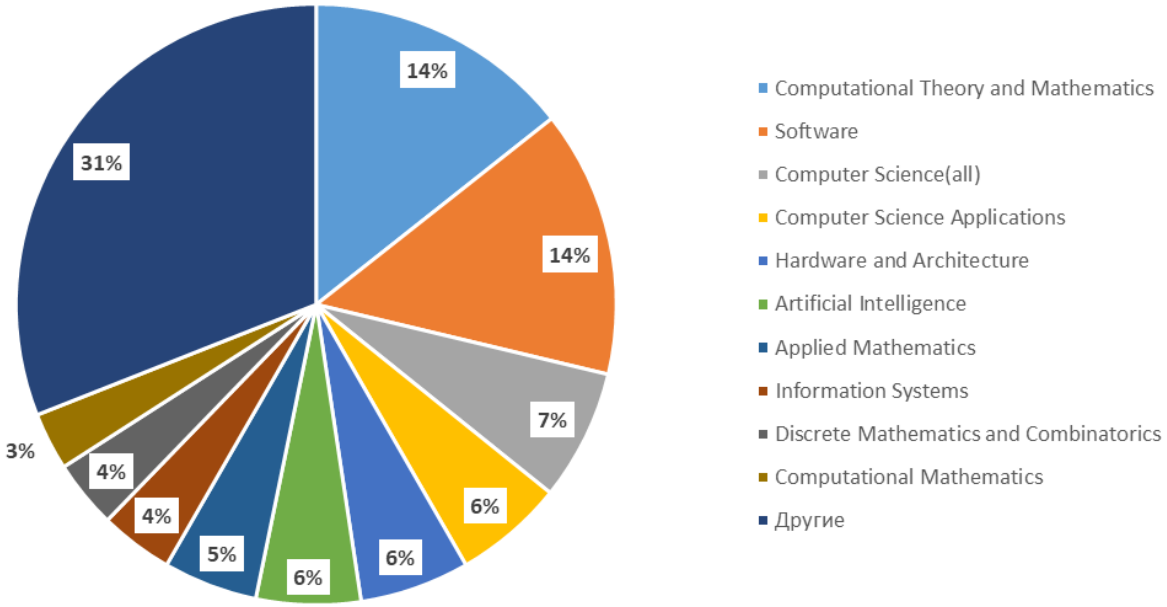

Рис. 3. Категории журналов, указываемые совместно с «Theoretical Computer Science»
Fig. 3. Categories of journals indicated in conjunction with “Theoretical Computer Science”

