УДК 004.81

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ПОИСКА ДЛЯ МАТЕМАТИЧЕСКОГО АРХИВА ПУБЛИКАЦИЙ

А. А. Насибулин[1] [0009-0005-9092-2520], О. М. Атаева[2] [0000-0003-0367-5575]

[1]*Московский физико-технический институт, г. Долгопрудный, Россия*

[2]*Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Россия*

[1]nasibulin.aa@phystech.edu, [2]oataeva@frccsc.ru

Аннотация

В работе проведено исследование, связанное с поиском схожих документов по математике. Разработан рекомендательный алгоритм нахождения похожих научных статей по данной тематике, использующий приоритетный поиск по математическим формулам с текстовым подкреплением.

Выполнен перевод текста из графического в текстовое представление через технологию OCR для последующего анализа и индексации. В процессе анализа реализовано разбиение текста на блоки с последующим извлечением из текста значимых формул, ключевых слов и фраз. В процессе индексации сформирована векторная база данных на основе векторных представлений формул, полученных через процесс эмбеддинга. Результаты индексации использованы при поиске статей, имеющих сходство с документом, подаваемым пользователем на вход алгоритма. Получен список похожих статей с сортировкой результатов по метрике близости векторных представлений формул.

Исходные данные представляют собой около 5000 научных статей, посвященных различным исследованиями по математической тематике и представленных в виде PDF-файлов.

Эксперимент проведен на основе данных конкретного контента библиотечной системы, но предложенная технология может быть распространена на другие библиотечные системы, в том числе содержащие статьи по другим тематикам, например, по физике и другим точным наукам.

Ключевые слова*: поиск по формулам, семантика, извлечение знаний, математический поиск, семантический поиск.*

ВВЕДЕНИЕ

В настоящее время остро стоит проблема поиска информации по научным статьям из-за роста объемов данных и стремительного увеличения сложности текстов в различных предметных областях, особенно в предметной области «математика». Классические алгоритмы лексического поиска по тексту (например, BM25), применяемые в большинстве рекомендательных систем, уступают алгоритмам семантического поиска с использованием Large Language Model (большая языковая модель, далее – LLM) [1]. Примером такого алгоритма является RAG с использованием LLM-модели Qwen3. Однако LLM-алгоритмы, как и классические, некорректно обрабатывают семантические значения математических формул, из-за чего в поисковой выдаче либо присутствуют некорректные совпадения формул по смыслу, либо отсутствуют похожие по смыслу формулы [2–4]. Отдельный поиск по формулам без использования текстового содержания хотя и эффективен для поиска по математическим статьям (например, Approach0 [12] достигает метрики качества $\mathrm{nDCG}' = 0.72$ на задаче поиска по формулам ARQMath-3), но все еще не задействует текстовые данные, использование которых может значительно улучшить поиск по статьям [5].

Таким образом, раздельное использование указанных выше алгоритмов поиска имеет ограничения, такие как потеря контекста из математических формул и текста или их некорректное распознавание. Поэтому для преодоления названных ограничений мы предлагаем комбинировать несколько из упомянутых подходов для достижения наилучших результатов поиска, т. е. создать систему гибридного поиска по текстам и формулам, которые в них содержатся. Таким образом, поставлена следующая задача: разработать и протестировать рекомендательный алгоритм нахождения похожих научных статей по математике, использующий приоритетный поиск по математическим формулам с текстовым подкреплением, т. е. реализовать поиск по математическим формулам, извлеченным из текста, с фильтрацией результатов этого поиска по семантическому сходству текстов. На вход алгоритма как для поиска, так и для индексирования подаются документы в распространенном формате – PDF. На выходе алгоритм выдает список найденных похожих статей, отсортированный по коэффициенту их совпадения с анализируемой статьей. Это и есть алгоритм поиска научных статей по математическим формулам с текстовым подкреплением.

БЛИЗКИЕ ПО ТЕМАТИКЕ ИССЛЕДОВАНИЯ

Для поиска по текстам, в том числе по научным текстам по математике, используют различные подходы поиска информации [5]. Наиболее распространенные варианты поиска информации можно разделить следующим образом: поиск по семантическим аннотациям из PDF-файлов [6], поиск с помощью RAG-систем [7] и поиск по онтологиям через граф знаний [8].

Системы Retrieval-Augmented Generation (генерация, дополненная поиском, далее – RAG) разделяются на два типа: классический RAG, в котором используется поиск по векторной базе данных (БД), и GraphRAG, использующий граф знаний. Оба типа систем на этапе анализа текстов при поиске или индексации не используют специфичную обработку формул, интерпретируя их как обычный текст, из-за этого поиск по формулам по существу не работает: например, после индексации более 5000 статей математического содержания системы не способны корректно различать формулы. Более того, из-за специфики хранения документов в RAG-системах поиск по точным совпадениям формул из проиндексированных документов невозможен.

Наиболее распространенные подходы такого поиска информации по онтологиям можно охарактеризовать следующим образом.

1. Поиск по графу знаний, создаваемому с использованием LLM [8–10]. Из-за того, что при построении графа знаний используются запросы к LLM, онтология, содержащаяся в ответе от нее, может быть некорректной (например, может быть извлечена только часть нужной формулы), так как у LLM-моделей есть склонность к ошибкам при ответе на математические вопросы [5].

