УДК 004.912
DOI: 10.25559/SITITO.18.202204.821-828

Оригинальная статья

Применение вероятносто-энтропийного подхода к отбору тематически схожих документов в информационной системе специального назначения

В. А. Попов[1*], Д. В. Крахмалёв[2], М. С. Чипчагов[2]

[1] ФГКВОУ ВО «Военная академия Ракетных войск стратегического назначения имени Петра Великого» Министерства обороны Российской Федерации, г. Балашиха, Российская Федерация
Адрес: 143900, Российская Федерация, Московская область, г. Балашиха, ул. Карбышева, д. 8
[*] spam@vpopov.su

[2] ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», г. Москва, Российская Федерация
Адрес: 125167, Российская Федерация, г. Москва, пр. Ленинградский, д. 49/2

Аннотация

Задача поиска тематически схожих документов, как одна из задач классификации текста, является одной из наиболее важных областей обработки естественного языка. В результате решения данной задачи, данные автоматически сортируются по заранее определенному набору классов. Поиск тематически схожих документов и классификация текста широко применяется в коммерческих приложениях, таких как фильтрация спама, принятие решений, извлечение информации из необработанных данных и многие другие приложения. В информационных системах специального назначения автоматическая классификация текста применяется для обработки информационных сообщений из открытых источников информации, устраняя необходимость использовать более дорогой и трудоемкий механизм ручной классификации данных. В настоящее время лучшие результаты в автоматической классификации текстов показывают методы основанные на нейронных сетях. Однако следует учитывать, что такие результаты получены на тестовых наборах содержащих десятки и сотни тысяч размеченных документов и в условиях постоянства набора классов. В работе предложен способ отбора тематически схожих документов, в основе которого лежит эталонный набор из нескольких десятков документов, относящийся к каждому конкретному классу. Эталонный набор документов представляется в виде ранжированного списка ключевых слов и словосочетаний (списка ключевых терминов). Место термина в данном списке (ранг термина) определяется путем расчета нескольких вероятностно-энтропийных показателей и последующего суммирования. Далее определяется близость к каждому классу, исходя из количества ключевых терминов каждого класса и итогового веса в документе, подлежащем классификации.

Ключевые слова: классификация текстовых документов, информационная система, обработка информации

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Попов В. А., Крахмалёв Д. В., Чипчагов М. С. Применение вероятносто-энтропийного подхода к отбору тематически схожих документов в информационной системе специального назначения // Современные информационные технологии и ИТ-образование. 2022. Т. 18, № 4. С. 821-828. doi: https://doi.org/10.25559/SITITO.18.202204.821-828

Original article

Application of Probability-Entropy Approach to the Selection of Thematically Similar Documents in the Information System Military Administration

V. A. Popov[a*], D. V. Krakhmalev[b], M. S. Chipchagov[b]

[a] The Military Academy of Strategic Rocket Troops after Peter the Great, Balashikha, Russian Federation
Address: 8 Karbysheva St., Balashikha 143900, Moscow Region, Russian Federation
[*] spam@vpopov.su

[b] Financial University under the Government of the Russian Federation, Moscow, Russian Federation
Address: 49/2 Leningradsky Prospekt, Moscow 125167, Russian Federation

Abstract

The task of finding thematically similar documents, as one of the tasks of text classification, is one of the most important areas of natural language processing. As a result of solving this problem, the data is automatically sorted according to a predefined set of classes. The search for thematically similar documents and text classification is widely used in commercial applications such as spam filtering, decision-making, extracting information from raw data, and many other applications. In special-purpose information systems, automatic text classification is used to process information messages from open sources of information, eliminating the need to use a more expensive and time-consuming manual data classification mechanism. Currently, the best results in automatic text classification are shown by methods based on neural networks. However, it should be taken into account that such results were obtained on test sets containing tens and hundreds of thousands of marked-up documents and under conditions of a constant set of classes. The article offers a method of selection of thematically similar documents, which is based on a reference set of several dozens of documents related to each specific class. The reference set of documents is presented as a ranked list of keywords and phrases (list of key terms). The place of a term in this list (the rank of a term) is determined by calculating several probabilistic-entropy indicators and subsequent summation. Next, proximity to each class is determined based on the number of key terms in each class and the final weight in the document to be classified.

Keywords: classification of text documents, information system, information processing

Conflict of interests: The author declares no conflict of interest.

