Методологические аспекты информационных и математических технологий

УДК 004.8+519.217 + 519.876.2
DOI:10.25729/ESI.2024.35.3.001

Интеллектуальный анализ данных при построении графа знаний мультидисциплинарного журнала

Атаева Ольга Муратовна[1], Массель Людмила Васильевна[2], Серебряков Владимир Алексеевич[1], Тучкова Наталия Павловна[1]

[1]ФИЦ «Информатика и управление» РАН, Россия, Москва, *oli.ataeva@gmail.com*
[2]Институт систем энергетики им. Л.А. Мелентьева СО РАН, Россия, Иркутск

Аннотация. В работе исследуется тематическое многообразие междисциплинарного журнала. Цель исследований составляет построение графа знаний журнала для тематического представления и систематизации электронного архива и новых публикаций журнала. Исходные данные представляют собой статьи журнала, посвященные различным информационным и математическим технологиям в науке и управлении, то есть междисциплинарным исследованиям. Предлагается систематизация текстов с помощью методов векторного анализа. В процессе тематического анализа контента журнала предлагается разбиение на рубрики, устанавливаются связи рубрик и статей с соответствующими описаниями специальностей ВАК. Для анализа тематики используется разведочный анализ исходных текстов, далее применяются методы интеллектуального анализа данных. Результаты разбиения предоставляются экспертам журнала, после чего вырабатывается решение о формировании тематической рубрики и включении в нее специальностей ВАК. Статьи журнала интегрируются в семантическую библиотеку LibMeta, в силу чего онтология библиотеки достраивается и формируется онтология журнала, и на этой основе строится граф знаний журнала. Предлагается процедура навигации по контенту журнала с помощью графа знаний в семантической библиотеке LibMeta, которая может стать основой для информационного сопровождения научных исследований и создания цифрового ассистента в междисциплинарной предметной области. Примеры приведены для конкретного контента журнала, но предложенная технология может быть распространена на другие журналы, так как большинство журналов, относящихся к нескольким специальностям ВАК, естественным образом захватывают несколько дисциплин.

Ключевые слова: граф знаний, семантическая библиотека, достраивание онтологии, кластеризация научных статей, суммаризация текста

Цитирование: Атаева О.М. Интеллектуальный анализ данных при построении графа знаний мультидисциплинарного журнала / О.М. Атаева, Л.В. Массель, В.А. Серебряков, Н.П. Тучкова // Информационные и математические технологии в науке и управлении, 2024. – № 3(35). – С. 5-19. – DOI: 10.25729/ESI.2024.35.3.001.

Введение. Задача тематической кластеризации обсуждается достаточно давно, начиная с создания первых библиографических массивов данных [1]. Для анализа научных публикаций эта задача будет ставиться снова и снова в процессе развития науки и появления новых предметов исследования. Кластеризация данных предполагает группировку данных на основе присущих им сходств без заранее определенных категорий [2, 3]. Наиболее известные научные библиографические ресурсы, такие как, WoS (https://clarivate.com/cis/solutions/web-of-science/), Scopus (https://www.scopus.com/), eLIBRARY (https://www.elibrary.ru/), предлагают поиск по ключевым словам и другим метаданным, которые, в свою очередь, связаны как с публикацией, так и с конкретной предметной областью (ПрО). В целом, эти ресурсы не предлагают сервиса по анализу содержания публикаций, исключение составляет zbMATH Open (https://zbmath.org/), где можно найти отзывы и рецензии. В то же время нарастающий объем публикаций оставляет необходимость *в предварительной оценке тематики* публикации прежде, чем купить или скачать ее для ознакомления. Работы по кластеризации научных тек-

стов с целью их тематического распределения в семантической среде продолжаются и нарастают с применением средств машинного обучения (МО) и моделей естественного языка [4, 5, 6]. Определенный скачок в этом направлении произошел в связи с применением векторных алгоритмов [7]. Почти через полвека появилась технология построения семантических векторных представлений, или эмбеддингов (embedding) [8, 9, 10], в том числе и для научных областей. Так, для eLIBRARY разработана нейронная сеть для получения эмбеддингов научных текстов [11, 12, 13] на основе аннотаций к статьям, с помощью которых можно решать различные задачи обработки текстов. Тем не менее, далеко не все ПрО и не все журналы в полном объеме охвачены этой технологией и проблема тематического поиска остается, а ее актуальность неизбежно связана с экспоненциальным ростом числа публикаций. К тому же есть междисциплинарные ПрО, для которых тоже необходимо составить картину принадлежности к множеству ПрО. Предлагаемое исследование посвящено именно междисциплинарной тематике журнала и кластеризации этого контента с целью выявления множеств научных статей, которые можно отнести к разным ПрО.