2. Поиск по базе знаний, состоящей из текстов в векторном представлении, полученном через эмбеддинги [11]. Этот подход используется при построении RAG-систем и он не подразумевает специфичную обработку для формул, из-за чего на большом массиве данных поиск по формулам становится невозможным.

Что касается алгоритмов поиска по формулам, то их можно разделить на три типа по виду представления формул [5].

1. Представление математических формул в виде дерева расположения символов (Symbol Layout Tree, SLT) или дерева операторов (Operator Tree, OPT). Это наиболее распространенный вид представления формул. Примеры современных алгоритмов с его использованием: Approach0 [12], BERT (модель

MathBERT [13]), Tangent-CFT [14]. Эти алгоритмы предназначены только для поиска по формулам TeX-представлении, без контекста (текста), который может быть представлен вместе с формулой или вместо нее.

2. Представление в текстовом виде – используется, например, в алгоритме поиска по самой длинной общей подпоследовательности (Longest Common Subsequence, LCS) [15]. В современных алгоритмах такой вид представления не используется, так как поиск «по формулам как по текстам» неэффективен из-за особенностей их представления [16], которые невозможно учесть при использовании лексического поиска.

3. Представление в виде LEAN-кода (https://lean-lang.org/). Примером алгоритма, производящего поиск формул в таком представлении, является LeanSearch (существует еще версия без LLM) [17]. Этот вид представления используется в рекомендательных системах для ответов на математические вопросы, зачастую с использованием LLM. Применение такого представления в настоящей работе не рассматривается.

Существуют также системы гибридного поиска, из которых лучшим по совокупности показателей (SOTA – State-of-the-Art) решением для поиска по текстам с формулами является MABOWDOR [18]. Эта система использует комбинацию технологий: алгоритм поиска по формулам на основе SLT-дерева Approach0 [12], собственную BERT-подобную модель Coco-MAE [18] для поиска по формулам и алгоритм лексического поиска по текстам BM25+ [5] для поиска по текстам. Отметим, что эта система не адаптирована к текстам на русском языке.

Все приведенные примеры алгоритмов предполагают использование материалов в текстовом виде с формулами, представленными в TeX-нотации, и, соответственно, проводят оценку качества работы алгоритмов на таких инструментах оценки качества математического поиска, как ARQMATH [19] и NTCIR-12 [20]. Так как нами рассматриваются тексты научных работ по математике, которые содержат формулы, необходимо учесть, что значительная часть хранимых работ представлена в виде изображений или соответствующих PDF-файлов. Поэтому в таких случаях нужно использовать технологию распознавания текста из изображений/документов (OCR), чтобы гарантировать возможность работы разрабатываемого поискового алгоритма со статьями, представленными как в текстовом виде, так и в виде изображений, в том числе в PDF-файлах со специфичной ко-

дировкой или нестандартным форматированием. Так, на основании анализа результатов работы инструмента оценки качества работы алгоритмов OCR OmniDocBench [21] (точнее, интерпретации данных, собранных с помощью этой утилиты) нами была выбрана SOTA-технология набора инструментов распознавания текстов (OCR) PaddleOCR (PP-StructureV3). Эта технология производит экспорт формул в TeX-формат и сохраняет информацию о структуре документа, позволяя свести представление документов к единому формату, что необходимо для корректной работы системы поиска похожих работ.

Таким образом, использование комбинации распознавания PDF-файлов и гибридного поиска по математическим формулам с текстовым подкреплением ранее не рассматривалось (были использованы только TeX-файлы).

В настоящем исследовании для поиска по формулам использована эмбеддинг-модель MathBERT [13]. Она находит похожие формулы по семантике. Эта модель была обучена на учебниках и статьях по математике с сайта arxiv.org [13]. Ее особенность состоит в том, что она позволяет производить семантический поиск по формулам через их векторизацию и последующий поиск по расстоянию между ними (например, используя векторную базу данных scann [24]), как в RAG-системах. Пример работы поиска по формулам представлен на рис. 1. В этом примере был использован индекс из 100 различных формул.

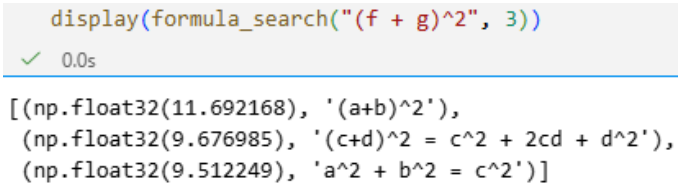


Рис. 1. Результат работы MathBERT.

Дополнительно была использована эмбеддинг-модель embeddinggemma-300m [22] для извлечения ключевых слов из текстов статей. Эта модель общего назначения, обученная для работы с мультиязычными текстами и имеющая лучшие показатели качества по сравнению с часто используемыми моделями bge-m3 и Qwen3Embedding 0.6B, при сниженном практически вдвое количестве параметров выдает при запросе более релевантные ключевые слова, как показано на рис. 2 (обеим моделям на вход подавалась аннотация статьи по семантическому поиску [23], дополнительный промпт не использован). Отметим, что

только модель embeddinggemma-300m экспортировала фразу «библиотека Libmeta». Экспорт этой фразы в данном примере примечателен тем, что ее наличие определяет возможность дальнейшего качественного поиска, так как она фактически является ключевой фразой для всего текста.

