Научная статья
УДК 811.112
DOI: 10.20323/2499-9679-2024-3-38-139
EDN: PUBMIE

Автоматическое извлечение ключевых терминов из корпуса научных статей в SCP

Татьяна Сергеевна Падерина
Младший научный сотрудник, Иркутский научный центр Сибирского отделения Российской академии наук. 664033, г. Иркутск, ул. Лермонтова, д. 134
jana-pad@mail.ru, https://orcid.org/0000-0002-2603-6242

Аннотация. Настоящая статья посвящена изложению теоретических и прикладных принципов работы по автоматическому извлечению терминов из научных текстов. Работа выполняется в рамках государственного задания по теме «Лингвосемиотическая гетерогенность научной картины мира: теоретическое и лингводидактическое описание». Цель исследования заключается в извлечении терминов из подготовленного корпуса научных текстов. Основной задачей на данном этапе исследования было выявить конкорданс определенной терминологии, то есть обозначить список всех употреблений заданного языкового выражения при помощи приложений для автоматической обработки текстов (АОТ). Практическим материалом являются научные статьи по направлению «Науки о Земле». Извлечение терминов при помощи автоматических систем является перспективным направлением современной прикладной лингвистики, так как существенно упрощает и ускоряет процесс создания терминосистем для узкоспециализированных предметных областей и для междисциплинарных направлений, которые находятся на стыке нескольких наук и требует определенного терминологического аппарата. Оценка рабочего процесса извлечения, проведенная с использованием большого набора данных, показала хорошую производительность для большинства типов данных. В этой статье мы описываем общую архитектуру рабочего процесса и предоставляем подробную информацию о реализации отдельных этапов. В результате проделанной работы отмечаем, что полностью перейти на автоматическую обработку текстов на данный момент весьма проблематично, так как полученные результаты не всегда являются точными и могут содержать ошибки. Перспектива исследования связана с адаптацией существующих моделей под определенные научное направления, создание цифровой языковой модели определенных терминосистем и её обучение.

Ключевые слова: терминология; извлечение терминов; автоматическая обработка текстов; термины-кандидаты; языковая модель; научная коммуникация

Для цитирования: Падерина Т. С. Автоматическое извлечение ключевых терминов из корпуса научных статей в SCP // Верхневолжский филологический вестник. 2024. № 3 (38). С. 139–144. http://dx.doi.org/10.20323/2499-9679-2024-3-38-139. https://elibrary.ru/PUBMIE

Original article

Automatic extraction of key terms from the scientific articles corpus in SCP

Tatiana S. Paderina
Junior researcher, Irkutsk scientific center, Siberian branch of Russian academy of sciences. 664033, Irkutsk, Lermontov st., 134
jana-pad@mail.ru, https://orcid.org/0000-0002-2603-6242

Abstract. This article focuses on presenting theoretical and applied principles of automatic term extraction from scientific texts. The work is carried out within the framework of the state assignment on «Linguosemiotic heterogeneity of the scientific worldview: theoretical and linguodidactic description». The study aims at extracting terms from a prepared corpus of scientific texts. The main task at this stage of the study is to identify the concordance of certain terminology, i.e. to list all uses of a given linguistic expression by means of automatic text processing (ATP) applications. The practical material is scientific articles on Earth Sciences. Term extraction by means of automatic systems is a promising direction in modern applied linguistics, as it significantly simplifies and accelerates the process of creating term systems for highly specialized subject areas as well as for interdisciplinary areas at the intersection of several sciences, and requires a certain terminological apparatus. Assessment of the extraction workflow using a large

dataset shows good performance for most data types. In this paper, the author outlines the overall architecture of the workflow and provides detailed information on implementing particular steps. As a result of the work done, we note that it is very problematic to completely switch to automatic text processing at the moment, as the results obtained are not always accurate and may contain errors. The research prospect is related to adapting existing models for certain scientific areas, creating a digital linguistic model of certain term systems and training it.

