УДК 004.422+004.9

ФОРМИРОВАНИЕ СТРУКТУРИРОВАННЫХ ПРЕДСТАВЛЕНИЙ НАУЧНЫХ ЖУРНАЛОВ ДЛЯ ИНТЕГРАЦИИ В ГРАФ ЗНАНИЙ И СЕМАНТИЧЕСКОГО ПОИСКА

О. М. Атаева[1] [0000-0003-0367-5575], М. Г. Кобук[2] [0009-0002-9834-8218]

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва, Россия

[1, 2]Московский университет имени С.Ю. Витте, г. Москва, Россия

[1]oataeva@frccsc.ru, [2]mikhail.kobuk@mail.ru

Аннотация

Работа посвящена проблеме развития библиотеки научных предметных областей SciLibRu, как продолжения семантического описания научных трудов проекта LibMeta. В основе этой библиотеки лежит концептуальная модель данных, структура и семантика которой сформированы на принципах онтологического моделирования. Такой подход обеспечивает строгое описание предметной области, формализацию взаимосвязей между сущностями и возможность дальнейшего автоматизированного анализа данных. Целью настоящего исследования были разработка и экспериментальное применение методов структуризации содержимого научных журналов в формате LaTeX для их интеграции в онтологию библиотеки и обеспечения семантического поиска.

Предложен алгоритм трансляции в формат XML данных, представленных множеством файлов, для интеграции в онтологию библиотеки. Реализован модуль векторного поиска, основанный на вычислении эмбеддингов с использованием языковых моделей. Выявлены закономерности распределения эмбеддингов и факторы, влияющие на точность ранжирования результатов поиска. Проведено тестирование двух названых компонентов.

Разработанный метод составляет основу для автоматического включения содержимого научных журналов в граф знаний SciLibRu и создания обучающих корпусов для языковых моделей, ограниченных рамками научных предметных областей. Полученные результаты способствуют развитию систем навигации по графу

© О. М. Атаева, М. Г. Кобук, 2025.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

знаний журналов, а также рекомендательных механизмов и инструментов интеллектуального поиска по русскоязычным научным текстам.

Ключевые слова*: полуструктурированные данные, онтология текста, LaTeX, векторное представление текста, полнотекстовый поиск, семантический поиск.*

ВВЕДЕНИЕ

Развитие цифровых библиотек нового поколения связано с переходом от классических репозиториев публикаций к семантическим системам, обеспечивающим смысловую интеграцию и навигацию по данным. В рамках этого направления разрабатывается библиотека научных предметных областей SciLibRu. В основе этой библиотеки лежит концептуальная модель данных, структура и семантика которой сформированы на принципах онтологического моделирования. На базе этой онтологии строится граф знаний, обеспечивающий представление семантических связей между объектами предметной области и поддержку интеллектуальной обработки данных. Онтологическое проектирование составляет часть технологии строгого описания предметных областей и взаимосвязей между сущностями, служит для дальнейшей автоматизации процессов анализа данных. В библиотеке SciLibRu предлагается создание унифицированного пространства знаний научных предметных областей на основе научных публикаций, энциклопедий, классификаторов и тезаурусов. Развитие экосистемы SciLibRu требует интеграции разнородных источников, представленные в различных форматах. Анализ разнородных представлений данных и их автоматическое включение в онтологию библиотеки составляет предмет исследования.

В рамках работы рассматриваются прикладные задачи: разработка алгоритма трансляции слабоструктурированных публикаций в формат XML; реализация модуля семантического полнотекстового поиска по полученным данным. Этот модуль предназначен не только для поиска по содержанию, но и для оценки семантических связей между элементами текста. Модуль семантического полнотекстового поиска применяется в дальнейшем для интеграции данных в онтологию SciLibRu с последующим использованием в работе с большими языковыми моделями (Large Language Models, LLM).

В настоящем исследовании рассмотрен процесс структуризации содержимого научных журналов, включающего метаданные журналов и публикаций, а также структурные и смысловые единицы публикаций, такие как разделы, определения, теоремы и т. д. Этот процесс является одним из этапов формирования корпусов данных для библиотечной экосистемы SciLibRu, перехода от документного представления к формату, пригодному для машинного анализа и семантического поиска.

1. ПОСТАНОВКА ЗАДАЧИ

Построение модуля семантического поиска по научным текстам основано на приведении исходных документов к унифицированной форме с сохранением логической и структурной целостности материала. В рамках этой задачи предложен алгоритм структуризации данных и преобразования исходных файлов научных статей в представление, пригодное для последующей векторизации и анализа.

Особенность настоящего исследования заключается в обработке составных документов (журнальных выпусков, сборников статей) без многократного запуска конвертирующих процедур. Перечисленные документы содержат данные в различных форматах и используют авторские варианты TeX-макрокоманд. Предобработка данных учитывает структуру составных документов и таким образом обеспечивает целостную интерпретацию взаимосвязей между файлами и позволяет избежать типичных ограничений существующих инструментов, таких как tex4ht [1] и LaTeXML [2].