	model_name	keywords
0	sci_rus_small	междисциплинарного журнала, другие журналы, исследуется тематическое, журнала предлагается, журнала тематического, междисциплинарным исследованиям, статьи журнала, онтология журнала, экспертам журнала, интегрируются семантическую
1	frida_model	междисциплинарного журнала, журнала тематического, семантическую библиотеку, семантической библиотеке, онтология журнала, предлагается систематизация, междисциплинарной предметной, статьи журнала, тематической рубрики, анализа тематики
2	embeddinggemma_model	междисциплинарного журнала, журнала тематического, междисциплинарным исследованиям, анализа тематики, междисциплинарной предметной, журналов относящихся, анализа контента, контенту журнала, библиотеке libmeta многообразие междисциплинарного
3	bge_m3_model	междисциплинарного журнала, онтология журнала, междисциплинарным исследованиям, анализа тематики, исследуется тематическое, журнала тематического, тематического анализа, журнала интегрируются, журнала вырабатывается, междисциплинарной предметной
4	qwen3embedding_model	междисциплинарного журнала, знаний журнала, журнала тематического, тематического анализа, междисциплинарным исследованиям, анализа тематики, статьи журнала, тематической рубрики, онтология журнала, исследуется тематическое

Рис. 2. Сравнение работы BERT-моделей.

ОПИСАНИЕ ДАТАСЕТА

В качестве корпуса научных статей по математике был использован датасет научных работ на русском языке по математике и физике, содержащих математические формулы. Датасет представляет собой набор PDF-файлов порядка 5000 статей. Все статьи взяты из журнала «Известия высших учебных заведений. Математика» с 1997 по 2007 г. (1136 статей) и журнала «Вестник Тамбовского университета. Серия: Естественные и технические науки» с 2000 по 2013 г. (3892 статьи). Датасет содержит как вложенные в текст формулы (далее – inline-формулы), так и выделенные формулы (далее – outline-формулы).

МЕТОДИКА ПОИСКА

Гибридный поиск был реализован с помощью алгоритма поиска похожих математических статей, схема которого представлена на рис. 3.

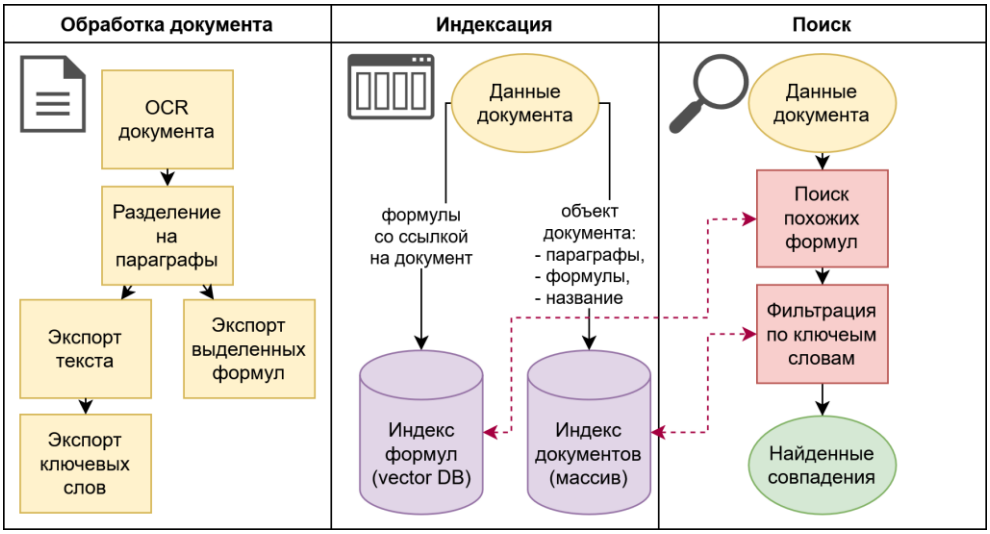


Рис. 3. Схема работы алгоритма

Документ можно разделить на несколько составляющих: текст (содержит inline-формулы), outline-формулы, таблицы, картинки (графики, схемы, фотографии и т. д.). Алгоритм в текущем виде работает только с текстом и формулами.

При обработке на вход алгоритму подается PDF-файл любого типа (страницы с произвольным текстовым и графическим содержанием, в том числе допустимо представление текстов в виде картинок или сканов). Из текста экспортируются и обрабатываются текст и формулы. Перечислим основные шаги работы алгоритма.

1. Обработка (анализ) документа. На этом этапе документ преобразуется в текстовое представление с помощью набора инструментов распознавания PaddleOCR, затем полученное текстовое представление конвертируется в объектное представление: текст документа разделяется на блоки (параграфы/значимые блоки текста и формулы), и из этих блоков с помощью BERT-модели EmbeddingGemma [22] производится экспорт ключевых слов. В сравнении с использованием всего текста как одного блока такой подход позволяет более точно извлекать контекст из каждого логически выделенного блока документа. Это достигается за счет уменьшенного количества текста на каждое вхождение при поиске. Пример обработки документа представлен на рис. 4.

2. Индексация. С помощью BERT-модели MathBERT [13] выполняется эмбеддинг найденных TeX-формул в векторные представления, которые добавляются в векторную базу данных scann [24] и в дальнейшем используются при выполнении поиска по документам. Каждая формула имеет ссылку на объект

блока документа, в котором она содержится. Используются только outline-формулы, т. е. формулы, выделенные из текста. Описанные данные представляют собой индекс математических формул. При проведении эксперимента индекс состоял из 144018 формул. Построение индекса заняло около 17 мин на процессоре AMD Ryzen 7840H.