For citation: Popov V.A., Krakhmalev D.V., Chipchagov M.S. Application of Probability-Entropy Approach to the Selection of Thematically Similar Documents in the Information System Military Administration. *Modern Information Technologies and IT-Education.* 2022;18(4):821-828. doi: https://doi.org/10.25559/SITITO.18.202204.821-828

Введение

Классификация текста – это процесс классификации документов по заранее определенным категориям на основе их содержания. Это автоматическое отнесение текстов на естественном языке к предопределенным категориям. Классификация текста является основным требованием систем поиска текста, которые извлекают тексты в ответ на запрос пользователя, и систем понимания текста, которые каким-либо образом преобразуют текст, например, создают резюме, отвечают на вопросы или извлекают данные. Существующие алгоритмы контролируемого обучения для классификации текста требуют достаточного количества документов для точного обучения. В настоящее время известно множество текстовых документов, доступных в электронном виде, количество которых ежедневно увеличивается. Такие документы представляют собой огромное количество информации, которая легко доступна [1], [2]. В поисках ценности в этой огромной коллекции требуется много работы по упорядочиванию документов, но это можно автоматизировать с помощью интеллектуального анализа данных – использования искусственного интеллекта. Точность и понимание таких систем в значительной степени влияют на их полезность [3-6]. Задача интеллектуального анализа данных состоит в том, чтобы автоматически классифицировать документы по предопределенным классам на основе их содержимого. Для автоматической классификации текста было разработано множество алгоритмов. Наиболее распространенные методы, используемые для этой цели, включают наивный байесовский классификатор, анализ ассоциативных правил, генетический алгоритм, дерево решений и т.д. Интеллектуальный анализ правил ассоциации [7] обнаруживает интересную ассоциацию или корреляцию между большим набором элементов данных. Обнаружение этих взаимосвязей между огромными объемами записей транзакций может помочь в процессе принятия многих решений. С другой стороны, наивный байесовский классификатор использует максимальную апостериорную оценку для изучения классификатора. Предполагается, что появление каждого слова в документе условно не зависит от всех других слов в этом документе, учитывая его класс. Хотя наивный Байес хорошо работает во многих исследованиях [8], [9] для точного обучения требуется большое количество обучающих документов. Генетический алгоритм начинается с начальной популяции, которая создается из случайно сгенерированных правил. Каждое правило может быть представлено строкой битов. Как правило, пригодность правила оценивается по точности его классификации на наборе обучающих примеров.

В данной статье представлен разработанный алгоритм классификации текстов. Вместо использования слов для получения набора признаков из предварительно классифицированных текстовых документов используется отношение слов, т.е. правила ассоциации. Затем концепция наивного байесовского классификатора используется для производных признаков, и, наконец, для окончательной классификации была добавлена концепция генетического алгоритма. Система, основанная на предложенном алгоритме, была внедрена и протестирована. Экспериментальные результаты показывают, что предлагаемая система работает как успешный текстовый классификатор.

Цель исследования

Анализ и мониторинг больших, постоянно растущих объёмов информации требует комплексного решения ряда подзадач:
– описание в краткой форме темы интересов пользователя;
– отбор документов, соответствующих теме;
– представление в наглядной форме результата анализа данных документов.

Существует ряд систем, частично позволяющих решить задачу поиска тематически схожих документов. В частности, в поисковых системах, основанных на Apache Lucene (полнотекстовый поиск), поиск подобных текстов производится, используя заранее определенный документ, и реализуется по методу «мешка слов». Несмотря на такие достоинства данного метода, как производительность и универсальность, его использование для анализа «эволюционирующих» тем, в которых со временем состав ключевых слов меняется, вызывает трудности.