1.1. Аналогичные исследования

В области обработки научных текстов в настоящее время преобладают подходы с использованием различных языковых моделей, а классические подходы семантического анализа данных остаются предметом исследования и применяются в специфических задачах, где важны эффективность и работа со структурой языка.

Применительно к семантическому преобразованию неструктурированного LaTeX-текста активная работа сейчас ведется с использованием конвертера tex4ht [1]. Это решение использует не исходный LaTeX-текст, а его промежуточное

представление, формируемое компилятором. При этом упрощается задача анализа текста, так как не требуется работы с авторскими макросами или сложными вложенными структурами документа. Однако в дальнейшем возникает необходимость повторения среды компиляции для каждого файла, что увеличивает процедуру предобработки, особенно при работе с большим количеством разнородных источников.

Второе известное решение, LaTeXML [2], которое поддерживается проектом arXiv и применяется для преобразования тяжелых PDF-версий статей в более легковесные HTML. С историей в более чем 20 лет разработки это самое зрелое и полное решение из применяемых в системах накопления текстов. Однако стремление разработчиков сделать LaTeXML универсальным инструментов оказалось одновременно и существенным недостатком: конфигурация оказалась сложна и избыточна для отдельных источников файлов. Обработка любых нестандартных макрокоманд или стилей приводит к необходимости модификации конфигурации этого инструмента. Расширение строгих шаблонов позволяет повысить строгость вывода, однако любое несоответствие, даже некритичное, прерывает процесс преобразования.

Несмотря на длительную историю развития методов накопления оцифрованной информации, остаются востребованными исследования в области алгоритмов предобработки научных текстов с целью их интеграции, семантического представления и использования для LLM.

2. АЛГОРИТМ СТРУКТУРИЗАЦИИ СОСТАВНЫХ ДОКУМЕНТОВ

Разработанный алгоритм структуризации для потоков цифровых архивов журналов состоит из нескольких последовательных этапов. Определяется структура входного набора файлов и формируется дерево включений для многофайловых структур. Главный файл, содержащий ссылки на другие, идентифицируется как корневой элемент, процесс структуризации выполняется для каждого главного файла отдельно. Далее производится регистрация пользовательских макрокоманд, определенных в каждом файле дерева, с возможностью настройки приоритетов по принадлежности и времени обнаружения. Это облегчает корректное разрешение конфликтов при объединении различных областей документа [3].

В итоге формируется абстрактное синтаксическое дерево (AST) [4, 5] с выборочным раскрытием макрокоманд, которое используется для генерации целевого формата базы знаний, строгого XML-представления, предназначенного для последующего анализа и индексирования.

Для оценки корректности работы алгоритма слабой структуризации была проведена серия экспериментов на тестовом множестве из более чем 1000 файлов, представленных в формате LaTeX. В результате 89 файлов не были идентифицированы как формирующие XML-вывод, поэтому не учитывались при анализе.

Оставшиеся 1010 документов были организованы в виде научных статей, объединенных в журнальные выпуски. Полученные результаты показали, что безошибочно конвертирован был 521 файл (≈ 51.6% от общего количества). Еще 169 файлов (≈ 16.7% от общего количества) содержали исправимые ошибки, связанные преимущественно с особенностями макрокоманд и вложенной структурой включений.

В 219 случаях (≈ 21.7%) алгоритм завершился с фатальной ошибкой, вызванной нарушением внутренней структуры исходного документа (чаще всего отсутствием точки входа или выхода для конструкций естественного языка, например для описаний теорем и др.). Еще 101 файл (≈ 10%) потребовал доработки модулей лексического анализа и парсера (синтаксического анализатора), отвечающих за обработку примитивов формата LaTeX и выявление зависимостей между макрокомандами.

Анализ выходных XML-файлов показал, что около 15% результатов содержат структурные неточности, а в трети случаев сохранялись остаточные LaTeX-конструкции. Эти дефекты не препятствуют общей интерпретации данных, но требуют уточнения механизма фильтрации и нормализации синтаксических элементов, что и было выполнено на этапе доработки алгоритма.

Доработка алгоритма была направлена на повышение точности и устойчивости парсинга. Основное внимание уделялось расширению покрытия примитивов языка LaTeX и внедрению более гибких механизмов конфигурации, позволяющих управлять процессом раскрытия макрокоманд без усложнения пользовательского интерфейса.

Для повышения надежности было предложено использование стекового контекстного механизма и опережающего просмотра (lookahead). Такой обновленный подход позволяет корректно обрабатывать вложенные конструкции и выявлять ошибки на ранних этапах синтаксического анализа. Улучшения позволили обеспечить сокращение числа ошибок доработки примерно на 6%, а также уменьшить количество некорректных преобразований, связанных с конфигурацией, на 3%. Количество фатальных ошибок снизилось до ≈ 15%, что свидетельствует о повышении общей стабильности алгоритма.

Однако оптимизация породила новые трудности. Применение «жадного» алгоритма (greedy algorithm) [5] слияния текстовых фрагментов привело к неравномерному распределению размера тегов <text>. Кроме того, в некоторых случаях тег опускается для заголовков и структурных элементов, таких как <theorem> или <lemma>, что вызывает пропуски текстовых блоков при последующем поиске.