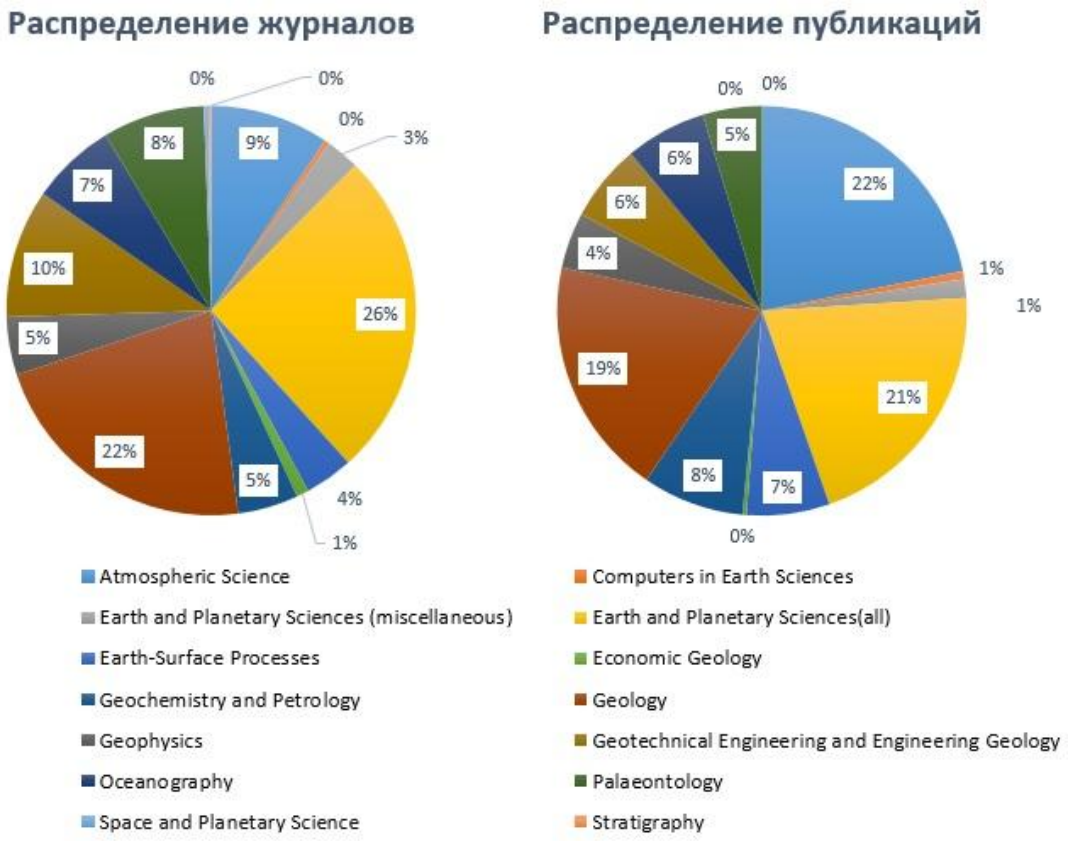

Рис. 4. Распределение журналов и публикаций по категориям области наук «Earth and Planetary Sciences»
Fig. 4. The distribution of journals and publications by category of subject area “Earth and Planetary Sciences”

На рис. 3 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 не только внутри области научных наук, но и целых научных направлений.

Область наук «Medicine» представлена в классификации ASJC 49 различными категориями. Всего моножурналов в этой области 7 301, 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 др.

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

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» (см. рис. 4).

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

Таким образом, формирования ядер по области наук с условием, что в него будет входить определенная доля статей по каждой из категорий, приписанной к этой области, является невыполнимой задачей.

Заключение

Результаты исследования показали, что автоматическое создание ядер для классификации аннотаций публикаций, индексируемых в ББД Scopus, является достаточно трудоемким процессом, а в ряде случаев невозможным вообще. Во-первых, это связано с ограничением на выгрузку данных, установленным в Scopus, и отсутствием названий категорий в Scopus Search API. Во-вторых, во многих областях наук полностью отсутствуют журналы и, соответственно, публикации, у которых указана только одна категория.

Также это указывает на то, что классификация, используемая в настоящий момент в Scopus и SciVal, может быть не до конца достоверной. Например, по данным SciVal по количеству публикаций категория «Theoretical computer science» находится на втором месте среди всех публикаций области «Mathematics». Наши исследования показали, что эта категория является одной из самых малочисленных категорий как с точки зрения присутствия журналов, так и публикаций только с этой категорией. Выделения публикаций из мультидисциплинарных журналов, у которых может быть конкретно эта категория, до настоящего момента проведено не было. Таким образом, многие исследования, основанные на использовании проклассифицированных по ASJC публикаций, могут иметь некоторые неточности.

Классификация публикаций методом, основанным на сжатии данных, по всем 26 областям наук невозможна в виду их обширности, а также изначальной классификации Scopus. Часто в разных областях наук находятся терминологически близкие категории, что затрудняет отнесение публикации к верной области. В основном в ядра попадают публикации лишь из нескольких категорий (зачастую это категория «general» или «all»), что не дает возможности сделать ядро терминологически «узким».

Таким образом, проведенное исследование и работы [1–3] показывают, что применение метода классификации, основанного на сжатии данных, а также других методов, базирующихся на лексической близости, к аннотациям публикаций, индексируемых в базе данных Scopus, возможно только к ограниченному числу терминологически различающихся категорий.

