DOI: 10.15514/ISPRAS–2020–32(6)–10

Иерархическая рубрикация текстовых документов

Д.И. Сорокин, ORCID 0000-0002-6466-3714 <dmitrii.sorokin@phystech.edu>
А.С. Нужный, ORCID 0000-0003-3319-2523 <nuzhny@ibrae.ac.ru>
Е.А. Савельева, ORCID 0000-0002-6562-8750 <esav@ibrae.ac.ru>
Институт проблем безопасного развития атомной энергетики РАН,
115191, Россия, г. Москва, ул. Большая Тульская, д. 52

Аннотация. В работе представлены алгоритм и компьютерная программа иерархической рубрикации текстовой документации. Программа позволяет структурировать неупорядоченный корпус документов в виде иерархии рубрик и визуализировать результат в виде интерактивной карты. Для каждой рубрики автоматически определяются ключевые слова, по которым находятся документы, отнесенные к ней. Анализ построенной иерархии тем позволяет оценить минимальную и максимальную допустимую глубину иерархии, соответствующие минимальному и максимальному количеству различных тем, содержащихся в корпусе документов. Программа апробирована на коллекции документов по захоронению радиоактивных отходов. Результаты тестирования программы показывают хорошее качество построенной иерархии рубрик. Программа может быть использована для ознакомления с коллекцией документов и для тематического поиска.

Ключевые слова: рубрикация; иерархическая кластеризация; обработка естественного языка; машинное обучение

Для цитирования: Сорокин Д.И., Нужный А.С., Савельева Е.А. Иерархическая рубрикация текстовых документов. Труды ИСП РАН, том 32, вып. 6, 2020 г., стр. 127-136. DOI: 10.15514/ISPRAS–2020–32(6)–10

Hierarchical Rubrication of Text Documents

D.I. Sorokin, ORCID 0000-0002-6466-3714 <dmitrii.sorokin@phystech.edu>
A.S. Nuzhny, ORCID 0000-0003-3319-2523<nuzhny@ibrae.ac.ru>
E.A. Saveleva, ORCID 0000-0002-6562-8750 <esav@ibrae.ac.ru>
Nuclear safety institute of the Russian Academy of Sciences,
52, Bolshaya Tulskaya st., Moscow, 115191, Russia

Abstract. Topic modeling is an important and widely used method in the analysis of a large collection of documents. It allows us to digest the content of documents by examination of the selected topics. It has drawbacks such as a need to specify the number of topics. The topics can become too local or too global, depending on that number. Also, it does not provide a relation between local and global topics. Here we present an algorithm and a computer program for the hierarchical rubrication of text documents. The program solves these problems by creating a hierarchy of automatically selected topics in which local topics are connected of the global topics. The program processes PDF documents split them into text segments, builds vector representations using word2vec model and stores them in a database. The vector embeddings are structured in the form of a hierarchy of automatically constructed categories. Each category is identified by automatically selected keywords. The result is visualized in an interactive map. Traversing the hierarchy of topics is done by zooming the map. An analysis of the constructed hierarchy of categories allows us to evaluate the minimum and maximum depth of the hierarchy corresponding to a minimum and a maximum number of different topics contained in the collection of documents. The program was tested on documents on deep nuclear waste disposal.

The results show good quality of the constructed hierarchy of topics and the program can be used for familiarization with the collection of documents and for thematic search.

Keywords: rubrication; hierarchical clustering; natural language processing; machine learning

For citation: Sorokin D.I., Nuzhny A.S., Saveleva E.A. Hierarchical rubrication of text documents. Trudy ISP RAN/Proc. ISP RAS, vol. 32, issue 6, 2020, pp. 127-136 (in Russian). DOI: 10.15514/ISPRAS–2020–32(6)–10

1. Введение

Рубрикация (отнесение документа к одной или нескольким категориям / рубрикам) является одним из наиболее распространенных методов систематизации не упорядоченной коллекции документов. Рубрикация позволяет определять набор тем содержащихся в документах и осуществлять быстрый поиск интересующей информации. В случае если темы рубрики заданы заранее, задачу рубрикации можно рассматривать как задачу классификации. Если же темы рубрик заранее не заданы, то рубрикацию можно рассматривать как задачу кластеризации. Отнесение текста к той или иной рубрике может быть выполнено на основании экспертной оценки, на основе правил или с использованием методов машинного обучения. Экспертная оценка большого корпуса текстов чаще всего является трудоемкой задачей. Рубрикация на основе правил требует подбора правил для каждой предметной области, так как одинаковые термины в разных предметных областях могут иметь различное значение. По этим причинам, в настоящее время большой интерес представляет разработка методов автоматической рубрикации с привлечением машинного обучения.

