ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ КАК ИНСТРУМЕНТ НАУЧНОГО ПОИСКА В IT-СФЕРЕ

***Бегларян М.Е.*[1], *Добровольская Н.Ю.*[2]**

Ключевые слова: *научная информация, научная статья, текст, тематика, тематическое моделирование, сервис, алгоритм, интерфейс, поиск, результат поиска.*

Аннотация

Цель работы: определить пути совершенствования подготовительного этапа создания научной статьи или научного доклада через новые информационные поисковые процессы, основанные на тематическом моделировании.

Методы: комплексный подход, связывающий основные элементы процесса поиска информации в различных IT-областях, в том числе в программировании; моделирование процесса отбора информации по ключевым признакам; методы компьютерной лингвистики, предполагающие выделение уровней текста, обладающих различной степенью развития и формализации: фонетический, морфологический, лексический, синтаксический и семантический; при реализации тематического моделирования применены семантические методы.

Результаты: разработан алгоритм, реализованный в виде сервиса, позволяющий выявить ключевые слова и соответствующие им тематики научных текстов в области программирования; основу алгоритма представляют различные методы построения тематических моделей; такой сервис позволяет систематизировать и классифицировать научные статьи и учебные материалы, как по программированию, так и в других IT-областях.

EDN: GJBKSH

Введение

Ключом к качественным знаниям является эффективный поиск научной информации. Причем если говорить о фундаментальных науках, то существует ряд признанных научных источников, к которым можно обратиться. Но если требуются знания в развивающихся областях, например, в *IT*-сфере, то новую информацию следует черпать из современных научных статей и монографий. Многие научные публикации снабжены списками *ключевых слов*, которые призваны указывать на узкую направленность работы, но не всегда точно отражают ее суть. Иногда ключевых слов к информационному блоку нет вообще, а заголовок статьи далеко не всегда отражает содержание. Решить проблему определения научной направленности текстов помогут алгоритмы *тематического моделирования*[3]. Тематическое моделирование — одно из направлений компьютерного анализа текстов[4].

Классификация научных текстов необходима в непрерывном образовании, в том числе и в дистанционном. Причем алгоритмы выделения тематик научных текстов могут быть полезны как преподавателям, готовящим учебные материалы для дистанционных курсов, так и учащимся, выполняющим поиск необходимой информации самостоятельно для выполнения заданий различной направленности.

Существует несколько алгоритмов тематического моделирования, результаты работы которых могут отличаться; предметной областью для тематического моделирования научных статей в данном исследовании выбрана *IT*-область, также будет проведен анализ результатов, полученных различными алгоритмами.

На *рис.* 1 представлена компонентная модель сервиса выявления тематик научных текстов.

[3] Тематическое моделирование (от англ. topic modelling) — способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов.

[4] Тематическое моделирование с помощью *Gensim*. 2019. URL: https://webdevblog.ru/tematicheskoe-modelirovanie-s-pomoshhju-gensim-python/ (дата обращения: 03.10.2023).

[1] ***Бегларян Маргарита Евгеньевна,*** *кандидат физико-математических наук, доцент, профессор кафедры социально-гуманитарных и естественнонаучных дисциплин Северо-Кавказского филиала Российского государственного университета правосудия, г. Краснодар, Российская Федерация.*

E-mail: rita_beg@mail.u

[2] ***Добровольская Наталья Юрьевна,*** *кандидат педагогических наук, доцент, доцент кафедры информационных технологий Кубанского государственного университета, г. Краснодар, Российская Федерация.*

E-mail: dnu10@mail.ru

Рис. 1. Компонентная модель сервиса

Анализ сервиса выявления тематик

Рассмотрим отдельные компоненты сервиса выявления тематик (*рис.* 1).

1. *Модуль пользовательского интерфейса*. Формирует поисковый запрос и отправляет его на сервер. Здесь же организуется удобный формат просмотра результатов поиска сервиса. Интерфейс выполнен на стеке технологий: *Vue* 3, *JavaScript*, *HTML*5, *CSS*3.