Список литературы

1. **Рябко Б. Я., Гуськов А. Е., Селиванова И. В.** Теоретико-информационный метод классификации текстов // Проблемы передачи информации. 2017. Т. 53, № 3. С. 100–111.
2. **Селиванова И. В., Рябко Б. Я., Гуськов А. Е.** Классификация посредством компрессии: применение методов теории информации для определения тематики научных текстов // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2017. № 6. С. 8–15.
3. **Селиванова И. В., Косяков Д. В., Гуськов А. Е.** Классификация научных текстов на основе компрессии аннотаций публикаций // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2019. № 12. С. 25–38.
4. **Cilibrasi R., Vitanyi P. M. B.** Clustering by Compression. *IEEE Transactions on Information Theory*, 2005, vol. 51, no. 4, p. 1523–1545. DOI 10.1109/tit.2005.844059
5. **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, no. 2, p. 347–364. DOI 10.1016/j.joi.2016.02.003
6. **Frédérique Bordignon.** Tracking content updates in Scopus (2011–2018): a quantitative analysis of journals per subject category and subject categories per journal. In: 17th International Conference on Scientometrics & Informetrics, ISSI. Rome, Italy, 2019, p. 1630.
7. **Miao Y., Keselj V, Milios E.** Document Clustering using Character N-grams: A Comparative Evaluation with Term-based and Word-based Clustering. URL: https://web.cs.dal.ca/eem/cvWeb/pubs/Miao-CIKM-2005.pdf
8. **Волкова Л. Л., Строганов Ю. В.** Об ассоциативных бинарных мерах близости документов: классификация и приложение к кластеризации // Новые информационные технологии в автоматизированных системах. 2014. № 17. С. 421–432.
9. **Baghel R., Dhir R.** A Frequent Concepts Based Document Clustering Algorithm. *International Journal of Computer Applications*, 2010, vol. 4, no. 5, p. 6–12, DOI 10.5120/826-1171
10. **Beil F., Ester M., Xu X.** Frequent Term-Based Text Clustering. In: Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD ‘2002). Edmonton, Alberta, Canada, 2002, p. 436–442. DOI 10.1145/775047.775110
11. **Deng Z.-H., Tang S.-W., Yang D.-Q., Zhang M., Li L.-Y., Xie K. Q.** A comparative study on feature weight in text categorization. In: APWeb, 2004, p. 588–597. DOI 10.1007/978-3-540-24655-8_64
12. **Агеев М. С., Добров Б. В.** Метод эффективного расчета матрицы ближайших соседей для полнотекстовых документов // Вестник Санкт-Петерб. ун-та. Серия 10. 2011. № 3. С. 72–84.
13. **Schaeffer S. E.** Graph clustering. *Computer Science Review*, 2007, vol. 1, no. 1, p. 27–64. DOI 10.1016/j.cosrev.2007.05.001
14. **Rujiang B., Junhua L.** A novel conception based text classification method. In: Proceedings of the IEEE international e-conference on Advanced Science and Technology, 2009, p. 30–34. DOI 10.1109/ast.2009.15
15. **Wang Z., Sun X., Zhang D., Li X.** An optimal SVM based text classification algorithm. In: Proceedings of the 5th IEEE international conference on Machine Learning and Cybernetics, 2006, p. 1378–1381. DOI 10.1109/icmlc.2006.258708
16. **Hu L., Huang M., Ke S. et al.** The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, 2016, vol. 1304. DOI 10.1186/s40064-016-2941-7
17. **Chung J., Tsay M.-Y.** A Bibliometric Analysis of the Literature on Open Access in Scopus. *Qualitative and Quantitative Methods in Libraries*, 2017, vol. 4, no. 4, p. 821–841.
18. **Martín-Martín A., Orduna-Malea E., Thelwall M., López-Cózar E. D.** Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 2018, vol. 12, no. 4, p. 1160–1177. DOI 10.1016/j.joi.2018.09.002

19. **Bathrinarayanan A. L., Vaithiyanathan V., Narayanan S.** Advanced Applied Mathematics Research Output Scientometric Analysis on SCOPUS Database. *International Journal of Pure and Applied Mathematics*, 2017, vol. 117, no. 13, p. 429–437.
20. **Bakri A., Azura N. M., Nadzar M. D., Ibrahim R., Tahira M.** Publication Productivity Pattern of Malaysian Researchers in Scopus from 1995 to 2015. *Journal of Scientometric Research*, 2017, vol. 6, no. 2, p. 86–101. DOI 10.5530/jscires.6.2.14