3. Поиск. Сначала производится семантический поиск схожих формул по ранее полученному индексу математических формул. Результатом являются сопоставления «формула – блок статьи». Далее блоки статей фильтруются по ключевым словам, т. е. происходит лексический поиск по словам. Из отфильтрованного списка блоков извлекаются ссылки на статьи, в которых они содержатся. Результатом поиска является список похожих статей, отсортированных по величине метрики их сходства. Такой метрикой служит коэффициент расстояния между векторными представлениями формул (важно – это не значение расстояния между векторами, а коэффициент, обратно пропорциональный расстоянию между двумя векторами). Этот коэффициент вычисляется с использованием библиотеки базы данных scann.

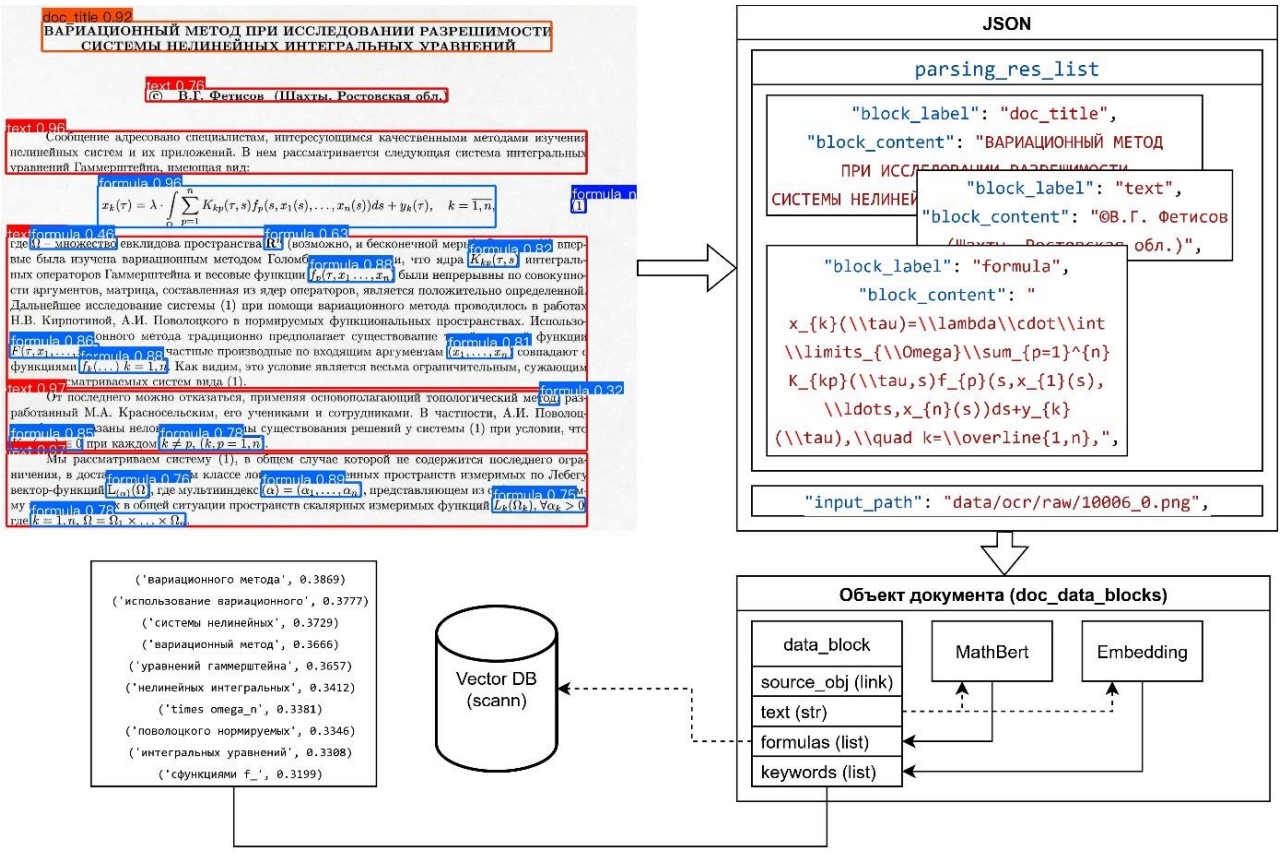


Рис. 4. Пример обработки документа.

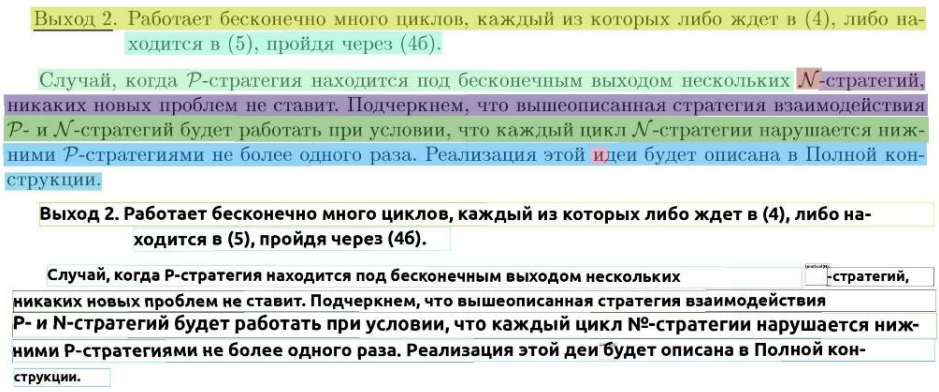
Далее отметим обнаруженные проблемы при разработке и тестировании алгоритма.