2. *Ядро сервиса*. Ядро выполняет обмен информацией всех модулей, принимает запросы пользовательского интерфейса, передает запрос в модуль поиска статей и собирает найденные статьи. Затем статьи поставляются модулю тематического моделирования для их классификации, формирования списка ключевых слов, построения финальных тематик. Полученные результаты тематического моделирования преобразуются в заданный формат и отправляются пользователю. Ядро сервиса выполнено на стеке технологий: *Python*, *FastAPI*.

3. *Модуль поиска статей*. Принимает поисковый запрос и с помощью парсеров[5] ищет как можно больше подходящих публикаций на самых популярных русскоязычных ресурсах (Хабр, Киберленинка и др.). Затем полученные статьи возвращаются в ядро сервиса для дальнейшей передачи модулю тематического моделирования. Стек используемых технологий: *Python*, *Scrapy*, *Beautiful soap* 4, *Selenium*.

4. *Модуль тематического моделирования*. Принимает полученные от ядра сервиса научные публикации. Статьи проходят несколько стадий предобработки: фильтрацию, удаление стоп-слов (бессмысленные, перегружающие текст слова), стемминг[6]. Затем обработанные статьи передаются предобученным алгоритмам тематического моделирования, которые определяют темы и ключевые слова каждого текста. Стек технологий: *Python*, *Sklearn*, *Pymystem*3. Для выявления тематик текстов в сервисе реализованы алгоритмы *BERT* (*Bidirectional Encoder Representations from Transformers*), алгоритм Дирихле *LDA* (*Latent Dirichlet allocation*), алгоритмы *TextRank* и *Rutermextract* [1—3, 7].

Выделим *основные направления* применения алгоритмов тематического моделирования научных статей, в том числе и применения предложенного сервиса, в дистанционном образовании (*рис.* 2).

Алгоритмы тематического моделирования способны выявлять основные темы и концепции, которые раскрываются в научных статьях [11—13]. Это позволит быстрее ориентироваться в большом объеме информации и понять, о чем идет речь в статьях, будут ли они полезны для собственного исследования.

Тематическое моделирование позволяет выявить связи между различными научными статьями. Полученные тематики научных материалов помогут определить схожие методы и подходы, которые используют различные авторы для решения одной задачи. Это может быть полезно для более глубокого понимания материала и проведения сравнительного анализа.

Алгоритмы выявления тематик текстов позволят принимать решения, которые будем называть *информированными,* т. е. реально соответствующими теме и направлению исследования. Они могут помочь выявить статьи, которые наиболее соответствуют интересам и потребностям аспиранта, ученого, специалиста. Преподаватели могут использовать тематическое мо-

[5] Парсеры (от англ. *parsing* ~ «структурирование; лексический разбор») — программы автоматизированного сбора и систематизации (структурирования) информации согласно заданным параметрам.

[6] Стемминг (от англ. stemming ~ находить происхождение) — эвристический, довольно грубый процесс нахождения основы слова путем отрезания «лишнего» от корня слов с возможной потерей словообразовательных суффиксов [10].