Key words: terminology; term extraction; automatic text processing; term candidates; language model; scientific communication

For citation: Paderina T. S. Automatic extraction of key terms from the scientific articles corpus in SCP. *Verhnevolzhski philological bulletin. 2024;(3):139–144. (In Russ.).* http://dx.doi.org/10.20323/2499-9679-2024-3-38-139. https://elibrary.ru/PUBMIE

Введение

Академический дискурс является важным каналом коммуникации в научном мире. Отслеживание последних научных открытий и достижений, результаты которых обычно публикуются в журналах или материалах конференций, является важнейшим аспектом исследовательской работы. Игнорирование этой задачи может привести к дефициту знаний, связанных с последними открытиями и научными тенденциями, что, в свою очередь, ведет к снижению качества собственного исследования, значительно усложняет оценку результатов и ограничивает возможность поиска новых интересных областей исследований, методов и задач. К сожалению, изучение научной литературы и, в частности, ознакомление с последними научными открытиями является сложным и чрезвычайно трудоемким процессом. Основная причина – огромный и постоянно растущий объем научной информации, а также тот факт, что публикации в основном доступны в виде неструктурированного текста.

Извлечение данных из статей и других документов является хорошо изученной проблемой. Более старые подходы предполагали, что на входе будут отсканированные документы, и были подготовлены для выполнения полной оцифровки из растровых изображений. В настоящее время приходится иметь дело с растущим количеством электронных документов, которые не требуют индивидуального распознавания символов.

Актуальность разработки и усовершенствования методов сбора, хранения и обработки информации с использованием автоматических программ подтверждается также стремительным развитием науки и применением цифровых технологий для работы с большими объемами данных. В рамках нашей профессиональной деятельности мы исследуем интерфейсы, которые могут значительно облегчить работу исследователям с научными текстами. Эти интерфейсы помогут найти статьи, соответствующие требуемому профилю, быстро просмотреть их для получения полного понимания содержания, осуществить поиск ключевых слов и интегрировать полученные знания в свои собственные исследовательские работы.

Теоретической базой нашего исследования послужили работы отечественных и зарубежных исследователей: изучение извлечения терминов на основе статистических показателей отмечаем в работах V. Stoykova, R. Stankovic, S. Pal и др. [Stoykova, Stankovic,2019; Das, Pal, et.al.; 2013]; на основе правил – W. Sha, T. Quirchmayer и др. [Sha, Hua;2021; Quirchmayr, Paech,et al.,2018]; метод машинного обучения освещен в работах таких исследователей как Е. П. Бручес, Т. В. Батура, M. Conrado, Ю. И. Бутенко, В. Э. Рогачева, K. Zhou и др. [Бручес, Батура, 2021;Бутенко, 2022; Рогачева, 2017; Conrado, 2013; Zhou, 2022; Schapire, 2000].

В качестве материала исследования для извлечения терминов был использован корпус статей за 2017–2022 гг. по направлению подготовки «Науки о Земле» из журналов *Landslides, Natural Hazards and Earth System Sciences, Journal of Geophysical research: Earth Surface, Geophysical Research Letters* и другие.

Принцип отбора тестов обусловлен задачами Государственного задания, которая выполняет лаборатория и более детально был описан ранее в нашей статье [Падерина, 2023]. Тексты, отобранные для корпуса написаны с соблюдение норм академического дискурса, опубликованы в журналах с высоким импакт-фактором (Scopus (Q1), Web of Science (WoS), написаны на английском языке. Такие требования к текстам обеспечивают отбор контента, наиболее соответствующего требованиям письменной научной коммуникации.

Результаты исследования

Раньше любая цифровая модель обучалась для решения конкретной задачи на основе предшествующих данных. Например, чтобы научить систему предсказывать погоду, нужно было собрать большую выборку прецедентов. Языковая модель строится аналогичным образом. Она способна «подсказать» слово в определенном контексте. В

последние годы языковые модели генерируют тексты вполне осмысленно.

Для автоматического сбора, анализа и извлечения данных из различных источников может применяться различный программный инструментарий и платформы. Среди наиболее популярных систем отметим следующие:

– GATE (General Architecture for Text Engineering) – одна из самых старых, открытая инструментальная система, предназначенная для обработки естественного языка (NLP – natural language processing) и информационного поиска с открытым исходным кодом. Она предоставляет набор инструментов, ресурсов и API (Application Programming Interface) для создания и оценки различных текстовых и лингвистических приложений. GATE позволяет работать с большими текстовыми корпусами, анализировать и извлекать информацию из них, а также создавать новые модели и алгоритмы для NLP. Система поддерживает различные языки и кодировки, а также включает в себя множество предобученных моделей и алгоритмов.

– NLTK (Natural Language Toolkit) – это открытый исходный код, который предоставляет инструменты для работы с текстом на естественном языке. Он включает в себя модули для токенизации, стемминга, лемматизации, синтаксического анализа, классификации текста и других задач.