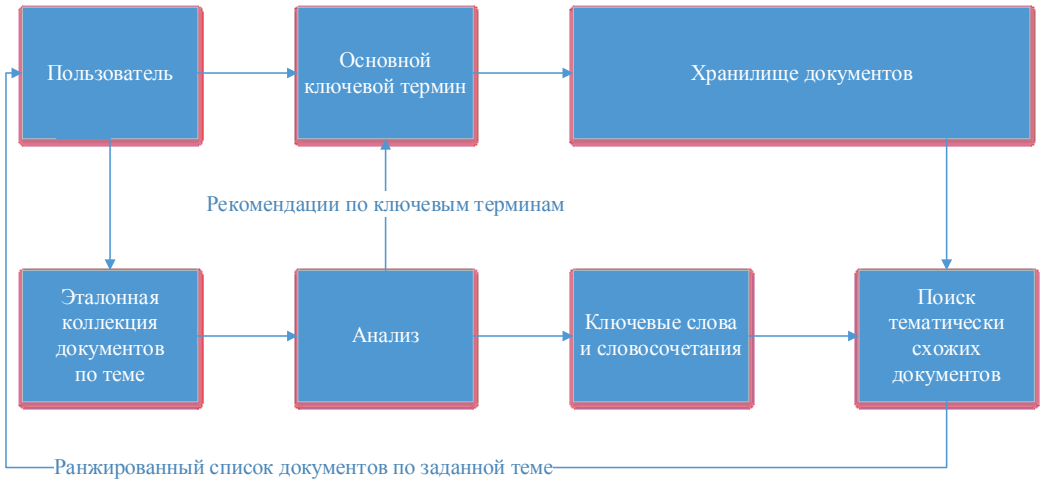
Другой подход заключается в представлении документов в виде вектора заданной размерности и использовании различных метрик близости двух векторов. В рамках этого подхода используются как статистические методы (LDA, PLSA), так и методы, основанные на нейронных сетях (Doc2vec). Недостатками такого подхода являются: требование к наличию большого корпуса для обучения алгоритма, невысокая точность и сложность определения необходимого уровня близости. Особенно остро данные недостатки проявляются при работе с военными текстовыми документами.

Для поиска и подбора тематически схожих документов в информационной системе органа военного управления предлагается использовать алгоритм на основе вероятностно-энтропийных и семантических индикаторов и эталонной коллекции текстов.

Данный алгоритм предложен Vidhya Govindaraju и Krishnan Ramanathan – «Similar document search and recommendation» и усовершенствован И. А. Молошников, А. Г. Сбоев и Д. В. Гудовских в работе «Алгоритм отбора тематически схожих документов с построением контекстно-семантического графа на основе вероятностно-энтропийного подхода» [10].

В дальнейшем предлагается улучшить данный алгоритм за счет использования корпуса военных текстовых документов для выделения общеупотребительных слов и учета последовательности слов при формировании словосочетаний (биграммы, триграммы) из ключевых слов. А также распространить применение данного алгоритма на классификацию (определение принадлежности к определенной группе) электронных текстовых документов в информационной системе органа военного назначения.

Общая схема поиска тематически схожих документов



Р и с. 1. Схема поиска тематически схожих документов
F i g. 1. Search scheme for thematically similar documents

Для анализа темы пользователь предоставляет коллекцию документов, отражающих предмет поиска, которая называется эталонной коллекцией [11-21]. Далее выделяются основные ключевые слова темы, основываясь на них и основном ключевом слове (запрос пользователя), производится отбор и ранжирование документов из хранилища (рис. 1).

Выделение ключевых слов и словосочетаний

При анализе происходит ранжирование слов и словосочетаний, отражающее их принадлежность к теме. Под словосочетанием, биграммой или триграммой, подразумевается комбинация из двух или трёх слов, встречающихся внутри одного предложения без учёта последовательности в предложении.
Входными данными для анализа служит пред обработанные документы из эталонной коллекции. В предобработку входит приведение слов к нормальной форме (лемматизация) (Mystem, UDPipe), разбиение текста на предложение (модуль АОТ), словарная фильтрация наиболее употребительных слов (предлогов, союзов и т. п.). Предобработка текста выполнятся автоматически, сторонним программным обеспечением [20], [21].
Далее определяются вероятностно-энтропийные и семантические индикаторы и рассчитывается ранга термина:
Это такие индикаторы, как:

1. Расстояние Кульбака-Лейблера, используемое для сравнения распределений терминов;
2. Информационная энтропия, отражающая равномерность распределения термина по документам эталонной коллекции;
3. Весы, основанные на распределении Бернулли;
4. Семантический алгоритме Гинзбурга, для определения близости двух слов.

На основе нормализованных значений указанных индикаторов вычисляется единый ранг для каждого термина. Слово «термин» обозначает «слово», в случае, если индикатор рассчитывается для ключевых слов, или «словосочетание», применительно к ключевым словосочетаниям.
Далее отбирается 100 слов с наивысшим рангом из которых формируются словосочетания (биграммы и триграммы) с ними без учёта последовательности слов в предложении. Далее вычисляются ранги для словосочетаний и выбираются биграммы и триграммы с наивысшим рангом.
Используя выделенные термины для каждого документа эталонной коллекции, рассчитывается суммарный вес документа, равный сумме рангов ключевых слов и словосочетаний, входящих в него. Минимальный суммарный ранг документа эталонной коллекции выбирается в качестве базового уровня для фильтрации нерелевантных текстов [22-25].

