



Рекомендательная система на основе обобщённого указателя журналов

© 2025, О.М. Атаева¹, Н.П. Тучкова¹✉, А.Г. Дегтев²

¹Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), Москва, Россия

²Московский физико-технический институт (национальный исследовательский университет), Долгопрудный, Московская область, Россия

Аннотация

Рассматривается тематическая классификация журналов на примере «Белого списка» – Единого государственного перечня научных журналов. Ставится задача автоматизации анализа тематического направления журналов. Используется Государственный рубрикатор научно-технической информации (ГРНТИ), классификаторы: Универсальная десятичная классификация (УДК), *Mathematics Subject Classification (MSC)* и др., а также онтология семантической библиотеки предметных областей *SciLibRu*. На основе данных о журналах «Белого списка» и источников в библиотеке *SciLibRu* составляется обобщённый указатель, который включается в граф знаний *SciLibRu*. Пользователи библиотеки *SciLibRu* получают возможность навигации по разным аспектам информации о журналах (тематика, категория и др.), что упрощает выбор журнала для возможной публикации. Приводится пример поиска журнала, основанный на семантическом анализе статьи для определения её тематической принадлежности к предметной области журнала из «Белого списка». Сформированный обобщённый указатель в библиотеке *SciLibRu* позволяет задавать на естественном языке запросы, связанные с выбором издания для публикации. Предложенная методология может быть распространена на другие предметные области (данные о конференциях и др.). Практическая значимость исследования состоит в автоматизации подбора тематики журнала для подготовленной научной статьи.

Ключевые слова: белый список журналов, онтология предметной области, рекомендательная система, классификатор, обобщённый указатель журналов, семантическая библиотека.

Цитирование: Атаева О.М., Тучкова Н.П., Дегтев А.Г. Рекомендательная система на основе обобщённого указателя журналов. *Онтология проектирования*. 2025. Т.15, №4(58). С.598-613. DOI: 10.18287/2223-9537-2025-15-4-598-613.

Финансирование: работа выполнена в рамках госзадания FFNG-2024-0003 по теме «Математические методы анализа данных и прогнозирования».

Вклад авторов: Атаева О.М. – обработка данных, подготовка примеров; Тучкова Н.П. – разработка структуры статьи, анализ источников; Дегтев А.Г. – подготовка примеров.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Тематический анализ научных публикаций – актуальное направление исследований, поскольку для составления обзоров и рецензий необходимо просматривать множество публикаций, чтобы найти тематически близкие. В связи с высокими требованиями к публикациям и конкуренцией в науке необходимо учитывать статус научных журналов (квартиль и другие наукометрические показатели). Актуальными становятся публикации в журналах Белого списка (БС), который планируется обновлять в реальном времени.

В данной работе предложен подход к *семантическому описанию* научных журналов БС и включение этих описаний в онтологию библиотеки *SciLibRu* для представления необходимой

информации в виде графа знаний (ГЗ) [1], где узлы представляют различные точки входа в указатель (логическую структуру) по этим данным. На базе объединённых данных формируется обобщённый (тематический) указатель (ОУ) журналов, который используется в качестве базы знаний (БЗ) рекомендательной системы. Данный подход позволяет определить с помощью навигации по ГЗ к каким предметным областям (ПрО) относится содержание научной работы и получить *список рекомендуемых журналов* из БС, наиболее близких по тематике этой работы. На основе предложенного ОУ разработана рекомендательная система, интегрированная в библиотеку *SciLibRu* и предоставляющая функцию подбора журнала, тематика которого наиболее соответствует содержанию рассматриваемой статьи. Практическая ценность предлагаемого решения заключается в упрощении и ускорении процесса выбора журнала для публикации за счёт использования знаний в ГЗ *SciLibRu*.

1 Постановка задачи

Поиск публикаций отличается от поиска журналов по многим признакам. Для поиска публикаций в библиографических базах и в библиотеках предлагается поиск по автору, названию, ключевым словам, индексу классификатора. Применение методов искусственного интеллекта расширило возможности поиска публикаций по тематическим признакам, признакам схожести, цитированию и другим связям [2, 3]. Поиск журналов связан с задачей выбора журнала для возможной публикации, который подходит тематически и соответствует некоторым показателям (специальностям Высшей аттестационной комиссии (ВАК) и др.). Тематика, квартиль, условия публикации, учитываются авторами в соответствии с индивидуальными потребностями. Классификацию по этим признакам можно сделать объективно на основе открытых данных, используя известные библиографические ресурсы (*scopus.com*, *www.webofscience.com*, *elibrary.ru*, *zbmath.org*, *mathnet.ru*, *cyberleninka.ru* и др.). В открытых и коммерческих разработках поля для поиска остаются такими же, как и для поиска публикаций: ключевые слова, названия, авторы, аннотации и полные тексты. Практика использования этих полей поиска без применения средств семантического анализа приводит к шуму в поисковой выдаче. Некоторые примеры таких ресурсов приведены в таблице 1.

Таблица 1 – Ресурсы тематического поиска журналов для публикации рукописи

№	Название	Тематический поиск	Доступ	Поля для поиска
1	<i>Master Journal List</i>	есть для английского языка / нет для РФ	платный	ISSN, название для журналов из коллекции <i>WoS</i>
2	<i>Elsevier Journal Finder</i>	есть для английского языка	бесплатный	данные рукописи (аннотация, ключевые слова, цель исследований), название журнала <i>Elsevier</i>
3	<i>Springer Journal Suggester</i>	есть для английского и немецкого языка	бесплатный	данные рукописи (название, аннотация, ключевые слова), название журнала <i>Springer</i>
4	<i>Wiley Journal Finder</i>	есть для английского языка	бесплатный	данные рукописи (название, аннотация) название журнала <i>Wiley</i>
5	<i>Web of Science Master Journal List</i>	есть для английского языка	бесплатный	данные рукописи (название, аннотация) название журнала из <i>WoS</i>
6	<i>Scopus</i>	есть для английского языка	бесплатный	выбор из тематического списка <i>Scopus</i> по типу изданий и квартилю
7	<i>Math-Net.Ru</i>	нет	бесплатный	название журнала из коллекции <i>Math-Net</i> по алфавиту или издательству
8	<i>eLibrary</i>	есть	бесплатный	тематика ГРНТИ, поля БС, <i>WoS</i> , <i>Scopus</i> , РИНЦ, а также квартиль, уровень, категория

После внедрения нейросети *SciRus-tiny* (разработана для семантического анализа научных текстов) [4] в библиотеке *eLibrary* (<https://www.elibrary.ru>) используется тематический поиск по верхнему уровню Государственного рубрикатора научно-технической информации (ГРНТИ). Журнал может быть отнесён к нескольким рубрикам. Отдельно в списке рубрик выделены мультидисциплинарные журналы. Внутри тематической рубрики можно указать поля БС, *WoS*, *Scopus*, РИНЦ, а также квантиль, уровень, категорию.

Из таблицы 1 видно, что библиотека *eLibrary* предоставляет поисковые поля для выбора журнала по тематике и другим признакам, а также сервис для сравнения журналов по показателям из списка *eLibrary*, но нет поиска журнала для предполагаемой публикации.

В данной статье предлагается семантическое описание журналов БС включить в онтологию и ГЗ библиотеки ПрО *SciLibRu* и создать набор данных с тематическим разбиением. Поиск по журналам можно провести как навигацию по ГЗ с целью подбора журнала, ПрО которого наиболее близка для предполагаемой публикации (рисунок 1). Для этого предлагается применить большие языковые модели (БЯМ) для обращения к ГЗ БС.