– spaCy – это гибкая платформа для обработки текста на естественном языке с открытым исходным кодом. Она поддерживает несколько языков и имеет обширную библиотеку инструментов для различных задач, включая извлечение сущностей, анализ тональности, морфологический анализ и многое другое.

– Simple Concordance Program (SCP) – многофункциональная программа, которая позволяет извлекать термины, словосочетания, задавать число слов словосочетаниях и т. д. К преимуществам данной программы относится возможность подключить stop-list (артикли, предлоги, общие слова, названия стран и т. д.). Программа извлекает термины в список, которые после этого можно отредактировать в ручном режиме. Программа может обрабатывать тексты на английском, французском, немецком, греческом, русском языках и т. д.

В нашем исследовании мы обратились к последней платформе (Simple Concordance Program (SCP)), так как она находится в свободном доступе и оптимально соответствует нашим задачам. Процесс обучения языковой модели долгосрочный и требует тщательной поэтапной подготовки. Вся работа в рамках нашего исследования была разделена на три ключевых этапа: первый этап был посвящен теоретико-методологическому описанию поставленной проблемы и изучение актуальности исследований в данном направлении. Актуальность подтверждается в том числе активным развитием цифровых аспектов прикладной лингвистики и открытием новых исследовательских лабораторий по изучению возможностей и развития искусственного интеллекта. Второй этап – подготовка корпусов текстов по направлениям заявленных в выполняемом нами государственном задании, отбор терминов-кандидатов для языковой модели, формирование лингвистических шаблонов. Третий этап – создание цифровой языковой модели, обучение этой модели и апробация.

В рамках проведенного исследования отбор терминов-кандидатов был рассмотрен на текстах, написанных на английском языке. Интерес обусловлен практическим применением полученной языковой модели в процессе подготовки аспирантов к сдаче кандидатского экзамена по английскому языку и для использования данной модели в их дальнейшей профессиональной деятельности (написание текстов для зарубежных журналов). Под языковой моделью мы понимаем статистическую модель, которая умеет прогнозировать вероятность последовательности слов в заданном тексте или предложении. Главная цель такой языковой модели состоит в том, чтобы понять и зафиксировать вероятностные связи между словами в языке и обучить подсказывать следующее слово (сочетание) на основе предыдущих. Для того, чтобы модель могла генерировать продолжение текста, необходимо ее «обучить» на достаточно большом объеме текстов, написанных на естественном языке.

Для подготовки корпуса научных текстов мы использовали приложение Semantic Scholar (https://www.semanticscholar.org/). По научному направлению «Науки о Земле» для первоначального этапа обучения модели нами было отобрано 17 оригинальных статей и 1 монография. Общий объем выборки составил 35 печатных листов.

Одна из основных сложностей работы с терминосистемами заключается в том, что без учета контекста достаточно трудно определить, является ли данное слово или словосочетание термином. Более того, в различных контекстах одно и то же выражение может быть обозначено как термин, и как обычное слово / словосочетание: например, *классификация, поток, язык, модель, классификатор и т.д.*

Извлечение терминов выполнялось в три этапа:

1 этап. Анализ научного текста и извлечение списка терминов-кандидатов (слова, словосочетания), которые будут соответствовать лексико-синтаксическим шаблонам;

2 этап. Фильтрация извлеченных терминов-кандидатов с помощью определенного списка стоп-слов (stop-word-list);

3 этап. Упорядочивание терминов-кандидатов по релевантности предметной области [Падерина, 2023].

Так как процесс обработки текстов трудоемкий, кропотливый и требующий определённого времени, в данной статье представлен пример работы с статьей *Debris flow behavior during the September 2013 rainstorm event in the Colorado Front Range, USA* [Schaefer, Santi, Duron, 2021]. В ходе работы по извлечению терминов мы выделили ключевые слова, которые обозначили как термины-кандидаты: *avulsion, debris-flaw, hazard, landslides, slope.*

После подготовки корпуса текстов мы использовали приложение Simple Concordance Program (далее SCP) для манипулирования данными. Для того, чтобы структурировать информацию при помощи системной программой SCP нам необходимо было преобразовать отобранный файл в реляционную форму. Программа предполагает работу либо с текстовыми документами с кодировкой ANSI, либо XML разметка, в нашем случае мы выбрали текстовый документ с кодировкой ANSI и загрузили его в систему. Система автоматически определила 1329 слов для поиска во встроенном словаре, поэтому мы обозначили для программы пять ключевых слов из списка терминов-кандидатов, которые был определен заранее (на предварительном этапе) (Рисунок 1).