1. PaddleOCR имеет редкие ошибки распознавания содержащихся внутри текста inline-формул: часть из текста «переходит» в корпус формулы, а часть формул распознается как обычный текст, как показано на рис. 5. Для решения этой проблем требуется либо дообучение моделей распознавания, либо использование нормализации текстов. Для уменьшения последствий этой проблемы использовано не точное сравнение ключевых слов, а расстояние редактирования между ними [25] (см. пример на рис. 6).

2. Использование всего текста документа без дробления на части для извлечения ключевых слов через BERT-модель неэффективно, в том числе из-за ложных срабатываний (например, обнаруженное моделью сочетание слов «функция для» не может считаться ключевым) и потери большой части информации. Поэтому документы разделялись на блоки (главы) для сужения контекста, извлекаемого из текста, что позволило модели более точно выделять релевантные ключевые слова, за счет чего улучшается качество поиска.

3. Если в индексируемой статьей не было найдено outline-формул, то эта статья не появляется в списке найденных документов. Для обхода этой проблемы мы попытались использовать inline-формулы: при их добавлении количество найденных статей увеличивается, но при этом нерелевантные статьи начинают считаться релевантными, из-за этого был оставлен поиск только по outline-формулам.

Из датасета, содержащего более 5000 статей по математике на русском языке, экспертами были отобраны 50 пар статей, схожих по тематике. При тестировании на вход алгоритму подавался первый документ из пары для поиска похожих текстов. Если в топ-10 похожих документов содержался второй документ из пары, то поиск считался успешным. Таким образом, при тестировании алгоритма были успешно найдены 35 пар похожих математических статей. Для сравнения: при использовании лексического поиска (BM25+) пар не было найдено вообще, а при поиске по алгоритму RAG (doc2vec) было успешно найдено только 18 пар. Это подтверждает эффективность предложенного подхода к поиску похожих документов по сравнению с другими популярными алгоритмами.



Выход 2. Работает бесконечно много циклов, каждый из которых либо ждет в (4), либо находится в (5), пройдя через (4б).

Случай, когда $\mathcal{P}$-стратегия находится под бесконечным выходом нескольких $\mathcal{N}$-стратегий, никаких новых проблем не ставит. Подчеркнем, что вышеописанная стратегия взаимодействия $\mathcal{P}$- и $\mathcal{N}$-стратегий будет работать при условии, что каждый цикл $\mathcal{N}$-стратегии нарушается нижними $\mathcal{P}$-стратегиями не более одного раза. Реализация этой идеи будет описана в Полной конструкции.

Выход 2. Работает бесконечно много циклов, каждый из которых либо ждет в (4), либо находится в (5), пройдя через (4б).

Случай, когда P-стратегия находится под бесконечным выходом нескольких -стратегий, никаких новых проблем не ставит. Подчеркнем, что вышеописанная стратегия взаимодействия P- и N-стратегий будет работать при условии, что каждый цикл №-стратегии нарушается нижними P-стратегиями не более одного раза. Реализация этой деи будет описана в Полной конструкции.

Рис. 5. Примеры проблем при распознавании формул.

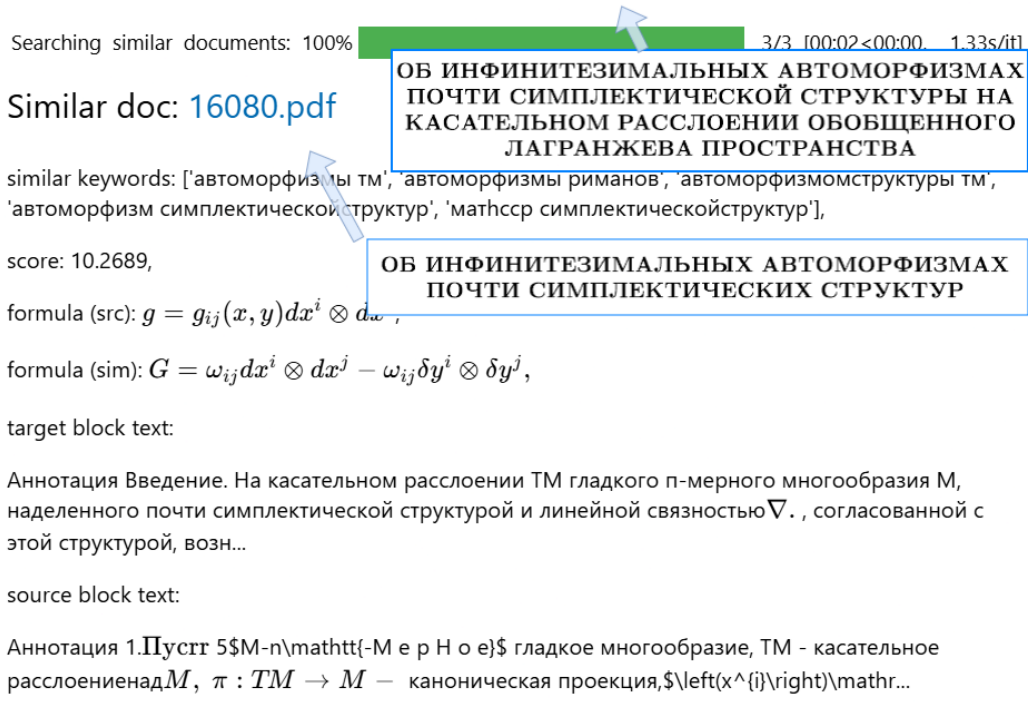
```
import difflib
from nltk.stem.snowball import SnowballStemmer
stemmer = SnowballStemmer("russian")

words1, words2 = ["примеры", "тестим", "маJIинное", "обучение"], ["пример0в", "тест", "машина", "обучаться"]
stem_words1, stem_words2 = [stemmer.stem(word) for word in words1],  [stemmer.stem(word) for word in words2]
for word in stem_words1:
    print(f"Word: {word}, Closest: {difflib.get_close_matches(word, stem_words2, cutoff=0.7)}")
```