Несмотря на возможность включения полного содержимого каждого тега в результирующий XML, на данном этапе основной интерес представляет текстовая составляющая документа, исключающая формулы и другие элементы разметки. Такое строгое ограничение не оптимально, но эффективно для задач полнотекстового и семантического поиска.

3. АРХИТЕКТУРА ПОИСКОВОГО МОДУЛЯ

Следующим этапом исследования было построение полнотекстового семантического поиска по сформированной XML-базе данных. Исходные файлы, полученные в результате трансляции LaTeX-документов, содержат структурированные теги, среди которых основной интерес представляет тег `<text>`. Он включает фрагменты основного текста и может содержать вложенные теги `<formula>`, которые на этом этапе анализа исключались из рассмотрения.

3.1. Построение векторной базы данных

Для каждого выделенного текстового фрагмента формировалось числовое векторное представление – эмбеддинг, описывающий смысловое содержание текста, при помощи языковых моделей. Это представление отражает семантическую близость между фрагментами. Предполагается, что полученные векторы

должны образовывать хорошо разделенные и интерпретируемые кластеры, в которых тексты, близкие по содержанию, группируются в пределах одного смыслового «облака». На рис. 1 представлен схематически процесс формирования векторных представлений и поиска по ним.

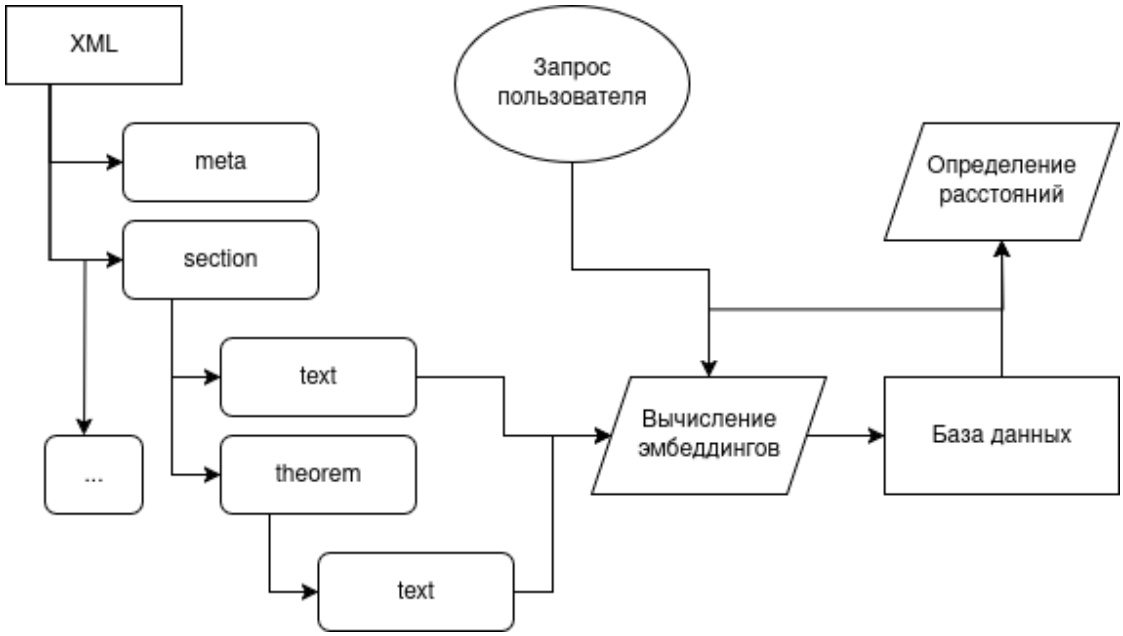

Рис. 1. Схематическое представление процесса формирования векторных представлений и поиска.

Опишем этот процесс на примере рассмотрения одной статьи. Каждый тег <text>, обнаруженный в выходном XML-файле, извлекается и передается в процедуру вычисления эмбеддингов, которая запускает выбранные методы векторизации, результат записывается в базу данных (БД) вместе с полным текстом фрагмента и метаданными источника. После заполнения БД, пользователь может осуществлять поиск. Для этого поисковый запрос (строка) векторизуется процедурой, которая применялась для вычисления эмбеддингов текста и сравнивается с содержимым БД.

Для вычисления эмбеддингов были использовано несколько языковых моделей: word2vec [6], GloVe [7], fastText [8], LaBSE [9], ruAdapt [10], ruBERT-base [11], ruBERT-tiny и ruGPT-3-medium [12]. В моделях первого типа (word2vec, GloVe, fastText) векторизация осуществлялась на уровне отдельных слов; итоговый вектор вычислялся как среднее арифметическое векторов всех слов фрагмента. В мо-

делях второго типа (LaBSE, ruAdapt, ruBERT, ruGPT-3) использовалось представление последнего скрытого состояния трансформера, что обеспечивало более сложное кодирование контекста [14, 15].

После получения векторов текстовых фрагментов формировалась векторная база данных. Запрос пользователя, представляющий собой строку произвольной длины, также преобразовывался в вектор (для этого применялась процедура со строго такими же параметрами, как для вычисления эмбеддингов), после чего вычислялось сходство между ним и векторами текстовых фрагментов. В качестве метрики использовалось косинусное расстояние [16–18], которое, несмотря на известные ограничения [2], остается одной из наиболее устойчивых и интерпретируемых мер близости в векторных пространствах.