Для формирования ОУ сведения о журналах БС разделяются на тематические кластеры, при этом каждый журнал связывается с классификаторами: Универсальной десятичной классификации (УДК), *Mathematics Subject Classification (MSC)*, БСК и онтологией *SciLibRu*. ОУ представлен в виде ГЗ, где узлы ГЗ – это точки входа ОУ. Для такого разделения используются алгоритмы автоматической классификации объектов (ААКО), которые обучаются на данных БС и *SciLibRu*.

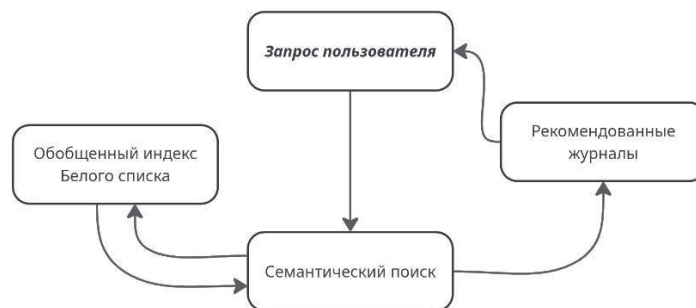


Рисунок 1 – Схема подбора журнала из *SciLibRu* для предполагаемой публикации

2 Семантическое описание журналов и их интеграция в онтологии *SciLibRu*

2.1 Библиотека *SciLibRu*

Для интеграции семантических образов журналов применяются методы онтологического проектирования [5-8] и их представление в виде ГЗ в библиотеке *SciLibRu* (ранее – проект *LibMeta*) [9], в которой используется технология описания ПрО научных журналов [10]. В библиотеке содержатся энциклопедии, тезаурусы, классификаторы и др., семантически связанные в *SciLibRu* [11]. На рисунке 2 показан пример поэтапного расширения ПрО «Математика» в *LibMeta*, когда в библиотеку добавляются данные об источниках (статьях журнала «Механика композиционных материалов и конструкций» (МКМК), <https://iampress.ru/>).

На разных этапах добавляются тематические разделы, термины и данные в виде публикаций. Онтологическое описание этих источников образует единый ГЗ, где понятия и объекты из разных источников семантически связаны между собой [9-12].

Известные системы рекомендаций включают совместную фильтрацию и рекомендации на основе содержания [13, 14] с использованием средств коммуникации с объектами, предназначенными для выбора (списки, изображения и т.д.). В данной работе выбор предлагается осуществлять с помощью навигации по ГЗ БС, а результат представляется в виде списка с обоснованием рекомендаций (рисунок 3).

Этап 1 Без классификации по подразделам	Предметная область "Математика"				
	Тезаурус "Математическая энциклопедия"				
Этап 2 Классификация и выделение подразделов	Классификаторы				
	MSC		УДК		ГРНТИ
	Подпространства предметной области "Математика"				
	Дифференциальные уравнения	Математическая физика	Вычислительная математика	Математическое моделирование	...
Этап 3 Выделение терминов для подразделов	Тезаурусы и словари подпространств предметной области "Математика"				
	Тезаурус ОДУ	Словарь спец функций	Словарь уравнений смешанного типа	Словарь механики сплошной среды	Словарь механики композитов
	Словарь фуллерены	Словарь наннотрубки	Словарь полимерных композитов		...
	Источники данных				
Этап 4 Классификация источников данных и обогащение описания предметной области	Неструктурированные			Структурированные	
	Журналы	Публикации	...	Репозитории	Mathnet

Рисунок 2 - Этапы расширения предметной области «Математика» на примере журнала МКМК (ОДУ - обыкновенные дифференциальные уравнения, Mathnet - Общероссийский портал Math-Net.Ru)

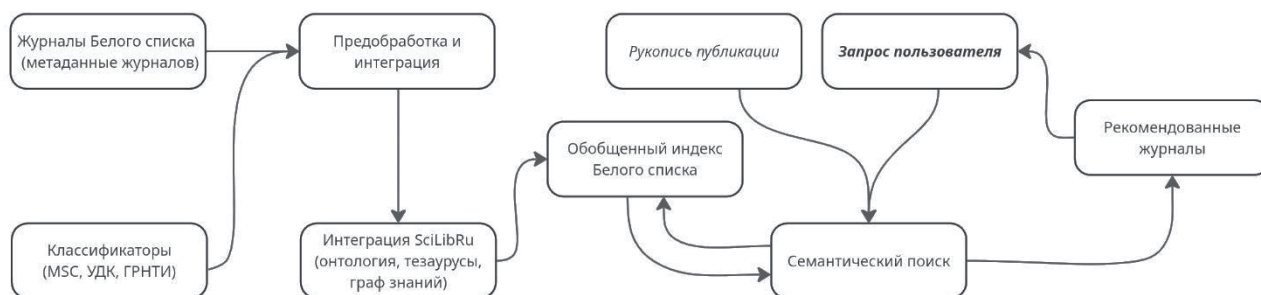


Рисунок 3 - Схема предлагаемого метода: интеграция данных БС и источников знаний в ОУ на базе ГЗ SciLibRu и использование ОУ для рекомендации журналов по пользовательскому запросу

Семантические описания журналов БС интегрированы в онтологию *SciLibRu*, встроены в ГЗ *SciLibRu* после предобработки и распределены по ПрО. Рукопись статьи проходит предобработку (семантическое сжатие текста) для выявления структуры текста, ключевых слов и связей с онтологией *SciLibRu*. Полученные связи позволяют обратиться к ГЗ *SciLibRu* и затем к ГЗ БС, чтобы получить список рекомендаций, близких к тематике рукописи журналов.

2.2 Интеграция семантических образов журналов в библиотеке *SciLibRu*

Интеграция БС в *SciLibRu* включает *предобработку данных* и *доставление онтологии SciLibRu* и ГЗ *SciLibRu*. Для построения ОУ научных журналов БС в *SciLibRu* используются сведения из открытых массивов научных данных: систем *OpenAlex* (данные на английском языке) и портала ВИНТИ (данные русским языком):

- метаданные журналов БС, которые содержат идентификаторы ISSN, ключевые слова (понятия ПрО) и наукометрические данные журналов (из *OpenAlex*);
- данные иерархических классификаторов (ГРНТИ, УДК, *MSC*), а также таблицы соответствия между ними (дополнительные сопоставления ВИНТИ).

Набор данных ГРНТИ использовался на русском и английском языках с автоматическим переводом. Кроме названия журнала в источниках представлены короткие описания основных понятий, соответствующих тематикам, освещаемым в журнале.

2.2.1 Предобработка журналов Белого списка и данных ГРНТИ

Использованные сведения о журналах БС содержали около 29000 наименований журналов с базовыми метаданными (ISSN, издатель, категория и др.) и списками ключевых понятий (в сумме более 500.000 ключевых слов на английском языке). Эти списки были очищены и унифицированы (нормализованы лексически, переведены на единый язык) в процессе предобработки. В результате сформирована обучающая выборка для тематической классификации журналов. Предобработка данных включает следующие шаги.

- *Лингвистическая нормализация.* Все рубрики и ключевые слова ГРНТИ переведены на английский язык, чтобы сформировать единый одноязычный набор данных для обучения модели классификации. Оригинальные данные БС на английском языке дополнены переводами рубрик ГРНТИ для сопоставления.
- *Структурирование классификаторов.* Отраслевые классификаторы (ГРНТИ, УДК, MSC) представлены в виде иерархий с определённой структурой кодов. Использованы таблицы соответствия между классификаторами (например, какие коды УДК соответствуют рубрикам ГРНТИ), в т.ч. взятые из открытых источников (ВИНИТИ).
- *Формирование обучающего набора.* Для экспериментов по классификации журналов по темам выбраны 64 рубрики верхнего уровня ГРНТИ, каждая из которых рассматривается как класс (категория).
- *Составление текстового описания рубрик.* Для каждой рубрики верхнего уровня ГРНТИ собрано множество текстовых данных, описывающих эту рубрику. В это множество вошли названия и описания рубрик, а также тексты подразделов второго и третьего уровней ГРНТИ, связанные с данной рубрикой. Для удобства изложения эти тексты названы «ключевые фразы рубрики», полагая, что совокупность таких фраз описывает соответствующую область знаний (тематику рубрики журнала).
- *Дополнение пропусков.* Поскольку классификатор ГРНТИ заполнен неравномерно, для заполнения «пустых» или слабо описанных рубрик использовались данные из описаний соответствующих разделов УДК. Информация о соответствии рубрик ГРНТИ и УДК составлена на основе открытых данных ВИНИТИ, а также использованы ранее установленные в *LibMeta* связи между рубриками различных классификаторов [12].