Расстояние Кульбака-Лейблера

Индикатор, основанный на Дивергенции Кульбака-Лейблера, рассчитывается для слов и словосочетаний согласно формуле

$$D(w) = \Sigma\, Pdoc(w;d) \bullet \ln((Pdoc(w;d)/Pn(d)) \quad (1)$$

Он характеризует различие между реальным распределением термина w с теоретическим, в соответствии с длиной документа (чем документ больше, тем больше в нём различных терминов, а значит, больше вероятность случайного попадания термина w в документ d).
Pn – вероятность встретить термин w во всей коллекции документов относительно длины документа d:
$Pn(d) = N(d) / \Sigma N(x)$,
где N(d) – общее количество терминов в документе d, сумма N(x) – общее количество терминов x во всей коллекции D.

Pdoc(w;d) – вероятность встречаемости термина w в документе d:
Pdoc(w;d) = tf(w;d) / F(w),
где tf(w; d) – встречаемость термина w в документе d, F(w) – встречаемость термина w в коллекции D.
Малое значение величины D(w) показывает, что данный термин характерен для представленной выборки. Это может быть общеупотребительное слово, фильтруемое за счёт других индикаторов, или тематическое ключевое слово. Большое значение говорит о не характерности данного термина для выборки.

Информационная энтропия

Информационная энтропия показывает равномерность распределения Pdoc(w;d) термина w в документах d коллекции D и рассчитывается по формуле:
H(w) =ΣPdoc(w;d) • ln (1 / Pdoc(w;d))
Если данный показатель большой, то термин равномерно представлен в коллекции документов, если он равен 0, то это означает, что термин w встречается только в одном документе. В эталонной тематической коллекции наиболее характеризующие ключевые слова равномерно распределены по набору документов.

Индикатор выделения общеупотребительных слов

Индикатор показывает отличие распределения слова w в эталонной коллекции и в Корпусе военных текстовых документов. Рассчитывается по формуле
r(w) = Pe(w) / Pk(w),
где Pe – относительная частота встречаемости термина в эталонной коллекции, а Pk – относительная частота встречаемости термина в Корпусе военных текстовых документов. Данный индикатор позволяет выделить большую часть общих слов, если они хорошо представлены в корпусе. Для общеупотребительных он будет иметь значение около 1, а для специализированных слов – много больше 1.

Алгоритм формирования ключевых словосочетаний

Алгоритм формирования ключевых словосочетаний состоит из нескольких шагов:

1. Выбрать 1000 наиболее частотных слов в эталонной коллекции в кандидаты на ключевые слова.
2. Рассчитать для кандидатов все описанные ранее индикаторы, за исключением индикатора связанности по Гинзбургу
3. Нормализовать полученные значения индикаторов. Нормализованные значения формируют одно значение для каждого кандидата, называемое ранг, отражающее принадлежность данного слова к теме
4. Слова с наивысшим значением ранга являются ключевыми словами темы
5. На основе выделенных ключевых слов формируются биграммы и триграммы. Эти словосочетания создаются из слов одного предложения без учёта последовательности и их положения в предложении.
6. Для ранжирования биграмм и триграмм рассчитываются ранее описанные индикаторы, включая индикатор связанности по Гинзбургу, и аналогично ключевым словам формируется ранг ключевого словосочетания.
7. Биграммы и триграммы с наивысшим значением ранга являются ключевыми словосочетаниями темы.

Результатом работы данного алгоритма является взвешенный относительно темы, заданной эталонной коллекцией, набор ключевых слов и словосочетаний.

Поиск тематически схожих документов

Используя полученный набор ключевых слов и словосочетаний рассчитывается суммарный вес документа из хранилища информации, равный сумме рангов ключевых слов и словосочетаний, входящих в него. Документы, вес которых превышает минимальный ранг, ранжируются в соответствии с их весом и выдаются пользователю.

Список использованных источников

[1] Loper E., Bird S. NLTK: the Natural Language Toolkit // Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics – Vol. 1 (ETMTNLP '02). USA : Association for Computational Linguistics, 2002. P. 63-70. doi: https://doi.org/10.3115/1118108.1118117

[2] Kamruzzaman S. M. Text classification using artificial intelligence // Journal of Electrical Engineering. 2006. Vol. EE 33, no. I. URL: https://arxiv.org/ftp/arxiv/papers/1009/1009.4964.pdf (дата обращения: 04.10.2022).