Для повышения наглядности и удобства восприятия результатов пользователю предоставлялся ранжированный список: тексты сортировались по возрастанию разницы между векторами запроса и найденных фрагментов, что позволяло выделить наиболее релевантные соответствия.

3.2. Результаты оценки моделей поиска

Для первичной оценки качества сформированных эмбеддингов использовалось подмножество из корпуса, включающее 2088 тегов `<text>`, извлеченных из 29 предварительно обработанных XML-файлов. На основе полученных векторов была выполнена двумерная проекция с использованием метода UMAP (Uniform Manifold Approximation and Projection) [19], что позволило визуально оценить распределение фрагментов текста в пространстве признаков (см. рис. 2, 3).

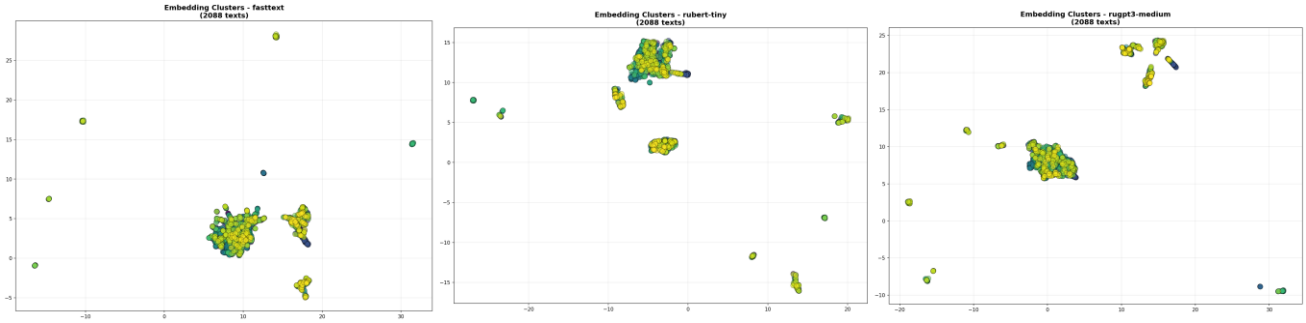

Рис. 2. UMAP-визуализация распределения эмбеддингов для fastText, ruBERT-tiny, ruGPT-3-medium.

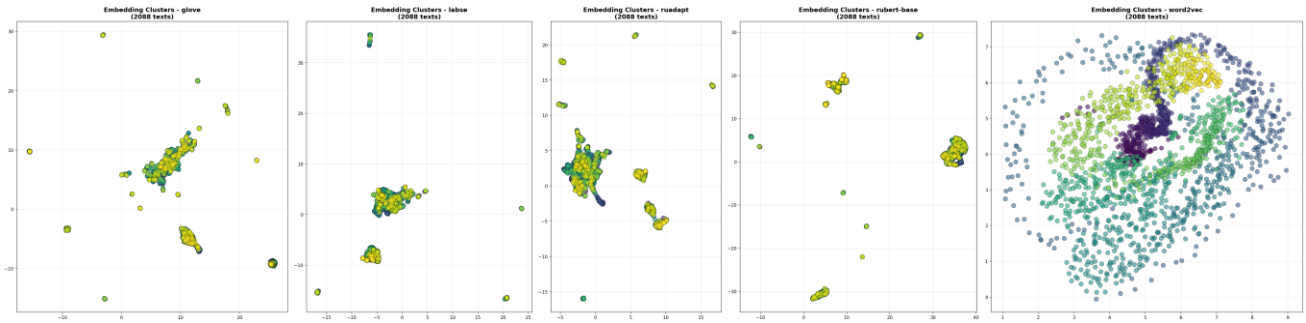

Рис. 3. UMAP-визуализация распределения эмбеддингов для GloVe, LaBSE, ruAdapt, ruBERT-base, word2vec.

Анализ показал, что наибольшая кластеризация эмбеддингов наблюдается при использовании модели word2vec, тогда как модели ruGPT-3-medium, ruBERT и ruAdapt продемонстрировали более выраженную способность к кластеризации по типам фрагментов, например к выделению аннотаций, подписей к изображениям и таблицам, а также основного научного текста.

Используемая модель word2vec-ruscorpora-300 [13] основана на морфологически аннотированном корпусе и применяет PoS-тегирование (Part of Speech), при котором каждый токен представлен в виде комбинации слова и части речи, например, «вода_NOUN» вместо «вода». Нестандартные токены, не входящие в словарь модели, маркировались как символы, что могло снижать точность представлений.

Качество поиска в целом оказалось ограниченным. Короткие запросы (из одного-двух слов) зачастую приводили к нерелевантным результатам, близким по эффективности к случайным совпадениям. Однако более длинные запросы (4–8 взаимосвязанных слов), особенно при использовании word2vec, позволили достигать смыслового соответствия между запросом и найденными фрагментами. При этом результаты содержали преимущественно длинные отрывки текста, включающие несколько смысловых единиц, что показано в табл. 1 на примере запроса *«поверхность шар»*.