Рисунок 1. Отбор терминов-кандидатов в приложении SCP

Как мы видим, программный интерфейс позволяет добавить дополнительные поля для систематизации терминов-кандидатов (частота, длина, шаблон). Результат поиска по заданным ключевым словам показал нам частоту употребления: *avulsion – 88* упоминаний, *debris-flaw – 45, hazard – 9, landslides – 17* и термин-кандидат *slope – 47*.

Список найденных примеров вхождения нужного токена в минимальном контексте был представлен в формате KWIC (Key Word in Context – ключевое слово в контексте). В режиме KWIC в каждом примере выделено центральное слово**.** Слова, расположенные левее (L) и правее (R) этого слова, формируют его левый и правый контекст (Рисунок 2).

Рисунок 2. Пример конкорданса для терминов-кандидатов

Изучение слов и их окружения (контекст) представляет интерес для нашего исследования, так как контекст позволяет рассмотреть разницу между терминами, которые на первый взгляд в переводе на русский язык могут быть синонимами

Так например, «Универсальный русско-английский словарь» [Универсальный русско-английский словарь] дает практически одинаковые определения для терминов *debris-flow* и *landslides* и обозначает их как *«обвал, оползень, оползание, сель»*, в том время как в специализированном словаре по геологии мы видим уточнение этих понятий, *debris-flow* авторы словаря рассматривают как *«течение/поток осколков, обломков, обломочного материала»* [Тимофеев, 1988, с. 124, 179], в то время как для термина *landslides* они дают определение *«оползень; обвал»* [Тимофеев, 1988, с. 239].

Изучение терминов в контексте позволить рассмотреть эту разницу и обучить языковую модель предлагать наиболее подходящий вариант при переводе.

Заключение

Проведя работу по извлечению терминов-кандидатов, мы отмечаем, что, несмотря на кажущееся разнообразие существующих программных цифровых продуктов для автоматической работы с текстами (первые прикладные исследования по извлечению информации из специализированных текстов относятся к началу 80-х годов XX века), задача автоматического извлечения терминов из научных текстов остается актуальной [Дементьева и др., 2022; Большакова и др., 2021; Grishman, 2010; Aggarwal, 2012; Kowsari, 2019; Шейко, 2023].

Это обусловлено рядом причин, среди которых можно выделить сложность и многоаспектность задачи работы с терминологией, разнообразие типов документов и языков, необходимость учета контекста и междисциплинарных связей. Решение

данной проблемы требует совместных усилий специалистов из разных областей, включая лингвистику, информатику, математику и т. д. Одним из перспективных направлений в этой области является использование искусственного интеллекта и машинного обучения для анализа и обработки естественного языка. Современные алгоритмы и методы машинного обучения позволяют создавать более точные и эффективные системы автоматической обработки текстов, способные обрабатывать большие объемы данных и извлекать из них ценную информацию. Однако, несмотря на значительные достижения в данной сфере, существует ряд ограничений, которые требуют дальнейшего исследования и разработки. К таким проблемам относятся:

– сложность и многоаспектность задач анализа текстов;

– неполнота и неоднозначность информации в текстах;

– необходимость учета контекста и семантических связей между словами и фразами;

– проблемы с интерпретацией результатов и оценкой качества работы алгоритмов.

В рамках нашего исследования на основании отобранных терминов-кандидатов производится создание и обучение языковой модели, которая позволит не только выделять термины, которые уже содержатся в подготовленном словаре, но и находить новые паттерны (новые слова и фразы), а также генерировать связный текст (предложения) по определённому научному направлению.

Библиографический список

1. Дементьева Я. Ю., Бручес Е. П., Батура Т. В. Извлечение терминов из текстов научных статей // Программные продукты и системы/Software & Systems. 2022. Т. 35. № 4. С. 689–697. DOI: 10.15827/0236-235X.140.689-697

2. Большакова Е. И., Семак В. В. Комбинирование методов для извлечения терминов из научно-технического текста // Интеллектуальные системы. Теория и приложения. 2021. № 25:4. С. 239–242.

3. Бручес Е. П., Батура Т. В. Метод автоматического извлечения терминов из научных статей на основе слабо контролируемого обучения // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19. № 2. С. 5–16. DOI 10.25205/1818-7900-2021-19-2-5-16

4. Бутенко Ю. И., Николаева Н. С., Карцева Е. Ю. Структурные модели англоязычных терминов для автоматической обработки корпусов научно-технических текстов // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2022. Т. 13. № 1. С. 80–95. DOI 10.22363/2313-2299-2022-13-1-80-95.