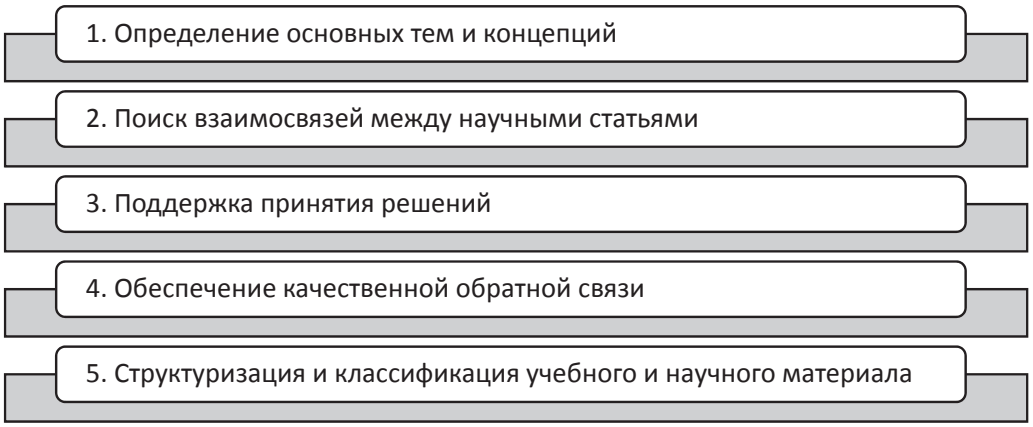

Рис. 2. Направления применения тематического моделирования в дистанционном образовании

делирование для анализа и оценки научных и реферативных работ обучаемых. Это позволит определить, насколько хорошо учащийся освоил определенную тему или концепцию, и предложить дополнительные материалы или рекомендации для дальнейшего изучения.

Алгоритмы тематического моделирования могут помочь обучающимся структурировать и классифицировать большой объем информации из научных статей. Алгоритмы и сервисы выявления тематик помогут выделить ключевые темы и понять, как различные идеи и концепции связаны между собой.

В целом тематическое моделирование научных статей и соответствующие сервисы могут значительно облегчить процесс как обучения, так и научных исследований, помогая обучающимся и научным работникам быстрее ориентироваться в огромных объемах информации и принимать в работу только выверенные информационные массивы.

Сравнительный анализ программных продуктов

Проведём сравнительный анализ существующих продуктов, решающих задачу тематического моделирования. К найденным в общедоступных источниках средствам относятся:

1. *MegaIndex* — онлайн сервис, решающий множество задач по продвижению сайтов; дополнительно позволяет выполнять тематическое моделирование контента (статей, документов, сайтов)[7].

На *рис*. 3 представлен интерфейс инструмента для определения тематики текста.

Сервис может принимать в качестве входных данных как сам текст, так и ссылку на сайт с текстом. Есть также возможность встраивания сервиса в другие информационные системы с помощью *http*-запросов. Представленная на *рис*. 3 статья описывает основы алгоритмов на примере сортировки вставками с использованием языка *C++*. Сервис почти верно определил тематику текста.

Использование сервиса не бесплатно. Длина текста для бесплатного анализа ограничена, программный интерфейс недоступен. Это делает затруднительным интеграцию и использование.

2. *Генератор ключевых слов* издательства «Молодой учёный»[8]. Сервис, позволяющий выделить ключевые слова в научных статьях.

Пользовательский интерфейс генератора представлен на *рис*. 4.

Результаты работы изображены на рис. 5.

Сервис справился с задачей хорошо, но присутствуют и незначимые ключевые слова: *часть*, *данные*.

Сервис не встраиваем, так как не предоставляет программный интерфейс, однако, в отличие от *MegaIndex,* не ограничивает длину текста для анализа.

3. Пакеты прикладных программ для тематического моделирования: *Gensim*, *Orange data mining*, *Sklearn* и др. Пакеты реализованы в основном на языке программирования *Python* [4].

Пакеты предоставляют множество инструментов для тематического моделирования. В них уже реализованы такие алгоритмы, как *LDA*[9], *LSA*[10] *и др. Главным недостатком* этих пакетов является то, что они требуют дальнейшей интеграции в разрабатываемые программы, а следовательно — знания *языков программирования*, в частности, *специального языка R* [6, 8, 9]. В исходном виде применять их для решения задач не представляется возможным.

Для определения ключевых тематик научных статей в нашем исследовании был рассмотрен алгоритм *BERT*. Это алгоритм глубокого обучения, предназначенный для обработки естественного языка. Его основное свойство — определение контекста и связей между

[7] Мега индекс. 2021. URL: https://www.megaindex.ru/ (дата обращения: 12.10.2023).

[8] Молодой учёный. 2020. URL: https://server.moluch.ru/key-words (дата обращения: 17.02.2023).