Так, при запросе *«теория вероятностей»* в число десяти наиболее близких фрагментов входили три последовательных абзаца одной статьи (объединенные как единый результат), где рассматривался метод теории массового обслуживания с многочисленными упоминаниями вероятностных подходов.

Табл. 1. Результаты ранжирования трёх наиболее близких векторов по запросу «поверхность шар» над word2vec

Ранжирование	Результат
1	Preview: «Таким образом, в однодипольной модели обратная задача решается следующим способом. Нужно найти на поверхности сферы две точки с максимальным значением радиальной компоненты магнитного поля»
2	Preview: «т. е. функция ширины экрана непрерывна и в окрестности точки»
3	Preview: «Функция ( ) выпукла вниз по»

3.3. Доработка алгоритма

Результаты первоначального эксперимента показали, что качество поиска заметно зависит от длины текстового фрагмента, используемого при векторизации: проявилась проблема «жадности» алгоритма, упомянутая выше. Несмотря на то что количество относительно длинных фрагментов было невелико (см. рис. 4), длинные участки текста, включающие несколько смысловых единиц, приводят к размыванию семантического центра и, как следствие, к снижению точности сопоставления запросов с релевантными контекстами.

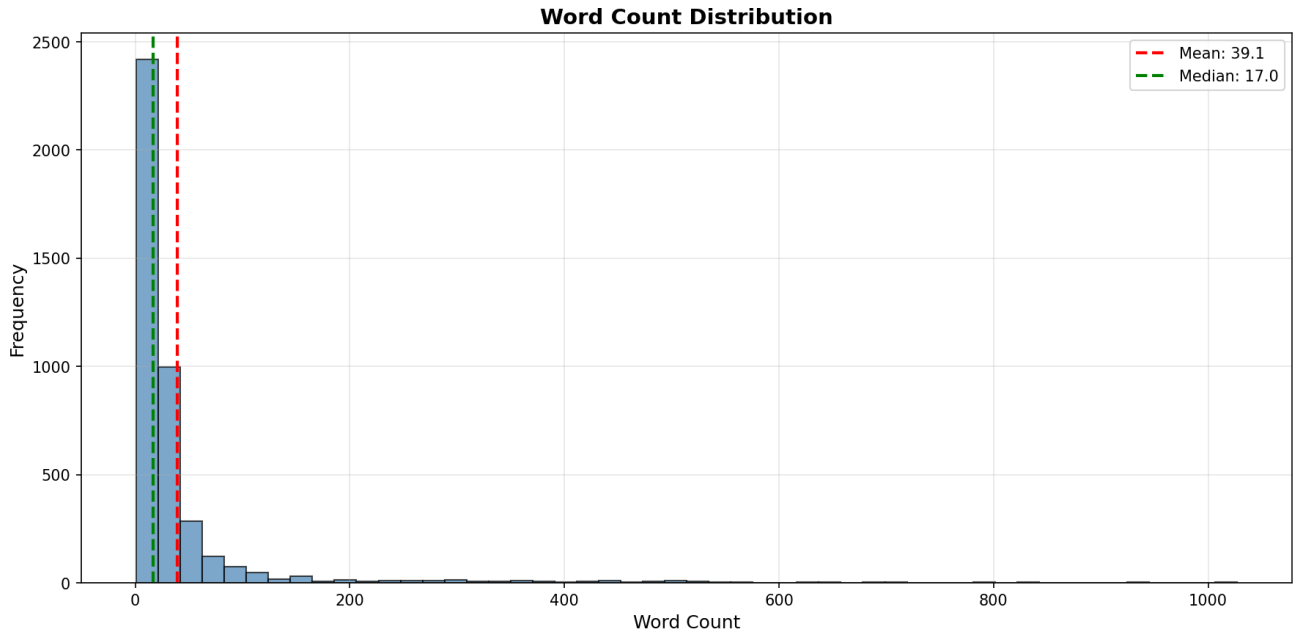

Рис. 4. Распределение количества слов в подмножестве из основного корпуса.

Для устранения этой проблемы была предложена двухступенчатая схема поиска, при которой наряду с усредненным вектором всего фрагмента дополнительно вычисляются векторы для каждого предложения, входящего в его состав.

В рамках эксперимента предложением считался текст, ограниченный знаками препинания (точкой, вопросительным или восклицательным знаком и т. д.). Такой подход позволяет повысить точность ранжирования и локализовать наиболее релевантное предложение в пределах найденного текста, однако все еще подвержен проблеме размывания семантического центра. Схема дополненного алгоритма представлена на рис. 5.