В результате получен *итоговый набор данных*, включающий примерно 8000 ключевых фраз для 64 тематических рубрик верхнего уровня ГРНТИ. Этот корпус текстов стал основой для обучения модели автоматической классификации объектов (журналов) по тематикам.

На рисунке 4 приведена иллюстрация *распределения количества ключевых фраз* по рубрикам верхнего уровня ГРНТИ (каждый столбец гистограммы соответствует одной рубрике ГРНТИ). Видно, что класс распределения ключевых фраз неравномерен – присутствует *дисбаланс классов*. В данной задаче это проявляется в склонности модели «игнорировать» малочисленные классы и чаще предсказывать рубрики, для которых в обучении было много примеров. Чтобы компенсировать этот эффект, применено *взвешивание классов* [15-17] при обучении моделей: меньшим по объёму классам назначаются повышенные веса ошибки.

В результате предобработки сформирован набор данных, который был использован ААКО тематической классификации журналов БС по рубрикам ГРНТИ верхнего уровня.

2.2.2 Интеграция в граф знаний LibMeta

Согласно [10] для каждого нового источника БС данных создаётся модель ГЗ источника в виде онтологии, включающая журналы, как узлы, и их связи. Онтологическая модель *SciLibRu* *достроена*: введён новый тип объектов «Журнал» и определены типы связей для этого типа объектов.

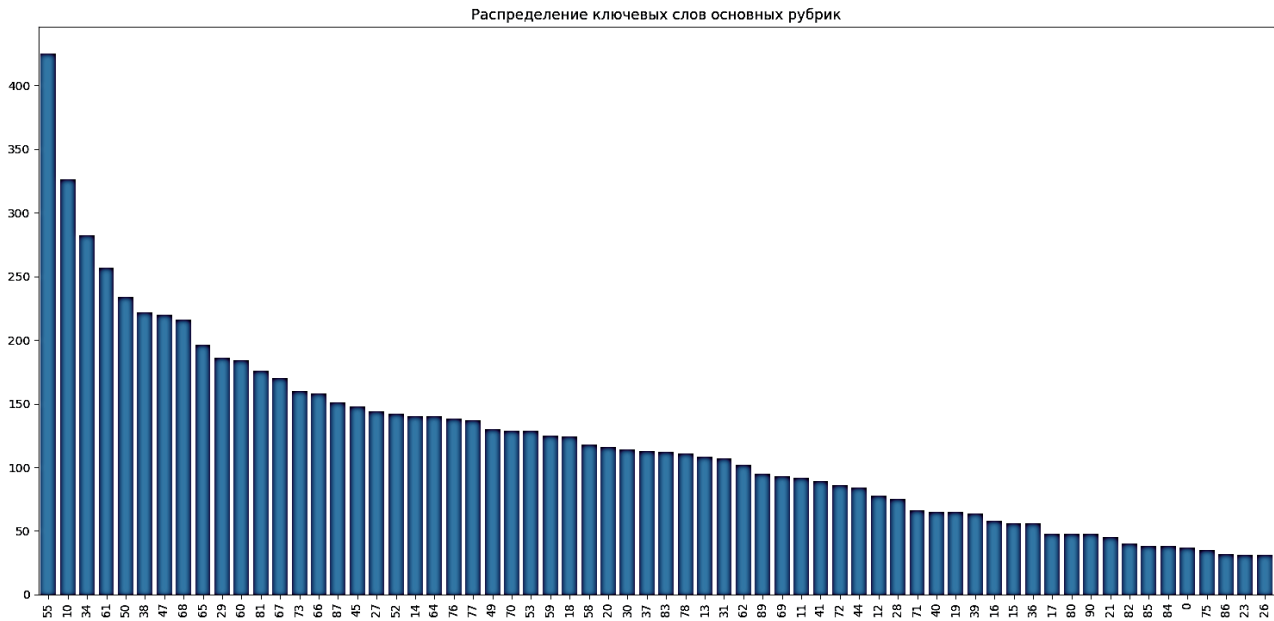


Рисунок 4 – Гистограмма распределения ключевых фраз по рубрикам верхнего уровня ГРНТИ

Для построения ГЗ БС формируются основные типы семантических связей:

- **Журнал ↔ Классификатор** – связь журнала с рубриками или категориями из внешних классификаторов (например, рубрика ГРНТИ, код УДК, или из списка ВАК);
- **Классификатор ↔ Классификатор** – взаимосвязи между классификационными системами (например, соответствие между рубрикой ГРНТИ и кодом MSC);
- **Ключевая фраза ↔ Журнал** – связь ключевого понятия (термина) с журналом, в котором это понятие присутствует как часть описания тематической области;
- **Ключевая фраза ↔ Классификатор** – связь понятий из онтологии *SciLibRu* с соответствующими рубриками классификаторов (например, термин из тезауруса, описывающий рубрику ГРНТИ);
- **Журнал ↔ Публикация** – связь между журналом и публикациями (статьями), включёнными в онтологию (данные о статьях из некоторых журналов загружены ранее [11]);
- **Публикация ↔ Ключевая фраза** – связь научной статьи с термином или ключевой фразой, обозначающей её тему (устанавливается при семантическом анализе текста статьи).

В *LibMeta* перечисленные связи были реализованы добавлением соответствующих *свойств и классов*. Таким образом, каждый журнал БС в ГЗ *SciLibRu* получил семантическое описание, включающее набор связей, отражающих все доступные тематические и предметные сведения о журнале: его рубрики ГРНТИ (по результатам классификации), связанные коды УДК и MSC (через сопоставления, установленные в *LibMeta* [13]), связи с существующими публикациями в *SciLibRu* (если таковые имеются по тем же тематикам), а также связи с терминологией библиотеки (ключевые слова из тезаурусов, энциклопедий и пр., если они совпадают или близки по смыслу к имеющимся в описании журнала).

Пример 1 показывает структуру фрагмента графа знаний БС и его связей в формате *RDF*, где журналы связаны с кодами ГРНТИ и *MSC* через свойства *hasClassification* и *hasMapping* (выполняется семантическое связывание данных при интеграции БС в *SciLibRu*).

```
Journal hasClassification GRNTI
GRNTI hasMapping MSC
Journal hasKeyword Keyword
Journal publishedIn Publisher
```

```

libmeta:journal/12345 rdf:type libmeta:Journal
libmeta:journal/12345 libmeta:issn "2313-1039"
libmeta:journal/12345 libmeta:title "Онтология проектирования"@ru
libmeta:journal/12345 libmeta:hasClassification libmeta:MSC_68
libmeta:journal/12345 libmeta:hasClassification libmeta:UDC_004
libmeta:UDC_004 skos:prefLabel "Информационные технологии..."@ru
libmeta:MSC_68 skos:prefLabel "Computer science"@en