В данной работе рассматривается задача рубрикации в применении к документации по темам, связанным с захоронением радиоактивных отходов (РАО). При обосновании безопасности захоронения РАО учитывается множество факторов, связанных с геологией, гидрогеологией, геомеханникой, теплофизикой и радиохимией. Многие рассматриваемые факторы близки друг к другу по смыслу и в зависимости от цели информационного поиска можно выделять как более общие темы, связанные, например, с теплопереносом и радионуклидами, гидрогеологией и радиохимией так и более частные темы. Необходимость рассматривать большое количество факторов при обосновании безопасности пункта захоронения РАО приводит к значительному объему разнообразной документации, состоящей из книг, научных статей, отчетов и служебных документов. Для решения данной задачи требуется разработка методов и создание автоматизированных средств рубрикации и поиска документов.

Для рубрикации на основе машинного обучения каждый документ представляется в виде вектора в n-мерном пространстве. Широкое распространение получили подходы к векторизации текстов основанные на использовании нейронных сетей (автоэнкодеров), строящих отображение из исходного пространства в пространство той-же размерности через промежуточное низко размерное представление. Одни из лучших результатов достигнуты с помощью автоэнкодеров таких как word2vec [1], FastText [2], GloVe [3] и doc2vec [4]. Эти методы позволяют автоматически получать семантически нагруженные векторные представления слов или фрагментов документов, а полученные векторы затем могут быть использованы для контекстного поиска [5] или кластеризации документов. Также в настоящее время большого успеха достигли языковые модели, построенные с помощью нейронных сетей на архитектуре трансформеров. Языковая модель BERT [6] показала один из лучших результатов в задачах направленных на понимание текстов таких как ответы на вопросы. Векторное представление, получаемое в результате обучения такой модели, также может быть использовано для кластеризации текстов. Однако данная задача не требует построения языковой модели и сравнимое качество кластеризации может быть достигнуто значительно более простым и менее ресурсоемким подходом с использованием автоэнкодеров [7].

На практике часто возникает необходимость рубрикации набора (корпуса) документов, когда рубрики не заданы или заданы не точно. В этом случае использование алгоритмов кластеризации позволяет распределять документы в кластеры так что признаки входящих в один кластер документов близки по заданной мере. Дополнительной сложностью при кластеризации является выбор числа кластеров для корпуса текстов с заранее неизвестным количеством различных тем. Если число кластеров окажется слишком велико, то каждый кластер будет содержать только очень узко специализированные тексты. А при слишком маленьком количестве кластеров в отдельный кластер могут попадать настолько разнородные тексты, что для них сложно выделить общую тему. Между этими значениями числа кластеров может существовать несколько уровней кластеризации, которые достаточно хорошо описывают как глобальные, так и более локальные темы, встречающиеся в корпусе текстов. Для построения иерархии (дерева) вложенных кластеров могут использоваться методы иерархической кластеризации [8]. Эти методы итеративно строят иерархию кластеров, позволяющую отслеживать в зависимости от уровня более или менее глобальные темы. Анализ получившейся иерархии выполняется с помощью ключевых слов определяемых для каждого кластера. Ключевые слова соответствуют темам документов, вошедших в кластер и позволяют судить о качестве кластеризации.

В работе описывается программное средство, позволяющее построить взаимосвязанную иерархию рубрик по коллекции документов и визуализировать эту иерархию в виде карты. Подход основан на векторном представлении документов и методах иерархической кластеризации. В качестве расстояния между векторами документов используется L2 норма. Тестирование программы производилось на основе коллекции из 200 документов по теме глубинного захоронения радиоактивных отходов [9]. Каждый документ разбивался на смысловые фрагменты, что привело к корпусу из 150 тысяч фрагментов текстов. Для построенной иерархии тем выполнена оценка точности в зависимости от уровня иерархии.