[3] Артеменко В. Б., Безденежных И. В., Василецкий Ю. Л. Система военно-научной информации Вооруженных Сил Российской Федерации: вчера, сегодня, завтра // Вооружение и экономика. 2022. № 3(61). С. 143-158. URL: https://elibrary.ru/zgtfxo (дата обращения: 04.10.2022).

[4] Горобец Е. А., Мамонтова А. В. Алгоритм автоматического поиска нестандартных словарных единиц при создании сводного словаря // Актуальные проблемы филологии и педагогической лингвистики. 2022. № 2. С. 131-142. doi: https://doi.org/10.29025/2079-6021-2022-2-131-142

[5] Сорокин А. Б., Кушнарев А. П. Морфологический анализатор текста для выявления полноты информации // Информационные технологии. 2018. Т. 24, № 11. С. 719-724. doi: https://doi.org/10.17587/it.24.719-724

[6] Жидков Р. Е., Викторов Д. С. Информационная технология верификации специального программного обеспечения автоматизированных систем военного назначения // Программные продукты и системы. 2019. Т. 32, № 2. С. 283-289. doi: https://doi.org/10.15827/0236-235X.126.283-289

[7] Narmadha D., NaveenSundar G., Geetha S. A novel approach to prune mined association rules in large databases // 2011 3rd International Conference on Electronics Computer Technology. Kanyakumari, India : IEEE Computer Society, 2011. P. 409-413. doi: https://doi.org/10.1109/ICECTECH.2011.5942031

[8] Text classification using the concept of association rule of data mining / C. M. Rahman, F. A. Sohel, P. Naushad, S. M. Kamruzzaman // Proceedings of International Conference on Information Technology. Kathmandu, Nepal, 2003. P. 234-241. doi: https://doi.org/10.48550/arXiv.1009.4582

[9] Kamruzzaman S. M., Farhana Haider. A Hybrid Learning Algorithm for Text Classification // Proceedings of the 3rd International Conference on Electrical & Computer Engineering (ICECE 2004). Dhaka Bangladesh, 2004. P. 577-580. doi: https://doi.org/10.48550/arXiv.1009.4574

[10] Молошников И. А., Сбоев А. Г., Гудовских Д. В. Алгоритм отбора тематически схожих документов с построением контекстно-семантического графа на основе вероятностно-энтропийного подхода // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2015. № 3. С. 64-70. URL: https://www.elibrary.ru/item.asp?id=25942829 (дата обращения: 04.10.2022).

[11] Лукашевич Н. В., Четверкин И. И. Построение модели для извлечения оценочной лексики в различных предметных областях // Моделирование и анализ информационных систем. 2013. Т. 20, № 2. С. 70-79. URL: https://elibrary.ru/item.asp?id=19544642 (дата обращения: 04.10.2022).

[12] da Silva N., Chrishman R. The role of frames in the organization of online dictionaries // Calidoscopio. 2018. Vol. 16, issue 3. P. 450-459. doi: https://doi.org/10.4013/cld.2018.163.09

[13] Introducing Baselines for Russian Named Entity Recognition / R. Gareev [и др.] // Computational Linguistics and Intelligent Text Processing. CICLing 2013. Lecture Notes in Computer Science ; ed. by A. Gelbukh. Vol. 7816. Berlin, Heidelberg : Springer, 2013. P. 329-342. doi: https://doi.org/10.1007/978-3-642-37247-6_27

[14] Panchendrarajan R., Amaresan A. Bidirectional LSTM-CRF for Named Entity Recognition // Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. Hong Kong : Association for Computational Linguistics, 2018. P. 531-540. URL: https://aclanthology.org/Y18-1061.pdf (дата обращения: 04.10.2022).