```

Пример 2 показывает *SPARQL*-запрос для выборки журналов по теме «Информационные технологии» и связанных кодов *MSC* (ГЗ *SciLibRu* поддерживает тематический поиск через формальные семантические запросы).

```

SELECT ?journal ?title ?msc
WHERE {
  ?journal a libmeta:Journal ;
    libmeta:title ?title ;
    libmeta:hasClassification ?grnti .
  ?grnti skos:prefLabel "Информационные технологии"@ru ;
    libmeta:hasMapping ?msc .
  ?msc skos:prefLabel ?mscLabel .
}

```

Пример 3 показывает *SPARQL*-запрос, позволяющий находить журналы, одновременно имеющие связи с ГРНТИ и *MSC*, что обеспечивает перекрёстную навигацию между различными классификационными системами.

```

SELECT ?journal ?grnti_code ?msc_code
WHERE {
  ?journal a libmeta:Journal ;
    libmeta:hasClassification ?grnti, ?msc .
  ?grnti a libmeta:GRNTI .
  ?msc a libmeta:MSC .
  ?grnti libmeta:code ?grnti_code .
  ?msc libmeta:code ?msc_code .
}

```

Пример 4 показывает запрос на извлечение всех ключевых слов журналов, относящихся к конкретной рубрике ГРНТИ (в *SciLibRu* можно использовать ГЗ для анализа семантического поля журналов и расширения онтологических связей).

```

SELECT ?journal ?keyword
WHERE {
  ?journal a libmeta:Journal ;
    libmeta:hasClassification libmeta:GRNTI_27.35 ;
    libmeta:hasKeyword ?keyword .
}

```

Установление связей между классификаторами и терминами выполнено автоматически (на основе подготовленных таблиц соответствий и алгоритмического поиска совпадений) и с привлечением экспертов на этапе наполнения ГЗ для проверки и добавления недостающих связей. В результате получен ОУ, представленный в виде фрагмента ГЗ *SciLibRu*. Этот ОУ включает узлы «Журнал», связанные множеством семантических отношений с узлами «Рубрика ГРНТИ», «Код УДК», «Код *MSC*», «Термин ПрО» и др. ОУ является единым указателем по научным журналам, где каждая «точка входа» – это некоторый объект или признак, по которому можно производить поиск и навигацию (тематика, категория, ключевое слово, ISSN, квартиль и т.п.).

3 Обобщённый указатель для рекомендательной системы

Для формирования ОУ массив сведений о журналах разделён на тематические кластеры, при этом каждый журнал связан с классификаторами УДК, MSC и онтологией SciLibRu.

3.1 Методы

Для решения задачи классификации по темам (отнесение журнала к одной или нескольким рубрикам верхнего уровня ГРНТИ) применены алгоритмы машинного обучения [18, 19]. Векторизация входных текстовых данных (описаний рубрик и журналов) проведена с использованием моделей «мешок слов» и *TF-IDF*, распределённых семантических представлений (эмбедингов) слов и текстов – модели *Word2Vec*, *FastText* – и специализированной модели *SciRus-tiny* [4]. *SciRus-tiny* является нейросетевой моделью, обученной на корпусе научных текстов, и хорошо учитывает специфику научной лексики [20–22].

В результате экспериментов выбрана наилучшая комбинация метода векторизации [23] и классификатора по критерию качества классификации. В таблице 2 приведены сводные показатели качества (*F1*-мера) для 10 моделей-классификаторов в сочетании с тремя методами векторизации. Для сравнения приведены результаты базовой модели (*DummyClassifier*), которая всегда предсказывает наиболее частый класс.

Наивысшее качество классификации на тестовой выборке показала модель логистической регрессии [24] при использовании эмбедингов *SciRus-tiny* ($F1 \approx 0.84$). Эксперименты показали, что на новых данных БС векторы *SciRus*

Таблица 2 – Результаты применения алгоритмов автоматической классификации объектов (ААКО)

№	Модель ААКО	<i>Word2Vec</i>	<i>Fasttext</i>	<i>SciRus</i>
1	<i>DummyClassifier</i>	0.05	0.05	0.05
2	<i>KNN</i>	0.35	0.25	0.70
3	<i>SVC</i>	0.33	0.23	0.81
4	<i>DecisionTree</i>	0.42	0.31	0.55
5	<i>ExtraTreesClassifier</i>	0.51	0.41	0.78
6	<i>RandomForest</i>	0.52	0.42	0.78
7	<i>LogisticRegression</i>	0.40	0.38	0.84
8	<i>LGBMClassifier</i>	0.52	0.46	0.79
9	<i>XGBClassifier</i>	0.52	0.44	0.76
10	<i>CatBoostClassifier</i>	0.37	0.28	0.63

лучше «улавливают» смысл и часто соотносят запрос с подходящей рубрикой. В описаниях журналов встречается много новых (не представленных в обучении) ключевых фраз, которые не всегда правильно соотнесены с нужной рубрикой [25, 26]. Эмбединги журналов формируются на основе названий, аннотаций и списков ключевых слов, связанных с каждым журналом, а эмбединги рубрик ГРНТИ строятся по множеству *ключевых фраз* соответствующих рубрик, включающих общепринятые формулировки, расширенные пояснения, ассоциированные ключевые слова, переводные аналоги, которые обогащают исходный набор данных. Т.е. эти эмбединги строятся на разных исходных данных.

Чтобы оценить вклад этапа обогащения данных, проведён эксперимент, в котором выполнено сравнение качества классификации для двух вариантов подготовки данных: без обогащения, используя только исходные описания рубрик ГРНТИ; с обогащением, используя описания, дополненные данными из УДК и терминологией *SciLibRu*. Результаты сравнения приведены в таблице 3, где показаны основные метрики на тестовом наборе: точность, полнота и *F1*-мера. Видно, что без обогащения данных модель не смогла классифицировать некоторые рубрики из-за отсутствия по ним информации (только 58 из 64 рубрик имели обучающие данные). Это привело к низкой полноте (не предсказывались «пустые» классы) и снижению усреднённых метрик. После обогащения все 64 рубрики получили описания; ка-

чество классификации существенно улучшилось – выросла полнота по редким рубрикам, увеличилась средняя $F1$.

Обучение модели логистической регрессии на описанном наборе (~8000 текстов, размерность эмбединга 300) заняло нескольких минут. Объём данных сравнительно невелик, поэтому затраты памяти и времени не стали ограничивающим фактором. Формирование эмбедингов *SciRus-tiny* для всех ключевых фраз и описаний журналов выполнялась примерно 15–20 минут. Таким образом, разработанный подход к классификации может быть масштабирован на большее число классов или документов, не требуя значительных вычислительных ресурсов.

Таблица 3 – Влияние обогащения данных на качество тематической классификации

Модель <i>SciRus-tiny</i>	Доля рубрик с данными	Точность	Полнота	$F1$
Логистическая регрессия – без обогащения	58 из 64	0.80	0.60	0.68
Логистическая регрессия – с обогащением	64 из 64	0.90	0.85	0.88

3.2 Граф знаний и обобщённый указатель журналов Белого списка

В результате исследований получено тематическое разбиение журналов БС по рубрикам ГРНТИ. При моделировании учтены наукометрические показатели, указанные в описаниях БС. В библиотеке *SciLibRu* создан набор данных из описаний журналов, снабжённых связями с классификаторами, рубрикаторами и ПрО онтологии *SciLibRu*, по которому осуществляется навигация на основе представления в виде ГЗ журналов БС. Ключевые слова из аннотаций журналов дополнены ключевых слов из соответствующих статей энциклопедий, тезаурусов и другого содержания *SciLibRu*. Это позволило сделать тематическое разбиение БС более детальным. Это позволяет создать ОУ журналов с навигацией через узлы ГЗ журналов БС, которые указывают на тематические и наукометрические показатели журнала.

На рисунке 5 представлен пример сведений о журнале из БС, загруженном в библиотеку *SciLibRu*, с установленными семантическими связями, которые были выявлены в процессе семантического анализа.