5. Падерина Т. С. Методы извлечения терминов в научных текстах (на материале статей по направлению науки о земле) // Казанский лингвистический журнал. 2023. Т. 6(3). С. 388–396. DOI /10.26907/2658-3321.2023.6.3.388

6. Рогачева В. Э. Методы извлечения терминологических единиц из корпуса сопоставимых текстов // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2017. № 2. С. 118–122.

7. Тимофеев П. П., Алексеев М. Н., Софиано Т. А. Англо-русский геологический словарь: Ок. 52 000 терминов / под ред. П. П. Тимофеева, М. Н. Алексеева. Москва, 1988б. 541 с.

8. Универсальный русско-английский словарь. URL: https://universal_ru_en.academic.ru/ (дата обращения: 30.04.2024).

9. Шейко А. М. Инструменты прикладной лингвистики в контроле качества перевода // Казанский лингвистический журнал. 2023. Т. 6. № 2. С. 282–293. DOI 10.26907/2658-3321.2023.6.2.282-293.

10. Aggarwal C.C., Zhai C. (2012). A Survey of Text Classification Algorithms. In: Mining Text Data. 2012. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_6

11. Conrado M., Pardo T., Rezende S. O. A machine learning approach to automatic term extraction using a rich feature set // Proc. NAACL. 2013. P. 16–23.

12. Das B., Pal S., Mondal S., Dalui D., Shome S.K. Automatic keyword extraction from any text document using N-gram rigid collocation // IJSCE. 2013. Vol. 3. No. 2. P. 238–242.

13. Grishman R. Information Extraction. In: The Handbook of Computational Linguistics and Natural Language Processing. A. Clark, C. Fox, and S. Lappin (Eds), WileyBlackwell, 2010. P. 515–530.

14. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D. Text Classification Algorithms: A Survey.// Information. 2019. Vol. 10 (4):150. https://doi.org/10.3390/info10040150

15. Quirchmayr T., Paech B., Kohl R. et al. Semi-automatic rule-based domain terminology and software feature-relevant information extraction from natural language user manuals // Empirical Software Engineering. 2018. Vol. 23. No. 6. P. 3630–3683. DOI: 10.1007/s10664-018-9597-6.

16. Schaefer L. N., Santi P. M., Duron T. C. Debris flow behavior during the September 2013 rainstorm event in the Colorado Front Range, USA // Landslides. 2021. Vol. 18(5). P. 1585–1595. doi:10.1007/s10346-020-01590-5

17. Sha W., Hua B., Linqi S. A Pattern and POS auto-learning method for terminology extraction from scientific text // Data and Information Management. 2021. Vol. 5. No. 3. P. 329–335. DOI: 10.2478/dim-2021- 0005.

18. Schapire R.E., Singer Y. BoosTexter: A Boosting-based System for Text Categorization. // Machine Learning. 2000. Vol. 39. Pp. 135–168. https://doi.org/10.1023/A:1007649029923

19. Stoykova V., Stankovic R. Using query expansion for cross-lingual mathematical terminology extraction. In:

AISC. 2019. P. 154–164. DOI: 10.1007/978-3-319-91189-2_16.

20. Zhou K., Li Y., Li Q. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. Proc. LX Annual Meeting of the Association for Computational Linguistics. 2022. Vol. 1. P. 7198–7211.

Reference list

1. Dement'eva Ja. Ju., Bruches E. P., Batura T. V. Izvlechenie terminov iz tekstov nauchnyh statej = Extracting terms from texts of scientific articles // Programmnye produkty i sistemy/Software & Systems. 2022. T. 35. № 4. S. 689–697. DOI: 10.15827/0236-235X.140.689-697

2. Bol'shakova E. I., Semak V. V. Kombinirovanie metodov dlja izvlechenija terminov iz nauchno-tehnicheskogo teksta = Combining methods for extracting terms from scientific technical texts // Intellektual'nye sistemy. Teorija i prilozhenija. 2021. № 25:4. S. 239–242.

3. Bruches E. P., Batura T. V. Metod avtomaticheskogo izvlechenija terminov iz nauchnyh statej na osnove slabo kontroliruemogo obuchenija = A method for automatically extracting terms from scientific articles based on low-control training // Vestnik NGU. Serija: Informacionnye tehnologii. 2021. T. 19. № 2. S. 5–16. DOI 10.25205/1818-7900-2021-19-2-5-16

4. Butenko Ju. I., Nikolaeva N. S., Karceva E. Ju. Strukturnye modeli anglojazychnyh terminov dlja avtomaticheskoj obrabotki korpusov nauchno-tehnicheskih tekstov = Structural models of English terms for automatic processing of scientific and technical text corpuses // Vestnik Rossijskogo universiteta druzhby narodov. Serija: Teorija jazyka. Semiotika. Semantika. 2022. T. 13. № 1. S. 80–95. DOI 10.22363/2313-2299-2022-13-1-80-95.