[15] Kuzmenko E. Morphological Analysis for Russian: Integration and Comparison of Taggers // Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science ; ed. by D. I. Ignatov, et al. Vol. 661. Cham : Springer, 2017. P. 162-171. doi: https://doi.org/10.1007/978-3-319-52920-2_16

[16] Cinque M., Corte R. D., Pecchia A. Entropy-Based Security Analytics: Measurements from a Critical Information System // 2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). Denver, CO, USA : IEEE Computer Society, 2017. P. 379-390. doi: https://doi.org/10.1109/DSN.2017.39

[17] Koltcov S. Application of Rényi and Tsallis entropies to topic modeling optimization // Physica A: Statistical Mechanics and its Applications. 2018. Vol. 512. P. 1192-1204. doi: https://doi.org/10.1016/j.physa.2018.08.050

[18] Yin C., Xi J. Maximum entropy model for mobile text classification in cloud computing using improved information gain algorithm // Multimedia Tools and Applications. 2017. Vol. 76, issue 16. P. 16875-16891. doi: https://doi.org/10.1007/s11042-016-3545-5

[19] Artemenko V. B., Bezdenezhnykh I. V. Prospective Areas for Development of the Military-Scientific Information System // Scientific and Technical Information Processing. 2021. Vol. 48, issue 1. P. 58-69. doi: https://doi.org/10.3103/S0147688221010093

[20] Комплекс вероятностно-энтропийных и интеллектуальных алгоритмов для эмотивно-тематического анализа эволюции общественного мнения в интернет сетях / И. А. Молошников, А. Г. Сбоев, Р. Б. Рыбка, Д. В. Гудовских // Вестник Национального исследовательского ядерного университета МИФИ. 2016. Т. 5, № 3. С. 262-267. doi: https://doi.org/10.1134/S2304487X16030081

[21] Двухуровневая модель нейронной сети глубокого обучения для задачи морфологического разбора предложений русского языка / И. А. Молошников, Р. Б. Рыбка, А. Г. Сбоев [и др.] // Вестник Национального исследовательского ядерного университета МИФИ. 2017. Т. 6, № 6. С. 555-562. doi: https://doi.org/10.1134/S2304487X17060086

[22] The System Adaptability Evaluation Index System of Military Communication Equipment System / J. Hou [и др.] // Proceedings of the 5th International Conference on Computer Science and Application Engineering (CSAE '21). New York, NY, USA : Association for Computing Machinery, 2021. Article number: 4. doi: https://doi.org/10.1145/3487075.3487079

[23] Dien T. T., Loc B. H., Thai-Nghe N. Article Classification using Natural Language Processing and Machine Learning // 2019 International Conference on Advanced Computing and Applications (ACOMP). Nha Trang, Vietnam : IEEE Computer Society, 2019. P. 78-84. doi: https://doi.org/10.1109/ACOMP.2019.00019

[24] Ashrafi M. Z., Taniar D., Smith K. A New Approach of Eliminating Redundant Association Rules // Database and Expert Systems Applications. DEXA 2004. Lecture Notes in Computer Science ; ed. by F. Galindo, M. Takizawa, R. Traunmüller. Vol. 3180. Berlin, Heidelberg : Springer, 2004. P. 465-474. doi: https://doi.org/10.1007/978-3-540-30075-5_45

[25] Text Classification Algorithms: A Survey / K. Kowsari [и др.] // Information. 2019. Vol. 10, issue 4. Article number: 150. doi: https://doi.org/10.3390/info10040150

Поступила 04.10.2022; одобрена после рецензирования 29.11.2022; принята к публикации 06.12.2022.

Об авторах:

Попов Владимир Александрович, адъюнкт кафедры, ФГКВОУ ВО «Военная академия Ракетных войск стратегического назначения имени Петра Великого» Министерства обороны Российской Федерации (143900, Российская Федерация, Московская область, г. Балашиха, ул. Карбышева, д. 8), **ORCID: https://orcid.org/0000-0002-6232-2481,** spam@vpopov.su

Крахмалёв Дмитрий Владимирович, доцент кафедры бизнес-информатики, ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации» (125167, Российская Федерация, г. Москва, пр. Ленинградский, д. 49/2), кандидат технических наук, доцент, **ORCID: https://orcid.org/0000-0001-5362-8676,** kdv@vpopov.su

Чипчагов Михаил Сергеевич, доцент Департамента анализа данных и машинного обучения, ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации» (125167, Российская Федерация, г. Москва, пр. Ленинградский, д. 49/2), кандидат технических наук, **ORCID: https://orcid.org/0000-0003-2038-9108,** hcip@vpopov.su

Все авторы прочитали и одобрили окончательный вариант рукописи.

References

[1] Loper E., Bird S. NLTK: the Natural Language Toolkit. In: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics – Vol. 1 (ETMTNLP '02). USA: Association for Computational Linguistics; 2002. p. 63-70. doi: https://doi.org/10.3115/1118108.1118117

[2] Kamruzzaman S.M. Text classification using artificial intelligence. *Journal of Electrical Engineering*. 2006;EE 33(I). Available at: https://arxiv.org/ftp/arxiv/papers/1009/1009.4964.pdf (accessed 04.10.2022).