2. Обзор методов иерархической кластеризации текстов

Методы иерархической кластеризации направлены на создание дерева вложенных кластеров. Для построения иерархической кластеризации выделяется два класса подходов агломеративные и дивизивные. При агломеративном подходе каждый документ изначально представляет собой отдельный кластер. Затем на каждом шаге выбираются два наиболее близких по заданной метрике кластера и объединяются в один. Новому кластеру ставится в соответствие вектор центра масс, входящих в него документов. При дивизивном подходе документы изначально представляют собой один кластер. На каждом шаге иерархической кластеризации один из кластеров разбивается на два. В отличии от методов с заранее заданным количеством кластеров таких как, например, k-means [10] иерархическая кластеризация не требует заранее определенных параметров и тем самым лучше подходит для реальных данных.

Применение иерархической кластеризации к текстовым документам рассматривалось в работах [11][13]. В работе [11] был предложен способ объединить агломеративную иерархическую кластеризацию с не иерархической кластеризацией. Было показано, что использование кластеризации алгоритмом k-means не уменьшает точность последующей агломеративной кластеризации, но позволяет ее значительно ускорить посредством уменьшения количества кластеризуемых данных. В работе [12] предложен метод агломеративной кластеризации в котором объединение в кластеры производится до тех пор, пока выполняется критерий «похожести». Данный подход выглядит похожим на подход применяемый в алгоритме dbscan [14]. В нем также вместо количества кластеров задается максимальное расстояние между двумя элементами, формирующими один кластер.

Важной задачей при анализе качества иерархической кластеризации является выделение ключевых слов для полученных кластеров. В работе [15] приведен подробный анализ

методов автоматического извлечения терминов. В зависимости от используемых предположений авторы разделяют их на следующие группы: методы, основанные на частотах слов; методы на основе контекстов вхождений; методы на основе тематических моделей; методы на основе внешних корпусов; методы на основе поисковых машин; методы на основе онтологий; методы на основе Википедии; методы на основе признаков. Для анализа иерархии кластеров нами был выбран метод называемый «частотность терминов – обратная частотность документов» (TF-IDF, Term frequency – inverse document frequency). TF-IDF метод вычисляет метрику как произведение нормированной частоты слова в документах данного кластера на инверсию частоты слова во всех документах

$$\mathrm{tf(t,D)} \cdot \ idf(t,D) = \frac{n_t}{\sum_k n_k} \cdot \log(\frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}),$$

где n_i- число вхождений слова i в документы кластера, $|D|$ число документов в корпусе, $|\{d_i \in D \mid t \in d_i\}|$, число документов в корпусе, в которых встречается слово t.

В работах [16][13] предложены специальные методы для извлечения ключевых слов при иерархической кластеризации документов. Предложенные методы могут улучшить качество выделения ключевых слов, но так как ключевые слова в большинстве случаев не известны, то для оценки их работы требуется дополнительная экспертная оценка.

3. Иерархическая кластеризация с использованием карты Кохонена

В данной работе также, как и в [5] используется модель векторизации текста word2vec. Этот подход позволяет получить семантически нагруженные векторные представления (эмбеддинги) для слов из корпуса текстов. Подход word2vec заключается в обучении нейронной сети предсказывать центральное слово исходя из контекста. В результате обучения в скрытом слое формируется сжатое представление контекста, в котором данное слово употребляется в корпусе. Это скрытое представление затем используется в качестве эмбеддинга.

Для визуализации, а также для уменьшения размерности и сокращения количества векторов перед иерархической кластеризацией используется карта Кохонена [17]. Карта представляет собой двумерную гексагональную сетку. Отличие карты Кохонена от других методов снижения размерности таких как t-sne [18] заключается в том, что отображение, построенное при обучении карты, может быть использовано для преобразования новых данных без необходимости проводить обучение повторно. Это позволит добавлять в уже построенную иерархию новые документы.