References

1. **Ryabko B. Y., Guskov A. E., Selivanova I. V.** Information-Theoretic Method for Classification of Texts. *Problems of Information Transmission*, 2017. vol. 53, no. 3, p. 294–304. DOI 10.1134/S0032946017030115
2. **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, no. 3, p. 120–126. DOI 10.3103/s0005105517030116
3. **Selivanova I. V., Kosyakov D. V., Guskov A. E.** Classification of Scientific Texts Based on the Compression of Annotations to Publications Texts. *Automatic Documentation and Mathematical Linguistics*, 2019, vol. 53, no. 6, p. 329–342. DOI 10.3103/S0005105519060062
4. **Cilibrasi R., Vitanyi P. M. B.** Clustering by Compression. *IEEE Transactions on Information Theory*, 2005, vol. 51, no. 4, p. 1523–1545. DOI 10.1109/tit.2005.844059
5. **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, no. 2, p. 347–364. DOI 10.1016/j.joi.2016.02.003
6. **Frédérique Bordignon.** Tracking content updates in Scopus (2011–2018): a quantitative analysis of journals per subject category and subject categories per journal. In: 17th International Conference on Scientometrics & Informetrics, ISSI. Rome, Italy, 2019, p. 1630.
7. **Miao Y., Keselj V, Milios E.** Document Clustering using Character N-grams: A Comparative Evaluation with Term-based and Word-based Clustering. URL: https://web.cs.dal.ca/eem/cvWeb/pubs/Miao-CIKM-2005.pdf
8. **Volkova L. L., Stroganov Yu. V.** Ob assotsiativnykh binarnykh merakh blizosti dokumentov: klassifikatsiya i prilozhenie k klasterizatsii. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh*, 2014, no. 17, p. 421–432. (in Russ.)
9. **Baghel R., Dhir R.** A Frequent Concepts Based Document Clustering Algorithm. *International Journal of Computer Applications*, 2010, vol. 4, no. 5, p. 6–12, DOI 10.5120/826-1171
10. **Beil F., Ester M., Xu X.** Frequent Term-Based Text Clustering. In: Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD ‘2002). Edmonton, Alberta, Canada, 2002, p. 436–442. DOI 10.1145/775047.775110
11. **Deng Z.-H., Tang S.-W., Yang D.-Q., Zhang M., Li L.-Y., Xie K. Q.** A comparative study on feature weight in text categorization. In: APWeb, 2004, p. 588–597. DOI 10.1007/978-3-540-24655-8_64
12. **Ageev M. S., Dobrov B. V.** Metod effektivnogo rascheta matritsy blizhaishikh sosedei dlya pol-notekstovykh dokumentov. *Vestnik Sankt-Peterburgskogo Universiteta. Seria 10*, 2011, no. 3, p. 72–84. (in Russ.)
13. **Schaeffer S. E.** Graph clustering. *Computer Science Review*, 2007, vol. 1, no. 1, p. 27–64. DOI 10.1016/j.cosrev.2007.05.001
14. **Rujiang B., Junhua L.** A novel conception based text classification method. In: Proceedings of the IEEE international e-conference on Advanced Science and Technology, 2009, p. 30–34. DOI 10.1109/ast.2009.15

15. **Wang Z., Sun X., Zhang D., Li X.** An optimal SVM based text classification algorithm. In: Proceedings of the 5th IEEE international conference on Machine Learning and Cybernetics, 2006, p. 1378–1381. DOI 10.1109/icmlc.2006.258708
16. **Hu L., Huang M., Ke S. et al.** The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, 2016, vol. 1304. DOI 10.1186/s40064-016-2941-7
17. **Chung J., Tsay M.-Y.** A Bibliometric Analysis of the Literature on Open Access in Scopus. *Qualitative and Quantitative Methods in Libraries*, 2017, vol. 4, no. 4, p. 821–841.
18. **Martín-Martín A., Orduna-Malea E., Thelwall M., López-Cózar E. D.** Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 2018, vol. 12, no. 4, p. 1160–1177. DOI 10.1016/j.joi.2018.09.002
19. **Bathrinarayanan A. L., Vaithiyanathan V., Narayanan S.** Advanced Applied Mathematics Research Output Scientometric Analysis on SCOPUS Database. *International Journal of Pure and Applied Mathematics*, 2017, vol. 117, no. 13, p. 429–437.
20. **Bakri A., Azura N. M., Nadzar M. D., Ibrahim R., Tahira M.** Publication Productivity Pattern of Malaysian Researchers in Scopus from 1995 to 2015. *Journal of Scientometric Research*, 2017, vol. 6, no. 2, p. 86–101. DOI 10.5530/jscires.6.2.14

Материал поступил в редколлегию
Received
08.04.2020

Сведения об авторе

Селиванова Ирина Вячеславовна, младший научный сотрудник ГПНТБ СО РАН (Новосибирск, Россия)

selivanova@spsl.nsc.ru
ORCID 0000-0001-8805-7631
ResearcherID I-9726-2018

Information about the Author

Irina V. Selivanova, Junior Researcher, SPSTL SB RAS (Novosibirsk, Russian Federation)

selivanova@spsl.nsc.ru
ORCID 0000-0001-8805-7631
ResearcherID I-9726-2018