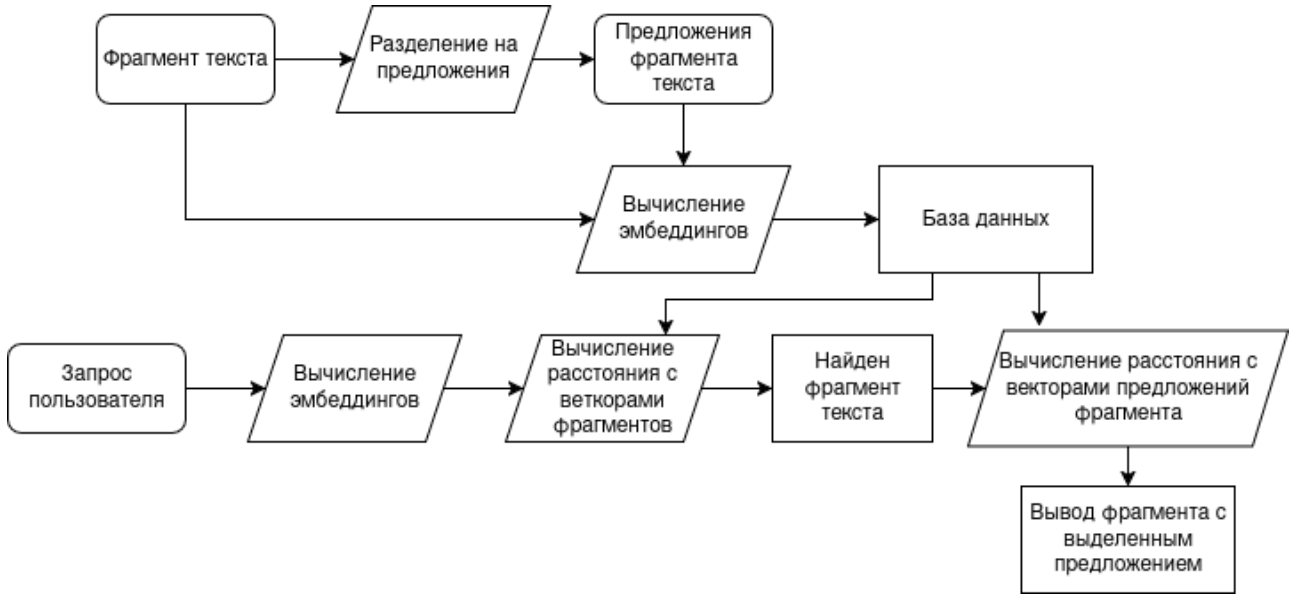

Рис. 5. Схема дополненного алгоритма.

3.4. Результаты анализа алгоритма поиска

Внедрение двухступенчатой системы поиска и ограничения фрагментов до уровня предложения позволило достичь заметного повышения точности ранжирования. В большинстве случаев система корректно выделяла предложение, наиболее близкое по смыслу к поисковому запросу, что обеспечило более интерпретируемые и релевантные результаты. Улучшение качества проявилось прежде всего в способности точнее определять локальные смысловые соответствия. В частности, при поиске по развернутым фразам релевантные тексты стали зани-

мать более высокие позиции в списке результатов, а в пределах найденных документов фокус сместился к предложениям, содержащим ключевые термины запроса.

Примечательно, что в ряде случаев в ранжировании наблюдались совпадения по артефактам, вызванные особенностями корпуса: например, включением отдельных знаков препинания или единичных слов. Подобные элементы могли появляться в результатах поиска из-за высокой частоты встречаемости и статистического сходства коротких векторов.

Качество ранжирования по-прежнему существенно зависит от длины поискового запроса. Более устойчивые результаты достигались при использовании либо частотных терминов корпуса, либо сложных запросов, включающих несколько взаимосвязанных слов. Такая зависимость наиболее выражена для моделей на основе контекстных эмбеддингов, тогда как word2vec демонстрирует большую стабильность при коротких и средних запросах.

Полученные результаты показывают, что использование представлений уровня предложения существенно повышает практическую применимость метода, особенно при работе с корпусами научных текстов, характеризующимися сложной внутренней структурой и разнообразием формулировок.

ЗАКЛЮЧЕНИЕ

Проведенное исследование представляет один из этапов работ в рамках построения библиотеки SciLibRu, направленный на автоматизацию интеграции содержимого научных журналов в ее онтологическую структуру. Разработанный алгоритм трансляции LaTeX-документов в структурированные XML-модели и последующий модуль семантического поиска подтвердили применимость гибридных алгоритмических и нейросетевых подходов для анализа научных текстов.

Полученные результаты, унифицированные наборы структурированных данных, эмбеддинговые представления текстов и прототип поискового механизма создают основу для последующего включения этих данных в граф знаний SciLibRu и их использования при обучении и адаптации языковых моделей LLM к русскоязычным предметным областям.

Развитие описанного направления включает совершенствование механизмов предобработки данных, повышение качества онтологического выравнивания

и разработку рекомендательных систем на основе объединения графа знаний и LLM. Таким образом, работа по структурированию и семантическому поиску журналов является частью комплексной стратегии SciLibRu по созданию интеллектуальной информационной среды для анализа, поиска и извлечения знаний из научных публикаций.

Дальнейшее развитие видится в нескольких направлениях: проведение настройки моделей на специализированных корпусах научных текстов; расширение набора признаков, используемых для ранжирования; а также интеграция механизма автоматического определения типа текстовых фрагментов (аннотация, основное содержание, подпись, теорема, лемма и др.). Такое развитие позволит повысить точность поиска и использовать его в системе навигации по графам знаний научных публикаций, основанных на семантическом описании.