ОУ включает связь журнала с рубрикой, полученной на основе обученной модели, и представляет собой часть ГЗ журналов БС, используемую на этапе рекомендаций.

Предлагаемый подход позволяет выполнять семантический поиск по ГЗ журналов и находить подходящие журналы, даже если в запросе пользователя не указаны точно те же слова, что и в описании журналов, поскольку ГЗ через связи и эмбединги найдёт близкие по смыслу совпадения.

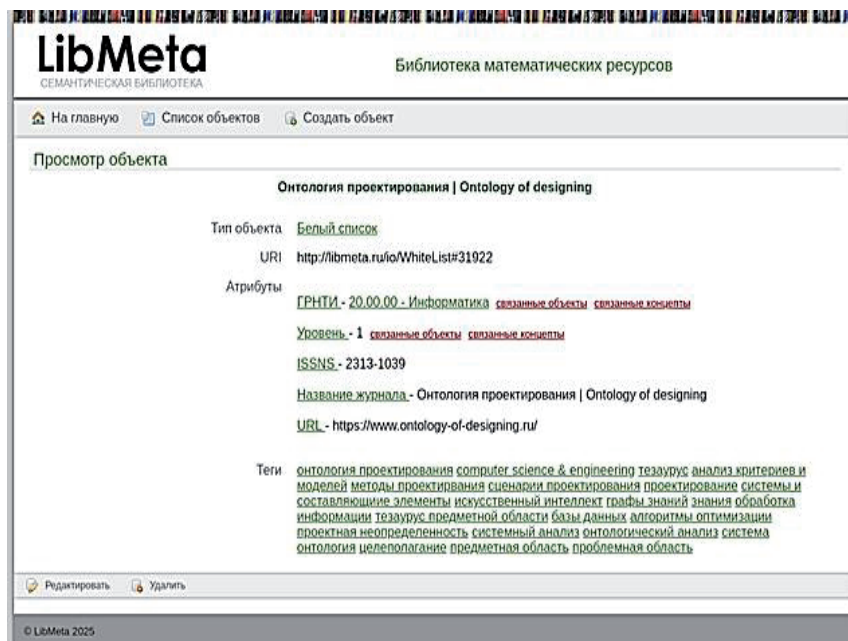


Рисунок 5 - Сведения о загруженном в библиотеку журнале Белого списка со связями

3.3 Пример использования обобщённого указателя

1) *Пользовательский запрос*. Пользователь формулирует запрос, содержащий предполагаемую тему научной работы (публикации), в свободной форме. Это может быть список ключевых слов или краткое описание. Цель запроса – определить, в каком журнале можно опубликовать (или найти для чтения) работу по данной теме. Например: «Где опубликовать работу на тему дифференциальные уравнения в приложениях?» или «Где искать публикации по теме дифференциальные уравнения в приложениях?» Такие запросы отражают ключевые понятия («дифференциальные уравнения», «приложения»), но не указывают прямо рубрику или название журнала.

2) *Семантический анализ запроса*. Поступивший запрос *очищается и нормализуется* (приводится к стандартной форме, устраняются стоп-слова, выполняется лемматизация ключевых терминов) и преобразуется в эмбединг тем же способом, который использовался при формировании эмбедингов описаний журналов и рубрик. В результате получается вектор, характеризующий семантическое содержание запроса.

3) *Поиск по семантической близости*. Выполняется основная часть рекомендации – сопоставление запроса с узлами ГЗ БС:

- вычисляется косинусное сходство между эмбедингом запроса и эмбедингами всех рубрик, осуществляется *сопоставление*, какой рубрике наиболее соответствует запрос;
- выбирается несколько рубрик с наибольшими значениями сходства (например, три наиболее близкие рубрики).

По каждой из выбранных рубрик из ГЗ БС извлекаются связанные журналы – те, которые были отнесены к данной рубрике (также можно добавить журналы, для которых косинусное сходство их эмбединга с эмбедингом запроса превышает заданный порог: если рубрика не совпала, это позволяет учесть случаи, когда журнал связан с близкой, но не точно совпадающей рубрикой).

4) *Формирование результата*. Пользователю отображается список найденных рубрик и соответствующих журналов, где заявленная тема представлена (или близка по смыслу). По каждому рекомендуемому журналу приводится основная информация: название, ISSN, тематические рубрики, а также ссылка на профиль журнала (в библиотеке *SciLibRu* или внешней системе, напр. *eLibrary*). *Обосновывается рекомендация*, т.е. указывается, почему этот журнал предложен (например, совпадение ключевых слов, высокая семантическая близость).

Использование ОУ позволяет найти информацию о подходящих журналах, даже если в запросе пользователя не упоминаются те же самые слова, что были использованы в описании журналов. Например, на запрос: «В каком журнале опубликовать статью с ключевыми словами *ontology design, semantic relationships, subject ontology, vector algorithms, knowledge graph*» пользователь получает в качестве ответа рубрику ГРНТИ «20.00.00 – Информатика» и перечень журналов данной тематики.

В случае ГЗ БС, интегрированного в библиотеку *SciLibRu*, язык запросов *SPARQL* используется для извлечения информации из графового представления данных (он представляет сложности для неподготовленного пользователя). Запрос на естественном языке составляет необходимое условие работы с данными ПрО. В [27] рассмотрено применение БЯМ к работе с ПрО математики в *SciLibRu* для автоматического формирования *SPARQL* запросов.

Можно показать, как на основе интеграции данных БС, ГРНТИ, УДК, *OpenAleph* в библиотеке *SciLibRu* выполнить поиск близкого по тематике журнала для возможной публикации рукописи на естественном языке. С этой целью проведено обогащение БС ключевыми словами (общая терминология, направления деятельности). Сформировано векторное представление обогащённого списка с использованием модели *SciRus*.

На рисунке 6 приведён пример визуализации запроса и ответов в векторном пространстве с использованием метода главных компонент [28]. В запросе: *knowledge graph, ontology, large language models* перечислены не только ключевые слова, но и тематика исследований. В ответ получен список журналов:

- 1) *Ontology of designing* – 0.91;
- 2) *Natural Language Processing Techniques* – 0.84;
- 3) *Logic, programming, and type systems* – 0.83;
- 4) *Topic Modeling* – 0.83;
- 5) *Advanced MIMO Systems Optimization* – 0.83;
- 6) *Handwritten Text Recognition Techniques* – 0.83,

где число указывает на степень близости эмбедингов запроса и журнала.

Близость запроса и ответов определяется на основе косинусного расстояния между запросом и результатами, который определяется выше некоторого порога k (на рисунке 6 $k=0.7$). Рисунок 6 сформирован автоматически путём преобразования эмбедингов журналов в пространство размерности 2. Метод главных компонент находит два направления (главные компоненты, оси для визуализации), вдоль которых данные имеют наибольшую изменчивость, т.е. сохраняет максимум информации из исходного набора векторов. Все объекты проецируются на две оси. Полученные координаты используются для построения точек на плоскости, где каждая точка соответствует журналу из БС. Близкие точки соответствуют семантически близким объектам (*крестики*). На рисунке 6 запрос помечен *звёздочкой*, а соответствующие ему близкие журналы – *крестиками*. Это представление векторизованных данных на плоскости позволяет получить представление о том, насколько связаны данные, а также насколько близки результаты к запросу и между собой. Видно, что *крестики* сгруппированы в одной зоне векторного пространства и близки к запросу (*звёздочке*). Иллюстрации на основе метода главных компонент позволяют качественно оценить достоверность полученных оценок близости эмбедингов запроса и журнала.