[9] LDA (англ. Latent Dirichlet Allocation) — метод латентного размещения Дирихле.

[10] LSA (англ. Latent Semantic Analysis) — алгоритм тематического моделирования.

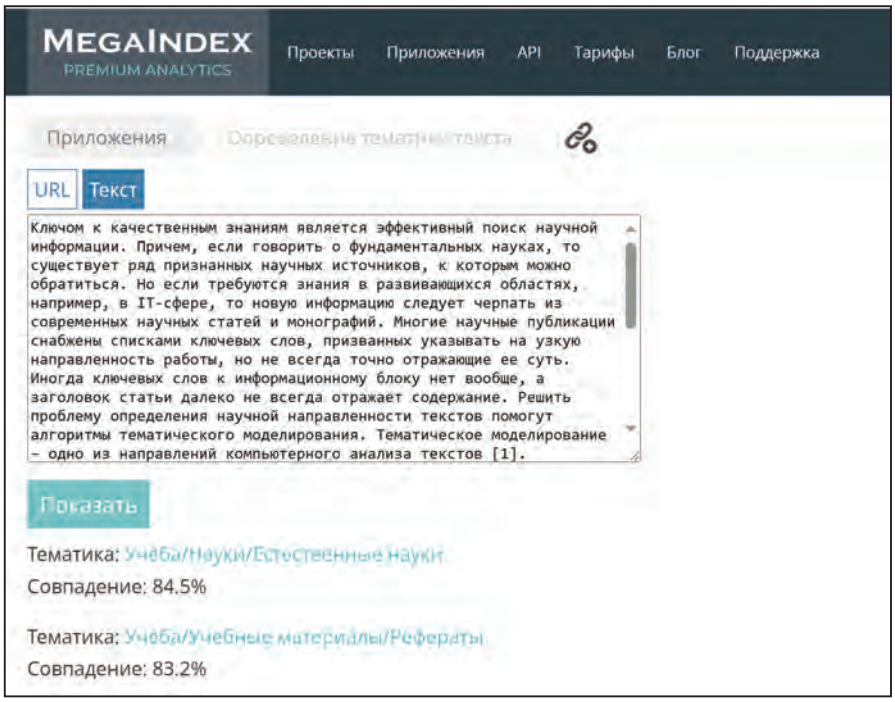

Рис. 3. Сервис MegaIndex

Рис. 4. Генератор ключевых слов издательства «Молодой учёный»

Рис. 5. Результаты работы генератора ключевых слов

словами в предложении, что дает ему преимущество по сравнению с традиционными методами обработки естественного языка. Разработчики *BERT,* помимо англоязычной модели, обучили также мультиязычную модель [5], в которую входит русский язык. В частности, компания «Сбер» предоставила в открытый доступ версию *ruBERT*. Использование *BERT* для задачи тематического моделирования включает следующие этапы: выделение эмбеддингов[11]; кластеризация; выделение тематик.

Под *эмбеддингом* (от англ. *embedding* — встраивание) понимают процесс преобразования слов или фраз в векторы чисел, которые в дальнейшем обрабатываются программой. На первом этапе использовалась мультиязычная модель *BERT*. Получившиеся векторы имеют большую размерность, так как учитывается контекст как до, так и после слова. Стоит обратить внимание, что слишком низкая размерность приводит к потере информации, а слишком высокая ухудшит результаты кластеризации. Для уменьшения размерности использовался пакет *UMAP* (*Uniform Manifold Approximation and Projection*) машинного обучения. Он хорошо справляется с этой задачей, при этом сохраняя значительную часть многомерной локальной структуры в меньшей размерности. После уменьшения размерности проведена кластеризация с помощью *HDBSCAN* (*Hierarchical Density-Based Spatial Clustering of Application with Noise*) — алгоритм кластеризации, основанный на плотности и не очень чувствительный к выбросам.