СПИСОК ЛИТЕРАТУРЫ

1. *Hoftich M.* TEX4ht: LATEX to Web Publishing // TUGboat. 2019. Vol. 40. No. 1. P. 76–81.

2. *Frankston C. et al.* Using HTML Papers on arXiv: Why It's Important, and How We Made It Happen // arXiv preprint 2024. https://doi.org/10.48550/arXiv.2402.08954

3. *Серебряков В.А., Галочкин М.П., Гончар Д.Р., Фуругян М.Г.* Теория и реализация языков программирования. 2-е изд. Москва: Изд-во МЗ-Пресс, 2006. 352 с. ISBN 978-5-4488-1013-8

4. *Хопкрофт Дж., Мотвани Р., Ульман Дж.* Введение в теорию автоматов, языков и вычислений. Москва: Изд-во Вильямс, 2002. 528 с. ISBN: 5-8459-0261-4

5. *Ахо А.В., Лам М.С., Сети Р., Ульман Дж.Д.* Компиляторы: принципы, технологии и инструментарий. 2-е изд. М.: Вильямс, 2008. 1184 с.

6. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed Representations of Words and Phrases and their Compositionality // Advances in Neural Information Processing Systems (NIPS 26). 2013. P. 3111–3119. https://dl.acm.org/doi/10.5555/2999792.2999959 (дата обращения: 08.11.2025)

7. *Pennington J., Socher R., Manning C.* GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. P. 1532–1543. https://doi.org/10.3115/v1/D14-1162

8. *Joulin A., Grave E., Bojanowski P., Mikolov T.* Bag of Tricks for Efficient Text Classification // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 2, Short Papers. Valencia, Spain, April 2017. P. 427–431. https://doi.org/10.18653/v1/E17-2068

9. *Feng F., Yang Y., Cer D., Arivazhagan N., Wang W.* Language-agnostic BERT Sentence Embedding // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Dublin, Ireland, May 2022. P. 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

10. *Zmitrovich D. et al.* A Family of Pretrained Transformer Language Models for Russian // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia, May 2024. P. 507–524. https://doi.org/10.48550/arXiv.2309.10931

11. *Kuratov Y., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019” Moscow, May–June 2019 https://doi.org/10.48550/arXiv.1905.07213

12. *Nikolich A., Puchkova A.* Fine-tuning GPT-3 for Russian Text Summarization // arXiv preprint 2021. https://doi.org/10.48550/arXiv.2108.03502

13. *Kutuzov A., Kuzmenko E.* WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // In: Ignatov D. et al. (eds.) Analysis of Images, Social Networks and Texts (AIST 2016). Communications in Computer and Information Science. Vol. 661. Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-52920-2_15

14. *Kasenchak R.T.* What is Semantic Search? and Why Is It Important? // Information Services and Use. 2019. Vol. 39. No. 3. P. 205–213. https://doi.org/10.3233/ISU-190045

15. *Shelke P. et al. A* Systematic and Comparative Analysis of Semantic Search Algorithms // International Journal on Recent and Innovation Trends in Computing and Communication. 2023. Vol. 11, No. 11s. P. 222–229. https://doi.org/10.17762/ijritcc.v11i11s.8094

16. *Weckmüller D., Dunkel A., Burghardt D.* Embedding-Based Multilingual Semantic Search for Geo-Textual Data in Urban Studies // Journal of Geovisualization and Spatial Analysis. 2025. Vol. 9. No.31. P. 1-18. https://doi.org/10.1007/s41651-025-00232-5

17. *Siddharth Pratap Singh*. Vector Search in the Era of Semantic Understanding: A Comprehensive Review of Applications and Implementations // International Journal of Computer Engineering and Technology. 2024. Vol. 15. No. 6. P. 1794–1805. https://doi.org/10.34218/IJCET_15_06_153

18. *Zhou Y. et al.* Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words // 2022. https://doi.org/10.48550/arXiv.2205.05092

19. *Healy J., McInnes L.* Uniform manifold approximation and projection // Nature Reviews Methods Primers. 2024, Vol. 4. No. 82. P. 1–15. https://doi.org/10.1038/s43586-024-00363-x

FORMATION OF STRUCTURED REPRESENTATIONS OF SCIENTIFIC JOURNALS FOR INTEGRATION INTO A KNOWLEDGE GRAPH AND SEMANTIC SEARCH

O. M. Ataeva[1] [0000-0003-0367-5575], M. G. Kobuk[2] [0009-0002-9834-8218]

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

[1, 2]S. Y. Witte University of Moscow, Moscow, Russia

[1]oataeva@frccsc.ru, [2]mikhail.kobuk@mail.ru

Abstract

This paper examines the development of the SciLibRu library of scientific subject areas, as a continuation of the semantic description of scientific works from the library LibMeta project. This library is based on a conceptual data model, the structure and semantics of which are formed based on the principles of ontological modeling. This approach ensures a strict description of the subject area, formalization of the relationships between entities, and the possibility of further automated data analysis. The goal of the study is to develop and experimentally apply methods for structuring scientific journal data in LaTeX format for their integration into the library ontology and to support semantic search.