✓ 0.0s

```
Word: пример, Closest: ['пример0в']
Word: тест, Closest: ['тест']
Word: маJIин, Closest: ['машин']
Word: обучен, Closest: ['обуча']
```

Рис. 6. Примеры сравнения слов.

Пример результата работы алгоритма приведен на рис. 7.



```
Searching for similar documents of data/pdf_tex/raw/16824.pdf...
```

Searching similar documents: 100% 3/3 [00:02<00:00, 1.33s/it]

ОБ ИНФИНИТЕЗИМАЛЬНЫХ АВТОМОРФИЗМАХ ПОЧТИ СИМПЛЕКТИЧЕСКОЙ СТРУКТУРЫ НА КАСАТЕЛЬНОМ РАССЛОЕНИИ ОБОБЩЕННОГО ЛАГРАНЖЕВА ПРОСТРАНСТВА

Similar doc: 16080.pdf

similar keywords: ['автоморфизмы тм', 'автоморфизмы риманов', 'автоморфизмомструктуры тм', 'автоморфизм симплектическойструктур', 'mathccp симплектическойструктур'],

score: 10.2689,

ОБ ИНФИНИТЕЗИМАЛЬНЫХ АВТОМОРФИЗМАХ ПОЧТИ СИМПЛЕКТИЧЕСКИХ СТРУКТУР

formula (src): $g = g_{ij}(x, y)dx^i \otimes d$,

formula (sim): $G = \omega_{ij}dx^i \otimes dx^j - \omega_{ij}\delta y^i \otimes \delta y^j$,

target block text:

Аннотация Введение. На касательном расслоении TM гладкого n-мерного многообразия M, наделенного почти симплектической структурой и линейной связностью∇. , согласованной с этой структурой, возн...

source block text:

Аннотация 1.Пусгг 5$M-n\mathtt{-M е р Н о е}$ гладкое многообразие, TM - касательное расслоениенадM, $\pi : TM \to M$ — каноническая проекция,$\left(x^{i}\right)\mathr...

Рис. 7. Пример работы алгоритма.

ЗАКЛЮЧЕНИЕ

Основным полученным результатом является успешное нахождение похожих математических статей на тестируемом приватном датасете научных публикаций на русском языке с использованием алгоритма, основанного на поиске по формулам с текстовым подкреплением. Алгоритм решает проблему поиска похожих научных статей на русском языке, которые содержат математические формулы, и в отличие от его аналогов позволяет производить поиск по текстам, не переведенным в текстовый формат: он успешно находит похожие статьи математического характера, представленные в PDF-формате. Полученные результаты могут стать отправной точкой для дальнейших исследований по разработке алгоритма поиска похожих статей по математике в PDF-формате.

СПИСОК ЛИТЕРАТУРЫ

1. *Stuhlmann L., Saxer M. A., Fürst J.* Efficient and Reproducible Biomedical Question Answering using Retrieval Augmented Generation // arXiv:2505.07917v2. https://doi.org/10.48550/arXiv.2505.07917

2. *Polyanin A.D., Shingareva I.K.* The similarity index of mathematical and other scientific publications with equations and formulas and the problem of self-plagiarism identification // arXiv:2110.03872. https://doi.org/10.48550/ arXiv.2110.03872

3. *Wang R. et al.* Evaluation of LLMs for mathematical problem solving // arXiv:2506.00309. https://doi.org/10.48550/arXiv.2506.00309

4. *Forootani A.A*. survey on mathematical reasoning and optimization with Large Language Models // arXiv:2503.17726. https://doi.org/10.48550/arXiv.2503.17726

5. *Zanibbi R. et al*. Mathematical Information Retrieval: Search and Question Answering // arXiv:2408.11646v3. https://doi.org/10.48550/arXiv.2408.11646

6. *Невзорова О.А., Николаев К.С.* Семантическое аннотирование математических формул в PDF-документах // Электронные библиотеки. 2022. Т. 25, № 6. С. 616–639. https://doi.org/10.26907/1562-5419-2022-25-6-616-639

7. *Chen E. et al.* Comparing RAG and GraphRAG for Page-Level Retrieval Question Answering on Math Textbook // arXiv:2509.16780. https://doi.org/10.48550/arXiv.2509.16780

8. *Feng X. et al*. Ontology-grounded automatic Knowledge Graph construction by LLM under wikidata schema // arXiv:2412.20942. https://doi.org/10.48550/arXiv.2412.20942

9. *Lippolis A.S. et al*. Ontology Generation using Large Language Models // arXiv:2503.05388. https://doi.org/10.48550/arXiv.2503.05388

10. *Khasanshin A. et al.* Indexing mathematical scholarly papers as linked open data // Proceedings of the Sixth Russian Young Scientists Conference in Information Retrieval (VI Russian Summer School in Information Retrieval), 2012. P. 24–34. https://doi.org/10.18653/v1/P19-1023

11. *Trisedya B.D. et al.* Neural relation extraction for knowledge base enrichment, in: A. Korhonen, D. Traum, L. Màrquez (Eds.) // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, P. 229–240. https://doi.org/10.18653/v1/P19-1023

12. *Zhong W., Xie Y., Lin J. et al.* Applying Structural and Dense Semantic Matching for the ARQMath Lab 2022, CLEF // CLEF (Working Notes). 2022. P. 147-170.