Для того чтобы извлечь темы, необходимо понять, чем один кластер отличается от другого и какими наиболее подходящими ключевыми словами представлен каждый кластер. Для решения этой задачи использована предложенная в 2020 г. Мартеном Гротендорстом метрика[12] *c-TF-IDF* (*class-based Term Frequency-Inverse Document Frequency*). Метрика *TF-IDF* вычисляется, отталкиваясь не от одного документа, а от всех документов одного кластера, объединив документы кластера в один и посчитав метрики *TF-IDF* для каждого слова. Таким образом, по величине метрики можно оценивать, насколько то или иное ключевое слово представляет тему и значимо в ней.

Метрика *c-TF-IDF* вычисляется по следующей формуле:

$$c - TF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_j^n t_j},$$

где i — номер класса; t_i — частота слова t в классе i; w_i — общее количество слов в классе i; m — общее количество документов.

Для формирования тематики выделено десять наиболее значимых слов. На *рис*. 6 приведен результат обработки научной статьи, содержащей информацию о *Node.js* (платформа для использования языка программирования *JavaScript* на стороне сервера).

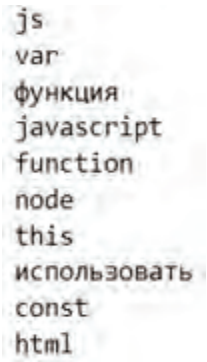

Рис. 6. Результат обработки текста о Node.js с использованием BERT

Модель *BERT* дает хорошие результаты и способна эффективно выделять тематики текста, что позволяет оценить содержание статьи без её прочтения.

Следующим используемым в сервисе алгоритмом выбран алгоритм латентного размещения Дирихле *LDA*. Для обучения алгоритма *LDA* необходимо составить терм-документную матрицу и заполнить её значениями *tf-idf*. Сам алгоритм *LDA* хорошо реализован в библиотеке *sklearn*.

Алгоритмы *TextRank* и *Rutermextract* также включены в сервис; эти алгоритмы не требуют предобучения. Алгоритм *TextRank* основан на представлении текста в виде неориентированного графа. В качестве *вершин* графа извлекаются предложения из текста, а в качестве *рёбер* — коэффициент похожести. Алгоритм *Rutermextract* основывается на модернизированном подсчёте вхождения слов и словосочетаний в текст.

Реализованные алгоритмы выявления тематик текстов интегрированы в сервисе с использованием современных технологий: *Vue* 3, *Pinia*, *Vite*.

Пример демонстрации работы сервиса

В примере рассматривается статья с открытого источника Хабр «Динамическая сборка и деплой *Docker*-образов с *werf* на примере сайта версионированной документации»[13].

Сайт выделяет ключевые слова статьи автоматически: *werf*, *GitLab CI*, сборка проекта, *continuous delivery*, *Docker*. Текст статьи содержит материал по программированию, включает множество специфических терминов и эти термины в основном составляют набор ключевых слов. Такой подход хорош для узких специалистов, но может стать неудобным для более широкой аудитории читателей. Проведено исследование текста статьи с помощью разработанного сервиса (*рис*. 7).

[11] Векторное представление слов (англ. *word embedding*) — общее название для различных подходов к моделированию языка и обучению представлений в обработке естественного языка, направленных на сопоставление словам из некоторого словаря векторов небольшого размера.

[12] URL: http://www.maartengrootendorst.com/blog/ctfidf

[13] URL: https://habr.com/ru/companies/flant/articles/478690/

Рис. 7. Тематики, выделенные с помощью сервиса

Проанализировав работу четырех алгоритмов, заложенных в основу сервиса, можно заметить, что наиболее близкие наборы к определенным на сайте Хабр выявлены алгоритмами *LDA* и *TextRank*, алгоритм *BERT* в качестве тематик использовал только русскоязычные термины, причем односложные. Алгоритм *Rutermextract* выделил словосочетания как основные тематики. В целом все алгоритмы имеют общее пересечение и отвечают смысловой направленности текста статьи.