An algorithm for translating data represented by multiple files into XML format is proposed for integration into the library ontology. A vector search module based on

embedding calculation using language models is implemented. Patterns in the distribution of embeddings and factors influencing the accuracy of search results ranking are identified. Testing of the two components is conducted.

The developed method forms the basis for automatically incorporating scientific journal data into the SciLibRu knowledge graph and creating training corpora for language models limited to scientific subject areas. The obtained results contribute to the development of journal knowledge graph navigation systems, recommendation engines, and intelligent search tools for Russian-language scientific texts.

***Keywords**: semi-structured data, text structuring, LaTeX, vector representations of text, full-text search, semantic search.*

REFERENCES

1. *Hoftich M.* TEX4ht: LATEX to Web Publishing // TUGboat. 2019. Vol. 40, No. 1. P. 76–81.

2. *Frankston C. et al.* Using HTML Papers on arXiv: Why It's Important, and How We Made It Happen // arXiv preprint 2024. https://doi.org/10.48550/arXiv.2402.08954 (In Russ.)

3. *Serebryakov V.A., Galochkin M.P., Gonchar D.R., Furugyan M.G.* Theory and Implementation of Programming Languages. 2nd ed. Moscow: MZ-Press, 2006. 352 p. (In Russ.)

4. *Hopcroft J., Motwani R., Ullman J.* Introduction to Automata Theory, Languages, and Computation. Moscow: Williams, 2002. 528 p. (In Russ.)

5. *Aho A.V., Lam M.S., Sethi R., Ullman J.D.* Compilers: Principles, Techniques, and Tools. 2nd ed. Moscow: Williams, 2008. 1184 p. (In Russ.)

6. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J*. Distributed Representations of Words and Phrases and their Compositionality // Advances in Neural Information Processing Systems (NIPS 26). 2013. P. 3111–3119. https://dl.acm.org/doi/10.5555/2999792.2999959 (date accessed: 08.11.2025)

7. *Pennington J., Socher R., Manning C.* GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. P. 1532–1543. https://doi.org/10.3115/v1/D14-1162

8. *Joulin A., Grave E., Bojanowski P., Mikolov T.* Bag of Tricks for Efficient Text Classification // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia, Spain, April 2017. P. 427–431. https://doi.org/10.18653/v1/E17-2068

9. *Feng F., Yang Y., Cer D., Arivazhagan N., Wang W.* Language-agnostic BERT Sentence Embedding // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Dublin, Ireland, May 2022. P. 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

10. *Zmitrovich D. et al.* A Family of Pretrained Transformer Language Models for Russian // arXiv preprint 2023. https://doi.org/10.48550/arXiv.2309.10931

11. *Kuratov Y., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // arXiv preprint 2019. https://doi.org/10.48550/arXiv.1905.07213

12. *Nikolich A., Puchkova A.* Fine-tuning GPT-3 for Russian Text Summarization // arXiv preprint 2021. https://doi.org/10.48550/arXiv.2108.03502

13. *Kutuzov A., Kuzmenko E.* WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // In: Ignatov D. et al. (Eds.) Analysis of Images, Social Networks and Texts (AIST 2016). Communications in Computer and Information Science. Vol. 661. Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-52920-2_15

14. *Kasenchak R.T.* What is Semantic Search? and Why Is It Important? // Information Services and Use. 2019. Vol. 39. No. 3. P. 205–213. https://doi.org/10.3233/ISU-190045

15. *Shelke P. et al. A* Systematic and Comparative Analysis of Semantic Search Algorithms // International Journal on Recent and Innovation Trends in Computing and Communication. 2023. Vol. 11, No. 11s. P. 222–229. https://doi.org/10.17762/ijritcc.v11i11s.8094

16. *Weckmüller D., Dunkel A., Burghardt D.* Embedding-Based Multilingual Semantic Search for Geo-Textual Data in Urban Studies // Journal of Geovisualization and Spatial Analysis. 2025. Vol. 9. No. 31. P. 1–18. https://doi.org/10.1007/s41651-025-00232-5

17. *Siddharth Pratap Singh*. Vector Search in the Era of Semantic Understanding: A Comprehensive Review of Applications and Implementations // International

Journal of Computer Engineering and Technology. 2024. Vol. 15. No. 6. P. 1794–1805. https://doi.org/10.34218/IJCET_15_06_153

18. *Zhou Y. et al.* Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words // 2022. https://doi.org/10.48550/arXiv.2205.05092

19. *Healy J., McInnes L.* Uniform manifold approximation and projection // Nature Reviews Methods Primers. 2024, Vol. 4. No. 82. P. 1–15. https://doi.org/10.1038/s43586-024-00363-x

СВЕДЕНИЯ ОБ АВТОРАХ

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, 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

КОБУК Михаил Геннадьевич – студент бакалавриата Московского университета имени С.Ю. Витте, область интересов NLP, анализ данных, системное программирование.

Mikhail Gennadievich KOBUK – bachelor at S. Witte University of Moscow. Research interests: NLP, data analysis, system programming.

email: mikhail.kobuk@mail.ru

ORCID*:* 0009-0002-9834-8218

Материал поступил в редакцию 8 ноября 2025 года