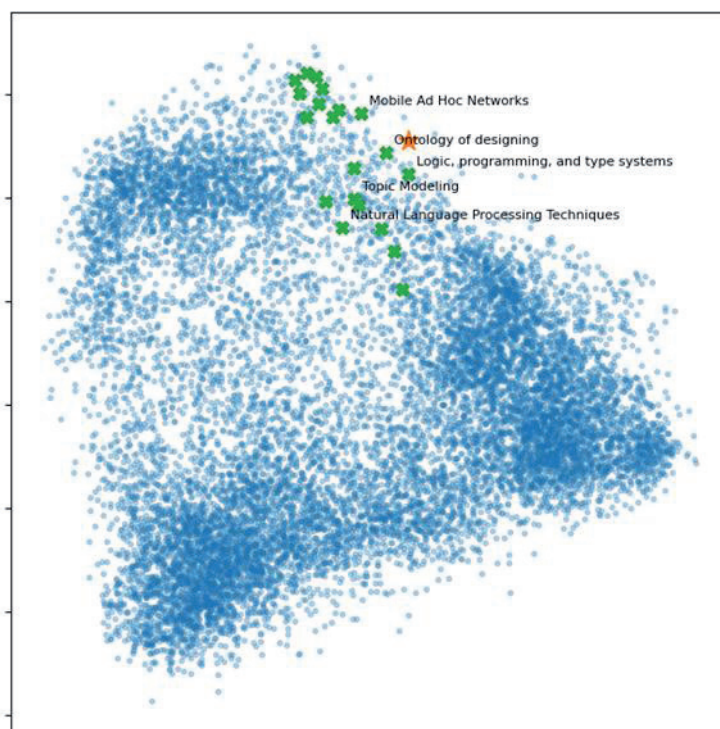


Рисунок 6 – Запрос: *knowledge graph, ontology, large language models* помечен звёздочкой. Результирующий список журналов, отмечен крестиками

4 Применение к другим областям знаний

Предложенная методика не привязана жёстко к сущности «журнал» онтологии *LibMeta* (*SciLibRu*) и может быть распространена на другие типы научных источников, например, «конференции». Если имеется аналогичный информационный ресурс, такой, как БС или реестр значимых конференций по различным ПрО, можно включить эти данные в онтологию *SciLibRu* согласно предложенной методике. Потребуется сформировать тематическое описа-

ние конференций (например, по ключевым словам из докладов, секциям и областям знаний, охватываемым конференцией), обучить модель классификации конференций по рубрикам ГРНТИ или иной рубрикации и построить ОУ конференций. В ГЗ *SciLibRu* появятся узлы «Конференция» со связями с темами и классификаторами, что даст возможность рекомендательной системе подбирать релевантные конференции для представления доклада.

Заключение

В работе предложен подход к интеграции и тематическому разбиению журналов с использованием методов онтологического моделирования и машинного обучения. На основе рубрикатора ГРНТИ и его соответствий с другими классификаторами сформирован ОУ для научных журналов из БС. В полученной онтологической модели информация о журналах (тематика, метрики, индексирование и пр.) связана семантическими отношениями, что обеспечивает несколько вариантов входа ОУ при поиске и навигации. Для автоматического отнесения журналов и публикаций к тематическим разделам применены методы обработки текстов: составлены тематические профили рубрик на основе ключевых фраз и на этих данных обучены ААКО. Эксперименты показали возможность классификации журналов по тематикам ГРНТИ с помощью ААКО, что объясняется использованием высокого уровня рубрикации и специальной обработки дисбаланса классов и использования профильных эмбедингов. Применение ГЗ для представления интегрированных данных о журналах позволяет создать рекомендательную систему, помогающую подобрать журнал для публикации. ГЗ обеспечивает навигацию по данным: пользователь может начать поиск с любой интересующей информации (тематика, ключевые слова статьи, наличие журнала в определённых базах, требуемый квартиль и др.) и получить набор релевантных изданий. Методы машинного обучения позволяют автоматизировать пополнение ГЗ новыми связями, определяя тематическую рубрику новой статьи или журнала, что снижает трудозатраты экспертов. Созданная рекомендательная система в сфере научных коммуникаций способна учитывать множество различных факторов, объяснять предложенные рекомендации и стать полезным инструментом для исследователей при планировании публикаций, облегчая выбор журнала.

Список источников

- [1] *Hogan A., Blomqvist E., Cochez M., et al.* Knowledge Graphs. *ACM Computing Surveys*. 2021. Vol.54, No.4. Article 71. P.1–37. DOI: 10.1145/3447772.
- [2] *Schuemie M.J., Kors J.A. Jane*: Suggesting Journals, Finding Experts. *Bioinformatics*. 2008. Vol.24, No.5. P.727–728. DOI: 10.1093/bioinformatics/btn006.
- [3] *Kosmulski M.* Generalized g-index. *Scientometrics*. 2025. V.130. P.531–536. DOI: 10.1007/s11192-024-05221-x.
- [4] *Vatolin A., Gerasimenko N., Ianina A. et al.* RuSciBench: Open Benchmark for Russian and English Scientific Document Representations. *Dokl. Math.* 110 (Suppl 1), S251–S260 (2024). DOI: 10.1134/S1064562424602191.
- [5] *Елизаров А.М., и др.* Онтологии математического знания и рекомендательная система для коллекций физико-математических документов. *Доклады РАН*. 2016. Т.467. №4. С.392–395. DOI: 10.7868/S0869565216100042.
- [6] *Bravo M., Hoyos Reyes L.F., Reyes Ortiz J.A.* Methodology for ontology design and construction. *Contaduría y Administración*. 2019. Vol.64, No.4. e134. DOI: 10.22201/fca.24488410e.2020.2368.
- [7] *Елизаров А.М., Жильцов Н.Г., Иванов В.В., и др.* Семантический рекомендательный сервис в профессиональной деятельности математика. *Ученые записки Института социальных и гуманитарных знаний*. 2015. Вып.1(13). С.190–197.
- [8] *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMathPRO ontology: a linked data hub for mathematics. *Communications in Computer and Information Science*. 2014. Vol.468. P.105–119. DOI: 10.1007/978-3-319-11716-4_9.