Заключение

Предлагаемый подход на основе различных алгоритмов построения тематических моделей позволяет выявить ключевые слова и соответствующие им тематики научных текстов по программированию. Результат работы сервиса позволяет систематизировать и классифицировать научные статьи и учебные материалы. Рассмотренные модели способны выделять тематики текста, что позволяет классифицировать содержание статьи без её прочтения.

На этапе реализации приложения использован язык программирования *Python* 3, а также библиотеки *sklearn*, *spacy*, *numpy*, *fastapi*.

Такой сервис может быть использован аспирантами, преподавателями и специалистами в *IT*-сфере для систематизации научных статей по программированию. При генерации датасета (набора данных) другой предметной области и обучения на нем реализованных тематических моделей можно расширить область выявления тематик и применять подобный сервис в других областях знаний.

Рецензент: ***Сухов Андрей Владимирович,*** *доктор технических наук, профессор, старший научный сотрудник 27 Центрального научно-исследовательского института Министерства обороны России, г. Москва, Российская Федерация.*
E-mail: avs57@mail.ru

Литература

1. Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // Машинное обучение и анализ данных. 2015. Т. 1. № 11. С. 1584—1618.
2. Апишев М. А. Эффективная реализация алгоритмов тематического моделирования // Труды Института системного программирования РАН. 2020. Т. 32. № 1. С. 223—240.
3. Бегларян М. Е., Добровольская Н. Ю. Формирование IT-компетенции юриста в цифровом пространстве // Правовая информатика. 2019. № 3. С. 60—70. DOI: 10.21681/1994-1404-2019-3-60-70 .
4. Воронцов К. В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека BigARTM. М. : Вильямс, 2023. 150 с. ISBN 978-5-534-09487-9.
5. Дударенко М. А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. 2015. Т. 16. С. 26—38.

6. Ерланова Р. Е., Нугуманова А. Б., Жантасова Ж. З., Байбурин Е. М. Тематическое моделирование текстовых учебных материалов по информатике средствами языка R // Известия АлтГУ. Математика и механика. 2018. № 4 (102). С. 68—72.
7. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды ИСП РАН. 2012. Т. 24. Вып. 4. С. 215—243.
8. Ловцов Д. А. Информационная теория эргасистем. Тезаурус : монография. М.: Наука, 2005. 248 с. ISBN 5-02-033779-X.
9. Ловцов Д. А., Богданова М. В., Лобан А. В., Паршинцева Л. С. Статистика (компьютеризированный курс) / Под ред. проф. Д. А. Ловцова. М. : РГУП, 2020. 400 с. ISBN 978-5-93916-834-2.
10. Ловцов Д. А., Бернацкая А. В. Русский язык и культура речи. М. : Форум, 2018. 156 с. ISBN 978-5-93673-189-1.
11. Тутубалина Е. В. Совместная вероятностная тематическая модель для идентификации проблемных высказываний, связанных нарушением функциональности продуктов // Труды ИСП РАН. 2015. Т. 27. Вып. 4. С. 111—128.
12. Машкин Д. О., Котельников Е. В. Извлечение аспектных терминов на основе условных случайных полей // Труды ИСП РАН. 2016. Т. 28. Вып. 6. С. 223—240.
13. Daud A., Juanzi L., Lizhu Z., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey. In: Proceedings of Frontiers of Computer Science in China. 2010. Pp. 280–225.

TOPIC MODELLING AS A SCHOLARLY SEARCH TOOL IN THE IT SPHERE

Margarita Beglarian, *Ph.D. (Physics & Mathematics), Associate Professor, Professor at the Department of the Humanities, Social and Natural Sciences Disciplines of the North-Caucasus Branch of the Russian State University of Justice, Krasnodar, Russian Federation.*
E-mail: rita_beg@mail.ru

Natal'ia Dobrovol'skaia, *Ph.D. (Paedadogy), Associate Professor at the Information Technology Department of the Kuban State University, Krasnodar, Russian Federation.*
E-mail: dnu10@mail.ru

Keywords: *research information, research paper, text, topic, topic modelling, service, algorithm, interface, search, search result.*

Abstract

Purpose of the paper: finding ways to improve the preparatory stage of writing a research paper or report using new information search processes based on topic modelling.