[3] Artemenko V.B., Bezdenezhnyh I.V., Vasiletskiy Yu.L. Military Scientific Information System of the Armed Forces of the Russian Federation: Yesterday, Today, Tomorrow. *Armament and Economics*. 2022;(3):143-158. Available at: https://elibrary.ru/zgtfxo (accessed 04.10.2022). (In Russ., abstract in Eng.)

[4] Gorobets E.A., Mamontova A.V. Algorithm of automatic search for non-standard vocabulary units when creating a comprehnsive dictionary. *Current Issues in Philology and Pedagogical Linguistics.* 2022;(2):131-142. (In Russ., abstract in Eng.) doi: https://doi.org/10.29025/2079-6021-2022-2-131-142

[5] Sorokin A.B., Kushnarev A.P. Morphological Text Analyzer for Revealing the Completeness of Information. *Informacionnye tehnologii* = Information Technologies. 2018;24(11):719-724. (In Russ., abstract in Eng.) doi: https://doi.org/10.17587/it.24.719-724

[6] Zhidkov R.E., Viktorov D.S., Zhidkov E.N. An information technology for verifying special software of military automated systems. *Software & Systems*. 2019;32(2):283-289. (In Russ., abstract in Eng.) doi: https://doi.org/10.15827/0236-235X.126.283-289

[7] Narmadha D., NaveenSundar G., Geetha S. A novel approach to prune mined association rules in large databases. In: 2011 3rd International Conference on Electronics Computer Technology. Kanyakumari, India: IEEE Computer Society; 2011. p. 409-413. doi: https://doi.org/10.1109/ICECTECH.2011.5942031

[8] Rahman C.M., Sohel F.A., Naushad P., Kamruzzaman S.M. Text classification using the concept of association rule of data mining. In: Proceedings of International Conference on Information Technology. Kathmandu, Nepal; 2003. p. 234-241. doi: https://doi.org/10.48550/arXiv.1009.4582

[9] Kamruzzaman S.M., Farhana Haider. A Hybrid Learning Algorithm for Text Classification. In: Proceedings of the 3rd International Conference on Electrical & Computer Engineering (ICECE 2004). Dhaka Bangladesh; 2004. p. 577-580. doi: https://doi.org/10.48550/arXiv.1009.4574

[10] Moloshnikov I.A., Sboev V.G., Gudovskikh D.V. Probabilistic-entropic algorythm of contextual semantic graph construction for selection of thematically similar texts. *Proceedings of Voronezh State University. Series: Linguistics and Intercultural Communication.* 2015;(3):64-70. Available at: https://www.elibrary.ru/item.asp?id=25942829 (accessed 04.10.2022). (In Russ., abstract in Eng.)

[11] Loukachevitch N.V., Chetviorkin I.I. Construction of a Model for the Cross-domain opinion word extraction. *Modeling and Analysis of Information Systems*. 2013;20(2):70-79. Available at: https://elibrary.ru/item.asp?id=19544642 (accessed 04.10.2022). (In Russ., abstract in Eng.)

[12] da Silva N., Chrishman R. The role of frames in the organization of online dictionaries. *Calidoscopio*. 2018;16(3):450-459. doi: https://doi.org/10.4013/cld.2018.163.09

[13] Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V. Introducing Baselines for Russian Named Entity Recognition. In: Gelbukh A. (ed.) Computational Linguistics and Intelligent Text Processing. CICLing 2013. Lecture Notes in Computer Science. Vol. 7816. Berlin, Heidelberg: Springer; 2013. p. 329-342. doi: https://doi.org/10.1007/978-3-642-37247-6_27

[14] Panchendrarajan R., Amaresan A. Bidirectional LSTM-CRF for Named Entity Recognition. In: Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. Hong Kong: Association for Computational Linguistics; 2018. p. 531-540. Available at: https://aclanthology.org/Y18-1061.pdf (accessed 04.10.2022).