Схема обработки документов представлена на рис. 1. Она состоит из двух основных блоков: предобработка документов и построение иерархии рубрик. Предобработка документов включает в себя считывание текста из исходного корпуса документов в формате PDF (tika.apache.org), разбиение текста документов на смысловые единицы – абзацы. Разбиение на абзацы необходимо так как большинство документов содержит в себе более одной темы. Текст каждого абзаца сохраняется в базу данных в формате sqlite (https://www.sqlite.org). Далее из текста удаляются стоп-слова – слова не несущие смысловой нагрузки такие как предлоги, союзы, междометия и производится нормализация слов – приведение слов к первому лицу, единственному числу, мужскому роду и нормальной форме глагола. Нормализация и удаление стоп-слов позволяет значительно уменьшить размер словаря и в конечном итоге увеличить качество рубрикации. На обработанных описанным выше способом фрагментах текстов обучается модель word2vec. После обучения модели для каждого абзаца вычисляется вектор центра масс, входящих в него векторов слов, полученных из word2vec и также сохраняется в базу данных. Данной подход соответствует приближению «мешка слов» в котором порядок слов во фрагменте текста не имеет значения, имеет значение лишь их наличие.

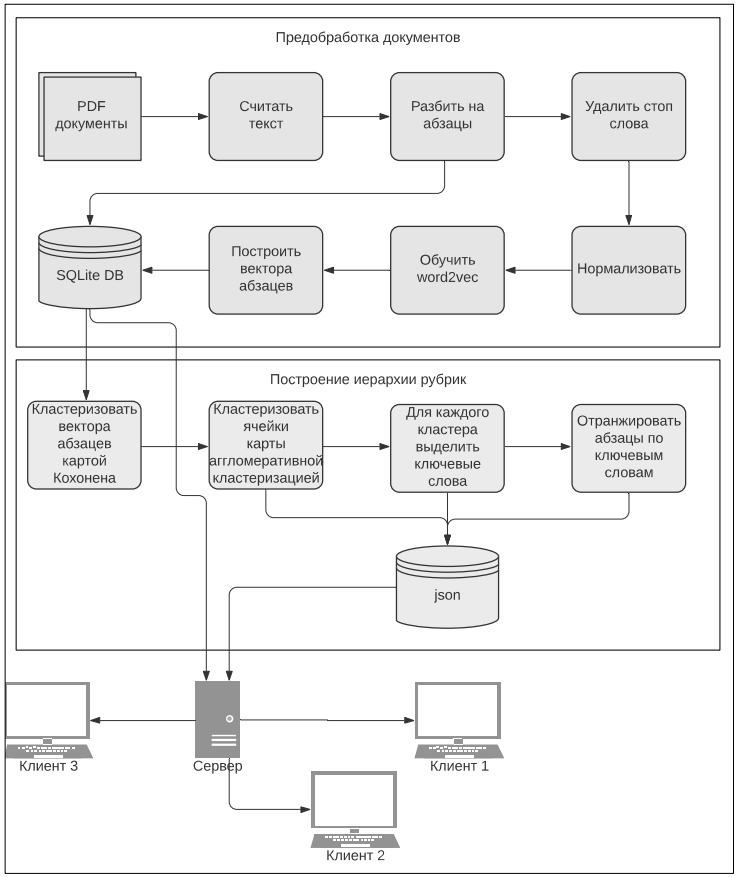

Рис. 1. Схема обработки документов
Fig. 1. Document preprocessing diagram

Далее осуществляется построение иерархии рубрик. На основе векторов абзацев из базы данных строится карта Кохонена. Ячейки карты Кохонена затем кластеризуются с помощью аггломеративной кластеризации (в реализации scikit-learn). Для каждого кластера выделяется набор ключевых слов на основе метрики $M = \frac{n_t^2}{\sum_k n_k}$. Для последующего поиска наиболее релевантных к теме кластера документов, фрагменты текстов, отнесенные к каждому кластеру, ранжируются на основании ключевых слов кластера с помощью алгоритма okapi-bm25 [19]. Результат кластеризации, ключевые слова кластеров и ранжированные идентификаторы документов сохраняются в файл в формате json (https://www.json.org).

База данных и json-файл затем используется в клиент-серверном приложении.