13. *Shen J. T. et al*. MathBERT: A Pre-trained Language Model for General NLP Tasks in Mathematics Education // arXiv:2106.07340. https://doi.org/10.48550/arXiv.2106.07340

14. *Mansouri B. et al.* Tangent-CFT: An embedding model for mathematical formulas //Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval. 2019. P. 11–18. https://doi.org/10.1145/3341981.3344235

15. *Kumar P., Agarwal A., Bhagvati C*. A structure based approach for mathematical expression retrieval // A Structure Based Approach for Mathematical Expression Retrieval // In: Sombattheera C., Loi N.K., Wankar R., Quan T. (Eds.) Multi-disciplinary Trends in Artificial Intelligence. MIWAI 2012. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2012. Vol. 7694. P. 23–34. https://doi.org/10.1007/978-3-642-35455-7_3

16. *Isele M.R.* Analyzing Similarity in Mathematical Content To Enhance the Detection of Academic Plagiarism // arXiv:1801.08439. https://doi.org/10.48550/arXiv.1801.08439

17. *Li I.R.* Towards Lightweight and LLM-Free Semantic Search for mathlib4 // AITP. 2025.

18. *Wei Zhong et al.* One Blade for One Purpose: Advancing Math Information Retrieval using Hybrid Search // In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23). 2023. P. 141–151. https://doi.org/10.1145/3539618.3591746

19. *Scharpf P. et al.* ARQMath Lab: An Incubator for Semantic Formula Search in zbMATH Open? // arXiv:2012.02413. https://doi.org/10.48550/arXiv.2012.02413

20. *Zanibbi R. et al.* NTCIR-12 MathIR Task Overview // NTCIR. 2016.

21. *Ouyang L. et al.* OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations // arXiv:2412.07626. https://doi.org/10.48550/arXiv.2412.07626

22. *Vera H.S. et al.* EmbeddingGemma: Powerful and Lightweight Text Representations // arXiv:2509.20354. https://doi.org/10.48550/arXiv.2509.20354

23. *Ataeva O.M. et al.* Data mining when constructing a knowledge graph of a multidisciplinary journal // Information and mathematical technologies in science and management. 2024. Vol. 3 (35). P. 5–19.

24. *Refahi S.M. et al.* Fast and Scalable Gene Embedding Search: A Comparative Study of FAISS // arXiv:2507.16978. https://doi.org/10.48550/arXiv.2507.16978

25. *Python developers.* Documentation of library difflib // Python 3.14.3 documentation.

DEVELOPMENT OF AN INTELLIGENT SEARCH SYSTEM FOR THE MATHEMATICAL ARCHIVE OF PUBLICATIONS

A. A. Nasibulin[1][0009-0005-9092-2520], O. M. Ataeva[2][0000-0003-0367-5575],

[1]*Moscow Institute of Physics and Technology, Dolgoprudny, Russia*

[2]*Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia*

[1]nasibulin.aa@phystech.edu, [2]oataeva@frccsc.ru

Abstract

A study was conducted on searching for similar documents. The goal was to create a recommendation algorithm for finding similar scientific articles in mathematics using a prioritized search of mathematical formulas with textual support.

The text was converted from graphical to textual representation using OCR technology for subsequent analysis and indexing. During the analysis process, the text was divided into blocks, followed by the extraction of significant formulas, keywords, and phrases from the text. During the indexing process, a vector database was formed based on vector representations of formulas obtained through the embedding process. The indexing results were used to search for articles that are similar to the document submitted by the user to the algorithm input. A list of similar articles is displayed with results sorted by the metric of closeness of vector representations of formulas.

The source data consisted of approximately 5,000 scientific articles devoted to various studies on mathematical topics and presented as PDF files. The experiment was conducted based on data from specific library system content, but the proposed technology can be extended to other library systems, including those containing articles on other topics, such as physics and other exact sciences.

***Keywords**: formula search, semantics, knowledge extraction, mathematical search, semantic search.*

REFERENCES

1. *Stuhlmann L., Saxer M.A., Fürst J.* Efficient and Reproducible Biomedical Question Answering using Retrieval Augmented Generation // arXiv:2505.07917v2. https://doi.org/10.48550/arXiv.2505.07917

2. *Polyanin A.D., Shingareva I.K.* The similarity index of mathematical and other scientific publications with equations and formulas and the problem of self-plagiarism identification // arXiv:2110.03872.
https://doi.org/10.48550/ arXiv.2110.03872

3. *Wang R. et al.* Evaluation of LLMs for mathematical problem solving // arXiv:2506.00309. https://doi.org/10.48550/arXiv.2506.00309

4. *Forootani A.A*. Survey on mathematical reasoning and optimization with Large Language Models // arXiv:2503.17726.
https://doi.org/10.48550/arXiv.2503.17726