Methods used: a complex approach connecting the basic elements of the information search process in different IT areas, including programming; modelling the information selection process based on key characteristics; computational linguistics methods for identifying text levels with different degrees of development and formalisation: phonetic, morphological, lexical, syntactic and semantic; semantic methods for implementing topic modelling.

Study findings: an algorithm implemented as a service is developed allowing to identify keywords and corresponding topics of research texts in the field of programming. Different methods for building topic models make up the algorithm basis. The service makes it possible to systematise and classify research papers and educational materials both in programming and in other IT areas.

References

1. Aisina R. M. Obzor sredstv vizualizatsii tematicheskikh modelei kollektsii tekstovykh dokumentov. Mashinnoe obuchenie i analiz dannykh, 2015, t. 1, No. 11, pp. 1584–1618.
2. Apishev M. A. Effektivnaia realizatsiia algoritmov tematicheskogo modelirovaniia. Trudy Instituta sistemnogo programmirovaniia RAN, 2020, t. 32, No. 1, pp. 223–240.
3. Beglarian M. E., Dobrovol'skaia N.Iu. Formirovanie IT-kompetentsii iurista v tsifrovom prostranstve. Pravovaia informatika, 2019, No. 3, pp. 60–70. DOI: 10.21681/1994-1404-2019-3-60-70 .
4. Vorontsov K. V. Veroiatnostnoe tematicheskoe modelirovanie: teoriia reguliarizatsii ARTM i biblioteka BigARTM. M. : Vil'iams, 2023. 150 s. ISBN 978-5-534-09487-9.
5. Dudarenko M. A. Reguliarizatsiia mnogoiazychnykh tematicheskikh modelei. Vychislitel'nye metody i programmirovanie, 2015, t. 16, pp. 26–38.

6. Erlanova R. E., Nugumanova A. B., Zhantasova Zh.Z., Baiburin E. M. Tematicheskoe modelirovanie tekstovykh uchebnykh materialov po informatike sredstvami iazyka R. Izvestiia AltGU. Matematika i mekhanika, 2018, No. 4 (102), pp. 68–72.
7. Korshunov A., Gomzin A. Tematicheskoe modelirovanie tekstov na estestvennom iazyke. Trudy ISP RAN, 2012, t. 24, vyp. 4, pp. 215–243.
8. Lovtsov D. A. Informatsionnaia teoriia ergasistem. Tezaurus : monografiia. M. : Nauka, 2005. 248 pp. ISBN 5-02-033779-Kh.
9. Lovtsov D. A., Bogdanova M. V., Loban A. V., Parshintseva L. S. Statistika (komp'iuterizirovannyi kurs). Pod red. prof. D.A. Lovtsova. M. : RGUP, 2020. 400 s. ISBN 978-5-93916-834-2.
10. Lovtsov D. A., Bernatskaia A. V. Russkii iazyk i kul'tura rechi. M. : Forum, 2018. 156 s. ISBN 978-5-93673-189-1.
11. Tutubalina E. V. Sovmestnaia veroiatnostnaia tematicheskaia model' dlia identifikatsii problemnykh vyskazyvanii, sviazannykh narusheniem funktsional'nosti produktov. Trudy ISP RAN, 2015, t. 27, vyp. 4, pp. 111–128.
12. Mashkin D. O., Kotel'nikov E. V. Izvlechenie aspektnykh terminov na osnove uslovnykh sluchainykh polei. Trudy ISP RAN, 2016, t. 28, vyp. 6, pp. 223–240.
13. Daud A., Juanzi L., Lizhu Z., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey. In: Proceedings of Frontiers of Computer Science in China, 2010. Pp. 280–225.