4. Анализ результатов

В случае построения иерархии рубрик нужно определять минимальную и максимальную глубину иерархии такую, чтобы выделенные рубрики не были слишком частными или же слишком общими. Для решения данной проблемы оценим качество выделения ключевых слов в зависимости от количества рубрик. Зависимость качества выделения ключевых слов

от шага иерархической кластеризации представлена на рис. 2. По оси абсцисс отложен шаг иерархической кластеризации: на нулевом шаге каждый документ представляет собой отдельный кластер, на последнем шаге все документы объединены в один кластер. По оси абсцисс справа отложено расстояние между двумя кластерами, объединенными в один на текущем шаге кластеризации. Видно, что это расстояние в начале растет слабо так как, объединяются близкие кластеры, а в конце растет значительно быстрее так как начинают объединяться сильно удаленные друг от друга кластеры.

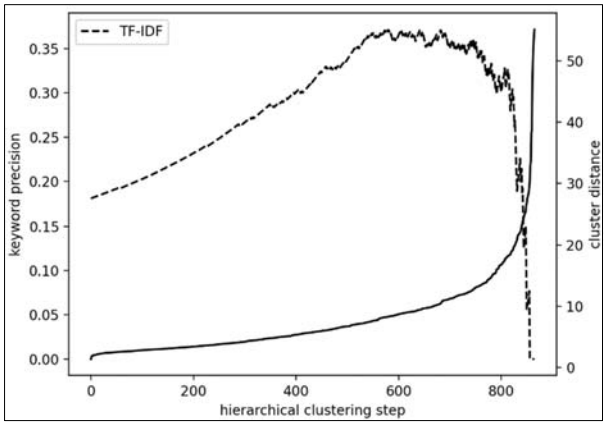

Рис. 2. Оценка качества иерархической кластеризации в зависимости от шага кластеризации
Fig. 2. Hierarchical clustering quality as a function of hierarchical clustering step

На оси ординат слева отмечено качество выделения ключевых слов – число правильно выделенных ключевых слов деленное на общее число выделенных ключевых слов в кластере. Для оценки качества необходимо знать «правильные» ключевые слова. В качества «правильных» ключевых слов рассматривались ключевые слова, выбранные экспертами [20]. Каждому фрагменту текста ставилось в соответствие подмножество (не более 10) ключевых слов, отсортированных по частотам появления во фрагменте текста.

Из рис. 2 видно, что при слишком большом размере кластера на последних шагах иерархической кластеризации сильно растет расстояние между кластерами и одновременно с этим сильно падает точность выделения ключевых слов. Также при слишком маленьком размере кластера получается достаточно низкая точность выделения ключевых слов. Между этими двумя областями находится зона, в которой точность выделения ключевых слов достигает максимума и некоторое время остается на этом уровне. Эта область соответствует иерархии тем, которые лучшим образом описывают данную коллекцию документов. Ограничение глубины иерархической кластеризации размерами близкими к этой области позволяет значительно улучшить качество получаемой иерархии.

5. Интерфейс

Предложенный алгоритм реализован в виде клиент-серверного приложения, написанного на flask (https://flask.palletsprojects.com). Клиентское окно программы состоит из двух частей (Рис. 3). Слева отображается гексагональная карта Кохонена в которой различные кластеры обозначены различным цветом. Приближение карты позволяет переходить между уровнями иерархической кластеризации. При наведении курсора на кластер, кластер подсвечивается и над курсором отображаются ключевые слова. При нажатии на кластер справа отображаются документы отнесенные к этому кластеру. Программа позволяет ознакомиться с иерархией тем содержащихся в документах и найти документы, отнесенные к выбранной рубрике.

Рис. 3. Окно программы иерархической рубрикации
Fig. 3. Hierarchical rubrication program window

Рассмотрим пример использования программы. Допустим пользователь ознакомился с рубриками верхнего уровня и его интересует рубрика «РАО, работа, система, безопасность, должный, год, программа, эксплуатация, оценка, требование». При приближении карты он видит, что эта рубрика в свою очередь состоит из двух рубрик:

- (1a) «РАО, отход, здание, контейнер, радиоактивный отход, оборудование, помещение, система, контроль, захоронение»;
- (1b) «Программа, безопасность, оценка, работа, год, должный, модель, эксплуатация, требование, решение».