5. *Zanibbi R. et al*. Mathematical Information Retrieval: Search and Question Answering // arXiv:2408.11646v3. https://doi.org/10.48550/arXiv.2408.11646

6. *Nevzorova O.A., Nikolaev K.S.* Semantic Annotation of Mathematical Formulas in PDF-Documents // Russian Digital Libraries. 2022. Vol. 25. No. 6. P. 616–639.

https://doi.org/10.26907/1562-5419-2022-25-6-616-639

7. *Chen E. et al.* Comparing RAG and GraphRAG for Page-Level Retrieval Question Answering on Math Textbook // arXiv:2509.16780. https://doi.org/10.48550/arXiv.2509.16780

8. *Feng X. et al*. Ontology-grounded automatic Knowledge Graph construction by LLM under wikidata schema // arXiv:2412.20942. https://doi.org/10.48550/arXiv.2412.20942

9. *Lippolis A.S. et al*. Ontology Generation using Large Language Models // arXiv:2503.05388. https://doi.org/10.48550/arXiv.2503.05388

10. *Khasanshin A. et al.* Indexing mathematical scholarly papers as linked open data // Proceedings of the Sixth Russian Young Scientists Conference in Information Retrieval (VI Russian Summer School in Information Retrieval), 2012. P. 24–34. https://doi.org/10.18653/v1/P19-1023

11. *Trisedya B.D. et al.* Neural relation extraction for knowledge base enrichment, in: A. Korhonen, D. Traum, L. Màrquez (Eds.) // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, P. 229–240. https://doi.org/10.18653/v1/P19-1023

12. *Zhong W., Xie Y., Lin J. et al.* Applying Structural and Dense Semantic Matching for the ARQMath Lab 2022, CLEF // CLEF (Working Notes). 2022. P. 147-170.

13. *Shen J.T. et al*. MathBERT: A Pre-trained Language Model for General NLP Tasks in Mathematics Education // arXiv:2106.07340. https://doi.org/10.48550/arXiv.2106.07340

14. *Mansouri B. et al.* Tangent-CFT: An embedding model for mathematical formulas //Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval. 2019. P. 11–18. https://doi.org/10.1145/3341981.3344235

15. *Kumar P., Agarwal A., Bhagvati C*. A structure based approach for mathematical expression retrieval // A Structure Based Approach for Mathematical Expression Retrieval // In: Sombattheera C., Loi N.K., Wankar R., Quan T. (Eds.) Multi-disciplinary Trends in Artificial Intelligence. MIWAI 2012. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2012. Vol. 7694. P. 23–34. https://doi.org/10.1007/978-3-642-35455-7_3

16. *Isele M.R.* Analyzing Similarity in Mathematical Content to Enhance the Detection of Academic Plagiarism // arXiv:1801.08439.

https://doi.org/10.48550/arXiv.1801.08439

17. *Li I.R.* Towards Lightweight and LLM-Free Semantic Search for mathlib4 // AITP. 2025.

18. *Wei Zhong et al.* One Blade for One Purpose: Advancing Math Information Retrieval using Hybrid Search // In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23). 2023. P. 141–151. https://doi.org/10.1145/3539618.3591746

19. *Scharpf P. et al.* ARQMath Lab: An Incubator for Semantic Formula Search in zbMATH Open? // arXiv:2012.02413. https://doi.org/10.48550/arXiv.2012.02413

20. *Zanibbi R. et al.* NTCIR-12 MathIR Task Overview // NTCIR. 2016.

21. *Ouyang L. et al.* OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations // arXiv:2412.07626. https://doi.org/10.48550/arXiv.2412.07626

22. *Vera H.S. et al.* EmbeddingGemma: Powerful and Lightweight Text Representations // arXiv:2509.20354. https://doi.org/10.48550/arXiv.2509.20354

23. *Ataeva O.M. et al.* Data mining when constructing a knowledge graph of a multidisciplinary journal // Information and mathematical technologies in science and management. 2024. Vol. 3 (35). P. 5–19.

24. *Refahi S.M. et al.* Fast and Scalable Gene Embedding Search: A Comparative Study of FAISS // arXiv:2507.16978. https://doi.org/10.48550/arXiv.2507.16978

25. Python developers. Documentation of library difflib // Python 3.14.3 documentation.

СВЕДЕНИЯ ОБ АВТОРАХ

НАСИБУЛИН Алексей Алексеевич – студент 2 курса магистратуры Московского физико-технического института по направлению «Науки о данных». Основные направления научных исследований: обработка естественного языка, компьютерное зрение, искусственный интеллект.

Aleksey Alekseevich NASIBULIN – second-year master's studentat MIPT in the field of «Data Science». Major fields of scientific research are Natural Language processing, computer vision and artificial intelligence.

email: nasibulin.aa@phystech.edu

ORCID: 0009-0005-9092-2520

АТАЕВА Ольга Муратовна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, к. т. н. Область научных интересов: системное программирование, базы данных, инженерия знаний и онтологии.

Olga Muratovna ATAEVA – senior researcher at the Dorodnicyn Computing Center, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences; Candidate of Technical Sciences (PhD). Research interests: systems programming, databases, knowledge engineering, ontologies. Number of scientific publications – 80.

email: oataeva@frccsc.ru

ORCID: 0000-0003-0367-5575

Материал поступил в редакцию 18 марта 2026 года