- [9] **Serebryakov V.A., Ataeva O.M.** Ontology-based approach to modeling of the subject domain “Mathematics” in the digital library. *Lobachevskii Journal of Mathematics*. 2021. Vol.42, No.8. P.1920–1934. DOI: 10.1134/S199508022108028X.
- [10] **Ataeva O., Kornet Yu.N., Serebryakov V., and Tuchkova N.** Approach to creating a thesaurus and a knowledge graph of an applied subject area. *Lobachevskii J. of Mathematics*, 2023. Vol.44, No.7. P.2577–2586. DOI: 10.1134/S1995080223070077.
- [11] **Атаева О.М., Массель Л.В., Серебряков В.А., Тучкова Н.П.** Интеллектуальный анализ данных при построении графа знаний мультидисциплинарного журнала. *ИМТ*, 2024, № 3(35). С.5-19. DOI: 10.25729/ESI.2024.35.3.001.
- [12] **Ataeva O., Serebryakov V., Tuchkova N., Strebkov I.** Ontology and Knowledge Graph of Mathematical Physics in the Semantic Library MathSemanticLib. 26th International Conference, DAMDID/RCDL 2024, Nizhny Novgorod, Russia, October 23–25, 2024, Revised Selected Papers. CCIS 2641/ P.48-63, 2025.
- [13] **Kobyshev K., Voinov N., Nikiforov I.** Hybrid image recommendation algorithm combining content and collaborative filtering approaches. *Procedia Computer Science*, 2021, Vol.193. P.200-209, DOI: 10.1016/j.procs.2021.10.020.
- [14] **Shuting Zhang, Kechen Liu, Zekai Yu, Bowen Feng, Zijie Ou.** Hybrid recommendation system combining collaborative filtering and content-based recommendation with keyword extraction. *The 4th International Conference on Computing and Data Science (CONF-CDS 2022)* 2022, P. 927-939. DOI: 10.54254/2755-2721/2/20220579.
- [15] **Haibo H., Yang B., Edwardo G., Shutao L.** ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *Proceedings of the International Joint Conference on Neural Networks*. 2008. P.1322-1328. DOI: 10.1109/IJCNN.2008.4633969.
- [16] **Chawla N.V. et al.** SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*. 2002. T.16. C.321-357. <https://arxiv.org/abs/1106.1813>.
- [17] **Johnson J.M., Khoshgoftaar T.M.** Survey on deep learning with class imbalance. *J Big Data*. 2019. Vol.6. P.1–54. DOI: 10.1186/s40537-019-0192-5.
- [18] **Shyrokykh K., Girnyk M., Dellmuth L.** Short text classification with machine learning in the social sciences: The case of climate change on Twitter. 2023. P.1–26. <https://arxiv.org/abs/2310.04452>.
- [19] **Gabi Goobar A., Regefalk D.** Classification of Short Text Messages Using Machine Learning. LU-CS-EX. 2020.
- [20] **Mikolov T., Chen K., Corrado G., Dean J.** Efficient Estimation of Word Representations in Vector Space. *Proc. of the International Conference on Learning Representations (ICLR)*. 2013.
- [21] **Feng X., Zhang H., Ren Y., et al.** The Deep Learning–Based Recommender System “Pubmender” for Choosing a Biomedical Publication Venue: Development and Validation Study. *Journal of Medical Internet Research*. 2019. Vol.21, No.5. P.e12957. DOI: 10.2196/12957.
- [22] **Liu C., Wang X., Liu H., et al.** Learning to recommend journals for submission based on embedding models. *Neurocomputing*. 2022. Vol.508. P.242–253. DOI: 10.1016/j.neucom.2022.08.043.
- [23] **Bouaguel W., Benyounes N., Ben Ncir C.E.** Enhancing research publication choices: A comparative study of journal recommender systems and their effectiveness. *International Journal of Advanced and Applied Sciences*. 2024. Vol.11, No.5. P.217–229. DOI: 10.1109/ICCPCT.2016.7530304.
- [24] **Zhang J.-C., Zain A.M., Zhou K.-Q., et al.** A review of recommender systems based on knowledge graph embedding. *Expert Systems with Applications: An International Journal*. 2024. Vol. 250. Issue C. DOI: 10.1016/j.eswa.2024.123876.
- [25] **Jurafsky D., Martin J.H.** Speech and Language Processing. Draft of January 12, 2025. <https://web.stanford.edu/~jurafsky/slp3/>.
- [26] **Ribeiro M.T., Singh S., Guestrin C.** “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. 2016. P.1135–1144. DOI: 10.1145/2939672.293977.
- [27] **Будзко В.И., Атаева О.М., Тучкова Н.П.** Автоматизация доступа к информации при навигации по данным семантической библиотеки и интеграции графа знаний с языковой моделью. *Системы высокой доступности*. 2025. Т.21. №2. С.5–11. DOI: 10.18127/j20729472-202502-0.
- [28] **Greenacre M., Groenen P.J.F., Hastie T. et al.** Principal component analysis. *Nat Rev Methods Primers*. 2022. Article number: 100. DOI: 10.1038/s43586-022-00184-w.

Сведения об авторах

Атаева Ольга Муратовна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, к.т.н. (2020), окончила математический факультет Северо-Осетинского государственного университета им. К.Л. Хетагурова, специалист в области системного программирования, баз данных, инженерии знаний и онтологии. AuthorID (РИНЦ): 129291; ORCID: 0000-0003-0367-5575. oataeva@frccsc.ru.



Тучкова Наталья Павловна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат физ.-мат. наук (2004), окончила факультет высшей математики и кибернетики МГУ им. М.В. Ломоносова. Специалист в области алгоритмических языков и информационных технологий. AuthorID (РИНЦ): 102162; Scopus Author ID: 15726569900. ORCID 0000-0001-5357-9640. ntuchkova@frccsc.ru. ✉



Дегтев Артём Геннадьевич (1995 г.р.), студент магистратуры Московского физико-технического института, направление «Науки о данных». AuthorID (РИНЦ): 1280137; ORCID: 0009-0006-6818-7782. Degtev_i_co@mail.ru.

Поступила в редакцию 15.09.2025, после рецензирования 24.10.2025. Принята к публикации 30.10.2025.



Scientific article

DOI: 10.18287/2223-9537-2025-15-4-598-613

Recommendation system based on a generalized journal index

© 2025, O.M. Ataeva¹, N.P. Tuchkova¹✉, A.G. Degtev²

¹ Federal Research Center Informatics and Management of the Russian Academy of Sciences (FRC IM RAS), Moscow, Russia

² Moscow Institute of Physics and Technology (National Research University), MIPT, Dolgoprudny, Russia

Abstract

This paper explores the thematic classification of scientific journals using the example of the "White List"—the Unified State Register of Scientific Journals. The study aims to automate the analysis of journal thematic orientations. For this purpose, the State Rubricator of Scientific and Technical Information (SRSTI), classifiers such as the Universal Decimal Classification (UDC) and the Mathematics Subject Classification (MSC), as well as the ontology of the SciLibRu semantic library of subject domains, are utilized. Based on data from journals included in the "White List" and from sources within the SciLibRu library, a generalized index has been developed and integrated into the SciLibRu knowledge graph. This enables users of the SciLibRu library to navigate various journal attributes, such as subject area and category, thus facilitating the selection of appropriate publication venues. An example is provided of how semantic analysis of an article can be used to assess its thematic relevance to journals from the "White List." The generalized index incorporated into SciLibRu also allows users to formulate natural language queries to identify suitable journals for publication. The proposed methodology can be extended to other domains, such as conferences proceedings. The practical significance of the study lies in the automation of journal topic selection for scientific manuscripts.

Keywords: White List of journals, domain ontology, recommendation system, classifier, generalized journal index, semantic library.

For citation: Ataeva OM, Tuchkova NP, Degtev AG. Recommendation system based on a generalized journal index [In Russian]. *Ontology of designing*. 2025; 15(4): 598-613. DOI: 10.18287/2223-9537-2025-15-4-598-613.

Funding: This work was completed as part of a state assignment FFNG-2024-0003 on the topic "Mathematical Methods of Data Analysis and Forecasting."

Authors' contributions: Ataeva O.M. – data processing, preparation of examples; Tuchkova N.P. – article structure development, source analysis; Degtev A.G. – preparation of examples.

Conflict of interest: The authors declare no conflict of interest.

List of figures and tables

Figure 1 – Scheme for selecting a journal from SciLibRu for a prospective publication

Figure 2 – Stages of expanding the subject domain "Mathematics" using the example of «Mechanics of composite materials and structures» journal (ODE – ordinary differential equations, MathNet – All-Russian portal Math-Net.Ru)

Figure 3 – Diagram of the proposed method: integration of "White List" (WL) data and knowledge sources into the generalized index (GI) based on the SciLibRu knowledge graph (KG), and use of the generalized index for recommending journals according to a user query

Figure 4 – Histogram of the distribution of key phrases across top-level SRSTI categories

Figure 5 – Information about a "White List" journal uploaded to the library with its connections

Figure 6 – Query: *knowledge graph, ontology, large language models* marked with an asterisk. The resulting list of journals is indicated by crosses

Table 1 – Resources for thematic journal search for manuscript publication

Table 2 – Results of applying algorithms for automatic classification of objects (AACO)