[15] Kuzmenko E. Morphological Analysis for Russian: Integration and Comparison of Taggers. In: Ignatov D.I., et al. (eds.) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. Vol. 661. Cham: Springer; 2017. p. 162-171. doi: https://doi.org/10.1007/978-3-319-52920-2_16

[16] Cinque M., Corte R.D., Pecchia A. Entropy-Based Security Analytics: Measurements from a Critical Information System. In: 2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). Denver, CO, USA: IEEE Computer Society; 2017. p. 379-390. doi: https://doi.org/10.1109/DSN.2017.39

[17] Koltcov S. Application of Rényi and Tsallis entropies to topic modeling optimization. *Physica A: Statistical Mechanics and its Applications.* 2018;512:1192-1204. doi: https://doi.org/10.1016/j.physa.2018.08.050

[18] Yin C., Xi J. Maximum entropy model for mobile text classification in cloud computing using improved information gain algorithm. *Multimedia Tools and Applications*. 2017;76(16):16875-16891. doi: https://doi.org/10.1007/s11042-016-3545-5

[19] Artemenko V.B., Bezdenezhnykh I.V. Prospective Areas for Development of the Military-Scientific Information System. *Scientific and Technical Information Processing.* 2021;48(1):58-69. doi: https://doi.org/10.3103/S0147688221010093

[20] Moloshnikov I.A., Sboev A.G., Rybka R.B., Gudovskikh D.V. Complex of Probabilistic-Entropy and Intelligent Algorithms for Emotiveness-Thematic Analysis of the Evolution of Public Opinion in the Internet Network. *Vestnik natsional'nogo issledovatel'skogo yadernogo universiteta "MIFI".* 2016;5(3):262-267. (In Russ., abstract in Eng.) doi: https://doi.org/10.1134/S2304487X16030081

[21] Moloshnikov I.A., Rybka R.B., Sboev A.G., Gudovskikh D.V., Ivanov I.I. Two-level model of a deep neural network for the morphological analysis of Russian sentences. *Vestnik natsional'nogo issledovatel'skogo yadernogo universiteta "MIFI".* 2017;6(6):555-562. (In Russ., abstract in Eng.) doi: https://doi.org/10.1134/S2304487X17060086

[22] Hou J., Wang R., Wang J., Yang Z., He D. The System Adaptability Evaluation Index System of Military Communication Equipment System. In: Proceedings of the 5th International Conference on Computer Science and Application Engineering (CSAE '21). New York, NY, USA: Association for Computing Machinery; 2021. Article number: 4. doi: https://doi.org/10.1145/3487075.3487079

[23] Dien T.T., Loc B.H., Thai-Nghe N. Article Classification using Natural Language Processing and Machine Learning. In: 2019 International Conference on Advanced Computing and Applications (ACOMP). Nha Trang, Vietnam: IEEE Computer Society; 2019. p. 78-84. doi: https://doi.org/10.1109/ACOMP.2019.00019

[24] Ashrafi M.Z., Taniar D., Smith K. A New Approach of Eliminating Redundant Association Rules. In: Galindo F., Takizawa M., Traunmüller R. (eds.) Database and Expert Systems Applications. DEXA 2004. Lecture Notes in Computer Science. Vol. 3180. Berlin, Heidelberg: Springer; 2004. p. 465-474. doi: https://doi.org/10.1007/978-3-540-30075-5_45

[25] Kowsari K., Meimandi K.J., Heidarysafa M., Mendu S., Barnes L., Brown D. Text Classification Algorithms: A Survey. *Information*. 2019;10(4):150. doi: https://doi.org/10.3390/info10040150

Submitted 04.10.2022; approved after reviewing 29.11.2022; accepted for publication 06.12.2022.

About the authors:

Vladimir A. Popov, Department Adjunct, The Military Academy of Strategic Rocket Troops after Peter the Great (8 Karbysheva St., Balashikha 143900, Moscow Region, Russian Federation), **ORCID: https://orcid.org/0000-0002-6232-2481,** spam@vpopov.su

Dmitry V. Krakhmalev, Associate Professor of the Department of Business Informatics, Financial University under the Government of the Russian Federation (49/2 Leningradsky Prospekt, Moscow 125167, Russian Federation), Cand.Sci. (Eng.), Associate Professor, **ORCID: https://orcid.org/0000-0001-5362-8676,** kdv@vpopov.su

Mikhail S. Chipchagov, Associate Professor of the Department of Data Analysis and Machine Learning, Financial University under the Government of the Russian Federation (49/2 Leningradsky Prospekt, Moscow 125167, Russian Federation), Cand.Sci. (Eng.), **ORCID: https://orcid.org/0000-0003-2038-9108,** hcip@vpopov.su

All authors have read and approved the final manuscript.