При последующих приближениях карты каждая из рубрик в свою очередь перейдет в составляющие подрубрики. Рубрика (1a) состоит из рубрик:

- (2aa) «Здание, помещение, система, контейнер, контроль, устройство, фильтр, машина, средство»;
- (2ab) «РАО, отход, радиоактивный отход, обращение, захоронение, хранение, переработка, упаковка, реактор, ТРО».

Рубрика (1b) состоит в свою очередь из рубрик:

- (2ba) «Программа, безопасность, требование, организация, управление, должный, МАГАТЭ, атомный энергия, документ» и
- (2bb) «Модель, оценка, облучение, доза, метод, условие, параметр, работа».

Видно, что рубрики верхнего уровня являются более общими. По мере увеличения глубины тема становится более конкретной и в тоже время прослеживается связь с родительской рубрикой. Когда пользователь определится с интересующей рубрикой и кликнет по ней справа откроется список фрагментов текстов, ранжированный, по ключевым словам, выбранной рубрики.

Код приложения выложен на github (https://github.com/dmitrySorokin/cluster_search). Там можно увидеть пример взаимодействия пользователя с интерфейсом программы. Ключевым отличием разработанной программы является использование карты Кохонена, которая позволяет не только визуализировать кластеры документов, но также отражает соотношения между ними: близкие на карте кластеры состоят из близких по темам документов; размер кластера отражает количество документов, отнесённых к нему. Это является ключевым отличием предложенного подхода по сравнению с программой Carrot2 (https://search.carrot2.org/#/search/web/hierarchical%20clustering/pie-chart), решающей похожую задачу с помощью круговых диаграмм. Размер кластеров позволяет оценивать

количество документов, отнесенных к заданной теме и находить недостаточно представленные в коллекции документов темы.

5. Заключение

В данной работе представлен автоматический способ рубрикации и визуализации не структурированного корпуса текстов. Предложенный алгоритм применен к коллекции документов по теме глубинного захоронения радиоактивных отходов. Полученный результат позволяет визуализировать эту коллекцию документов в виде карты и производить поиск по автоматически выделенным темам. Анализ полученных результатов показал хорошее соответствие автоматически построенной иерархии рубрик, рубрикам, заданным экспертными ключевыми словами. Также была выделена область иерархической кластеризации, которая лучшим образом соответствует иерархии тем в коллекции документов. Данная программа может быть использована для первичного ознакомления с содержанием коллекции документов и поиска по рубрикам. В дальнейшем планируется перейти от описания темы с помощью ключевых слов к методам суммаризации текстов.

Список литературы / References

[1]. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. In Proc. of the International Conference on Learning Representations, Workshop Track, 2013, 12 p.

[2]. Bojanowskij P., Grave E., Joulin A., and Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, vol. 5, 2017, pp. 135-146.

[3]. Pennington J., Socher R., Manning C. GloVe: Global Vectors for Word Representation. In Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532-1543.

[4]. Le Q., Mikolov T. Distributed representations of sentences and documents In Proc. of the 31st International Conference on Machine Learning, 2014, pp. 1188-1196.

[5]. Нужный А.С., Сорокин Д.И. Создание программы интеллектуального анализа текстовой документации по вопросам захоронения РАО. Труды МФТИ, том 12, № 1(45), 2020 г., стр. 104-111 / Nuzhny A.S., Sorokin D.I. Development of a text-mining program for analysis of documentation on the disposal of radioactive wasteproblem. Proceedings of MIPT, vol. 12, № 1(45), 2020, pp. 104-111 (in Russian).

[6]. Devlin J., Chang M.-W., Lee K., Toutanova K. Bert: Pre-training of Deep Bidirectional Transformers for Language understanding. In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2019, pp. 4171-4186.

[7]. Sia S., Dalmia A., Mielke S.J. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1728–1736.

[8]. Mullner D. Modern hierarchical, agglomerative clustering algorithms. arXiv:1109.2378v1, 2011, 29 p.