5. Paderina T. S. Metody izvlechenija terminov v nauchnyh tekstah (na materiale statej po napravleniju nauki o zemle) = Methods of extracting terms from scientific texts (based on articles in the field of earth science) // Kazanskij lingvisticheskij zhurnal. 2023. T. 6(3). S. 388–396. DOI /10.26907/2658-3321.2023.6.3.388

6. Rogacheva, V. Je. Metody izvlechenija terminologicheskih edinic iz korpusa sopostavimyh tekstov = Methods for extracting terminological units from a corpus of comparable texts // Vestnik Voronezhskogo gosudarstvennogo universiteta. Serija: Lingvistika i mezhkul'turnaja kommunikacija. 2017. № 2. S. 118–122.

7. Timofeev P. P., Alekseev M. N., Sofiano T. A. Anglo-russkij geologicheskij slovar': Ok. 52 000 terminov = English-Russian dictionary of geology: about 52 000 terms / pod red. P. P. Timofeeva, M. N. Alekseeva. Moskva, 1988b. 541 s.

8. Universal'nyj russko-anglijskij slovar'. = Universal Russian-English dictionary. URL: https://universal_ru_en.academic.ru/ (data obrashhenija: 30.04.2024).

9. Shejko A. M. Instrumenty prikladnoj lingvistiki v kontrole kachestva perevoda = Applied linguistics tools in translation quality control // Kazanskij lingvisticheskij zhurnal. 2023. T. 6. № 2. S. 282–293. DOI 10.26907/2658-3321.2023.6.2.282-293.

10. Aggarwal C.C., Zhai C. (2012). A Survey of Text Classification Algorithms. In: Mining Text Data. 2012. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_6

11. Conrado M., Pardo T., Rezende S. O. A machine learning approach to automatic term extraction using a rich feature set // Proc. NAACL. 2013. R. 16–23.

12. Das B., Pal S., Mondal S., Dalui D., Shome S.K. Automatic keyword extraction from any text document using N-gram rigid collocation // IJSCE. 2013. Vol. 3. No. 2. P. 238–242.

13. Grishman R. Information Extraction. In: The Handbook of Computational Linguistics and Natural Language Processing. A. Clark, C. Fox, and S. Lappin (Eds), WileyBlackwell, 2010. R. 515–530.

14. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D. Text Classification Algorithms: A Survey.// Information. 2019. Vol. 10 (4):150. https://doi.org/10.3390/info10040150

15. Quirchmayr T., Paech B., Kohl R. et al. Semi-automatic rule-based domain terminology and software feature-relevant information extraction from natural language user manuals // Empirical Software Engineering. 2018. Vol. 23. No. 6. P. 3630–3683. DOI: 10.1007/s10664-018-9597-6.

16. Schaefer L. N., Santi P. M., Duron T. C. Debris flow behavior during the September 2013 rainstorm event in the Colorado Front Range, USA // Landslides. 2021. Vol. 18(5). P. 1585–1595. doi:10.1007/s10346-020-01590-5

17. Sha W., Hua B., Linqi S. A Pattern and POS autolearning method for terminology extraction from scientific text // Data and Information Management. 2021. Vol. 5. No. 3. P. 329–335. DOI: 10.2478/dim-2021- 0005.

18. Schapire R.E., Singer Y. BoosTexter: A Boosting-based System for Text Categorization. // Machine Learning. 2000. Vol. 39. Pp. 135–168. https://doi.org/10.1023/A:1007649029923

19. Stoykova V., Stankovic R. Using query expansion for cross-lingual mathematical terminology extraction. In: AISC. 2019. P. 154–164. DOI: 10.1007/978-3-319-91189-2_16.

20. Zhou K., Li Y., Li Q. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. Proc. LX Annual Meeting of the Association for Computational Linguistics. 2022. Vol. 1. P. 7198–7211.

Статья поступила в редакцию 14.05.2024; одобрена после рецензирования 10.06.2024; принята к публикации 20.06.2024.
The article was submitted on 14.05.2024; approved after reviewing 10.06.2024; accepted for publication on 20.06.2024