Table 3 – Impact of data enrichment on the quality of thematic classification

References

- [1] **Hogan A, Blomqvist E, Cochez M, et al.** Knowledge Graphs. *ACM Computing Surveys*. 2021; 54(4): Article 71. P.1–37. DOI: 10.1145/3447772.
- [2] **Schuemie MJ, Kors JA.** Jane: Suggesting Journals, Finding Experts. *Bioinformatics*. 2008; 24(5): 727–728. DOI: 10.1093/bioinformatics/btn006.
- [3] **Kosmulski M.** Generalized g-index. *Scientometrics*. 2025; 130: 531–536. DOI: 10.1007/s11192-024-05221-x.
- [4] **Vatolin A, Gerasimenko N, Ianina A, et al.** RuSciBench: Open Benchmark for Russian and English Scientific Document Representations. *Dokl. Math.* 110 (Suppl 1), S251–S260 (2024). DOI: 10.1134/S1064562424602191.
- [5] **Elizarov AM, et al.** Ontologies of mathematical knowledge and a recommendation system for collections of physical and mathematical documents [In Russian]. *Reports of the Russian Academy of Sciences*. 2016; 467(4): 392–395. DOI: 10.7868/S0869565216100042.
- [6] **Bravo M, Hoyos Reyes LF, Reyes Ortiz JA.** Methodology for ontology design and construction. *Contaduría y Administración*. 2019; 64(4): e134. DOI: 10.22201/fca.24488410e.2020.2368.
- [7] **Elizarov AM, Zhiltsov NG, Ivanov VV, et al.** Semantic recommendation service in the professional activity of a mathematician [In Russian]. *Scientific Notes of the Institute of Social and Humanitarian Knowledge*. 2015; 1(13): 190–197.
- [8] **Nevzorova O, Zhiltsov N, Kirillovich A, Lipachev E.** OntoMathPRO ontology: a linked data hub for mathematics. *Communications in Computer and Information Science*. 2014; 468: 105–119. DOI: 10.1007/978-3-319-11716-4_9.
- [9] **Serebryakov VA, Ataeva OM.** Ontology-based approach to modeling of the subject domain "Mathematics" in the digital library. *Lobachevskii Journal of Mathematics*. 2021; 42(8): 1920–1934. DOI: 10.1134/S199508022108028X.
- [10] **Ataeva O, Kornet YuN, Serebryakov V, Tuchkova N.** Approach to creating a thesaurus and a knowledge graph of an applied subject area. *Lobachevskii J. of Mathematics*. 2023; 44(7): 2577–2586. DOI: 10.1134/S1995080223070077.
- [11] **Ataeva OM, Massel LV, Serebryakov VA, Tuchkova NP.** Intelligent data analysis in constructing a knowledge graph of a multidisciplinary journal [In Russian]. *IMT*. 2024; 3(35): 5-19. DOI: 10.25729/ESI.2024.35.3.001.
- [12] **Ataeva O, Serebryakov V, Tuchkova N, Strebkov I.** Ontology and Knowledge Graph of Mathematical Physics in the Semantic Library MathSemanticLib. 26th International Conference, DAMDID/RCDL 2024, Nizhny Novgorod, Russia, October 23–25, 2024, Revised Selected Papers. CCIS 2641, 2025. P.48-63,
- [13] **Kobyshev K, Voinov N, Nikiforov I.** Hybrid image recommendation algorithm combining content and collaborative filtering approaches. *Procedia Computer Science*, 2021; 193: 200-209, DOI: 10.1016/j.procs.2021.10.020.
- [14] **Zhang S, Liu K, Yu Z, Feng B, Ou Z.** Hybrid recommendation system combining collaborative filtering and content-based recommendation with keyword extraction. *The 4th International Conference on Computing and Data Science (CONF-CDS 2022)* 2022, P. 927-939. DOI: 10.54254/2755-2721/2/20220579.
- [15] **Haibo H, Yang B, Eduardo G, Shutao L.** ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *Proceedings of the International Joint Conference on Neural Networks*. 2008. P.1322-1328. DOI: 10.1109/IJCNN.2008.4633969.

- [16] **Chawla NV, et al.** SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*. 2002; 16: 321-357. <https://arxiv.org/abs/1106.1813>.
- [17] **Johnson JM, Khoshgoftaar TM.** Survey on deep learning with class imbalance. *J Big Data*. 2019; 6, 27. DOI: 10.1186/s40537-019-0192-5.
- [18] **Shyrokykh K, Girnyk M, Dellmuth L.** Short text classification with machine learning in the social sciences: The case of climate change on Twitter. 2023. <https://arxiv.org/abs/2310.04452>.
- [19] **Gabi Goobar A, Regefalk D.** Classification of Short Text Messages Using Machine Learning. LU-CS-EX. 2020.
- [20] **Mikolov T, Chen K, Corrado G, Dean J.** Efficient Estimation of Word Representations in Vector Space // Proc. of the International Conference on Learning Representations (ICLR). 2013.
- [21] **Feng X, Zhang H, Ren Y, et al.** The Deep Learning-Based Recommender System “Pubmender” for Choosing a Biomedical Publication Venue: Development and Validation Study. *Journal of Medical Internet Research*. 2019; 21(5): e12957. DOI: 10.2196/12957.
- [22] **Liu C, Wang X, Liu H, et al.** Learning to recommend journals for submission based on embedding models. *Neurocomputing*. 2022; 508: 242–253. DOI: 10.1016/j.neucom.2022.08.043.
- [23] **Bouaguel W, Benyounes N, Ben Ncir CE.** Enhancing research publication choices: A comparative study of journal recommender systems and their effectiveness. *International Journal of Advanced and Applied Sciences*. 2024; 11(5): 217–229. DOI: 10.1109/ICCPCT.2016.7530304.
- [24] **Zhang J-C, Zain AM, Zhou K-Q, et al.** A review of recommender systems based on knowledge graph embedding. *Expert Systems with Applications*. 2024; 250(C). DOI: 10.1016/j.eswa.2024.123876.
- [25] **Jurafsky D, Martin JH.** Speech and Language Processing. Draft of January 12, 2025. <https://web.stanford.edu/~jurafsky/slp3/>.
- [26] **Ribeiro MT, Singh S, Guestrin C.** “Why Should I Trust You?”: Explaining the Predictions of Any Classifier // Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. 2016. P.1135–1144. DOI: 10.1145/2939672.293977.
- [27] **Budzko VI, Ataeva OM, Tuckkova NP.** Automation of access to information navigating through semantic library data and integrating a knowledge graph with a language model [In Russian]. *High Availability Systems*. 2025; 21(2): 5–11. DOI: 10.18127/j20729472-202502-0.
- [28] **Greenacre M, Groenen PJF, Hastie T, et al.** Principal component analysis. *Nat Rev Methods Primers*. 2022. Article number: 100. DOI: 10.1038/s43586-022-00184-w.

About the authors

Olga Muratovna Ataeva – Senior Researcher at the Dorodnicyn Computing Centre of the FRC CSC RAS, Candidate of Technical Sciences (2020). She graduated from the Faculty of Mathematics, North Ossetian State University named after K.L. Khetagurov. Her areas of expertise include system programming, databases, knowledge engineering and ontology. Author ID (RSCI): 129291; ORCID: 0000-0003-0367-5575. oataeva@frccsc.ru.

Natalia Pavlovna Tuckkova – Senior Researcher at the Dorodnicyn Computing Centre of the FRC CSC RAS, Candidate of Physical and Mathematical Sciences (2004), graduated from the Faculty of Computational Mathematics and Cybernetics of the Lomonosov Moscow State University. Her research interests include algorithmic languages and information technologies. Author ID (RSCI): 102162; ORCID 0000-0001-5357-9640. ntuckkova@frccsc.ru. ✉

Artem Gennadievich Degtev (b. 1995), Master's student at MIPT, specializing in Data Science. AuthorID (РИНЦ): 1280137. ORCID: 0009-0006-6818-7782. Degtev_i_co@mail.ru.

Received September 15, 2025. Revised October 22, 2025. Accepted October 30, 2025.