[9]. Свительман В.С., Савельева Е.А., Бутов Р.А., Линге Ин.И., Дорофеев А.Н., Тихоновский В.Л. Информационно-аналитическая платформа программы исследований по обоснованию долговременной безопасности российского ПГЗРО. Радиоактивные отходы, № 2 (3), 2018 г., стр. 79-87 / Svitelman V.S., Dorofeev A.N., Saveleva E.A., Butov R.A., Linge I.I., Tikhonovsky V.L. Informational and Software Environment of the Russian Deep Geological Repository Research Program. Radioactive Waste, № 2 (3), 2018, pp. 79–87 (In Russian).

[10]. Jin X., Han J. K-Means Clustering. In Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining, Springer, 2011.

[11]. Bouguettaya A., Yu Q., Liu X., Zhou X., Song A. Efficient agglomerative hierarchical clustering. Expert Systems with Applications, vol. 42, issue 5, 2015, pp. 2785-2797.

[12]. Peng T., Liu L. A novel incremental conceptual hierarchical text clustering method using CFu-tree. Applied Soft Computing, vol. 27, 2015, pp. 268-278.

[13]. Nagarajan R., Nair S.A.H., Puviarasan N., Aruna P. Document clustering using agglomerative hierarchical clustering approach (AHDC) and proposed TSG keywords extraction method. IJRET: International Journal of Research in Engineering and Technology, vol. 05, issue 18, 2016, pp. 118-124.

[14]. Ester M., Kriegel H., Sander J., and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. of the 2nd ACM International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226-231.
[15]. Астраханцев Н.А., Федоренко Д.Г., Турдаков Д.Ю. Методы автоматического извлечения терминов из коллекции текстов предметной области. Программирование, том 41, № 6, 2015 г., стр. 33-52 / Astrakhantsev N.A., Fedorenko D.G., Turdakov D.Yu. Methods for automatic term recognition in domain-specific text collections: A survey. Programming and Computer Software, vol. 41, № 6, 2015, pp. 336-349.
[16]. Peganova I., Rebrova A., and Nedumov Y. Labelling Hierarchical Clusters of Scientific Articles. In Proc. of the 2019 Ivannikov Memorial Workshop (IVMEM), 2019, pp. 26-32.
[17]. Kohonen T. Self–Organizing Maps. Springer, 1997, 426 p.
[18]. van der Maaten L., Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research, vol. 9, 2008, pp. 2579-2605.
[19]. Robertson S.E., Walker S., Beaulieu M. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In Proc. of the Seventh Text Retrieval Conference, 1998, pp. 253-264.
[20]. Рукавичникова А.А., Валетов Д.К., Бутов Р.А., Свительман В.С. Средства тематической кластеризации документов для систематизации библиографической информации по вопросам ПГЗРО. Сборник трудов XIX научной школы молодых ученых ИБРАЭ РАН, 2018 г., стр. 145-148 / Rukavichnikova A.A., Valetov D.K., Butov R.A., Svitelman V.S. Tools for thematic clustering of documents for systematization of bibliographic information on the issues of PGWDF. In Proc. of the XIX Scientific School of Young Scientists IBRAE RAS, 2018, pp. 145-148 (in Russian).

Информация об авторах / Information about authors

Дмитрий Игоревич СОРОКИН – инженер. Научные интересы: обработка естественного языка, глубокое обучение, обучение с подкреплением.

Dmitry Igorevich SOROKIN – engineer. Research interests: natural language processing, deep learning, reinforcement learning.

Антон Сергеевич НУЖНЫЙ – кандидат физико-математических наук, старший научный сотрудник. Научные интересы: теория машинного обучения, некорректные задачи, обработка естественного языка, распознавание образов.

Anton Sergeevich NUZHNY – Ph.D. in Physical and Mathematical Sciences, senior researcher. Research interests: theory of machine learning, ill-posed problems, natural language processing, pattern recognition.

Елена Александровна САВЕЛЬЕВА – кандидат физико-математических наук, заведующая лабораторией геостатистического моделирования. Научные интересы: статистические методы анализа данных, чувствительность модели к ее параметрам, неопределенность при моделировании.

Elena Alexandrovna SAVELEVA – Ph.D. in Physical and Mathematical Sciences, head of geostatistical laboratory. Research interests: statistical methods of data analysis, sensitivity analysis, uncertainty in modeling.