Новый этап развития методов искусственного интеллекта (ИИ) связан с появлением новых эффективных нейросетевых методов для обработки естественного языка. Но, несмотря на это, особые трудности возникают при обработке научных статей из-за специфичной для ПрО терминологии и сложных идей, часто встречающихся в научной литературе [14,15]. Однако, использование графов знаний (ГЗ), основанных на онтологии ПрО, в совокупности с размеченными статьями при тонкой настройке нейросетевых методов, позволяет улучшить результаты обработки и извлечения смысла текстов, а поиск становится более обоснованным [16, 17, 18].

ГЗ обычно представляется как набор объектов, связанных между собой. Объекты представляют собой узлы этого графа, а связи представляются в виде ребер. Для некоторой ПрО ГЗ ограничивается определенным типом объектов и заданным набором связей. Формально при этом ПрО задается в виде онтологии $<R, A, C, I>$, которая представляет собой типы объектов R, их атрибуты A и отношения C, а также функции интерпретации I этих отношений. Таким образом, *множество* $\{R, A, C, I\}$ задает описание структуры ПрО.

Представление знаний в виде онтологии ПрО и навигация с помощью ГЗ представляет собой технологию поиска, при которой можно оставаться в рамках тематики и опираться на достоверные данные, связанные с энциклопедиями, словарями, тезаурусами, классификаторами и первоисточниками.

Работа направлена именно на задачу построения ГЗ *междисциплинарной* ПрО контента журнала (лов) [17, 18]. С этой целью был проведен интеллектуальный анализ поступающих публикаций для определения их тематической принадлежности. Интеллектуальный анализ текстов подразумевает совместную работу специалистов ПрО и средств ИИ. Для описания ПрО привлекается контент семантической библиотеки LibMeta (libmeta.ru) и классические первоисточники, рекомендованные экспертами. Начальное описание ПрО, как семантически связанной структуры данных, опирается на классификаторы УДК, MSC, рубрики журналов, специальности ВАК журналов, специальные словари и предметные указатели монографий.

Построенный ГЗ журнала интегрируется с энциклопедиями LibMeta и онтологией ИИ в LibMeta. Добавление контента журнала влечет за собой процедуру достраивания онтологии семантической библиотеки.

1. Граф знаний журнала. В работе исследуются проблемы построения и наполнения ГЗ на примере семантической библиотеки LibMeta. Используется подход онтологического проектирования для построения ГЗ. При построении онтологии графа определяются типы узлов и связи. При наполнении графа предполагается добавление новых узлов и связей в соответствии с описанием в онтологии и в соответствии с добавленным контентом журнала.

Научный журнал может быть посвящен отдельной теме (научной дисциплине) или охватывать ряд предметов (междисциплинарность). И в том, и другом случае тематика журнала, как правило, подпадает под несколько специальностей ВАК. В силу развития науки в целом выбор специальностей зависит от тенденций в исследованиях [19, 20] и от позиции редакции. Научная статья – это исследование, которое посвящено какой-то проблеме (задаче) и предполагает структуру изложения, из которой читателю должна быть понятна суть проблемы и пути ее исследования. Современные библиографические базы данных предлагают различные варианты поиска публикаций, включая использование отзывов, ссылок на аналогичные (с позиции базы данных) публикации и пр. В целом, это путь для помощи пользователю в навигации по библиографическим данным. Большие информационные поисковые ресурсы (google и др.) используют связи данных и строят для навигации ГЗ. Эта технология, основанная на семантическом анализе контента и нейросетевых алгоритмах, также применима для полнотекстовых научных библиотек, а именно, информационного сопровождения автора, создания цифрового ассистента научных исследований.

Цифровой ассистент (семантическая библиотека) выполняет задачу информационного сопровождения автора и редактора в подборе УДК, ключевых слов, аналогичных статей, суммаризации, автоматического аннотирования и др.

Для реализации подхода навигации по научному контенту с помощью ГЗ требуется значительная подготовка текстов, что составляет существенное отличие от текстов, с которыми работают поисковые агрегаторы в интернет. Несмотря на известную структуру научных статей, в них необходимо выделить *предмет* исследования и *идеи* его реализации [14]. Поэтому особую важность приобретает тематическая кластеризация и соответствие специальностям Высшей аттестационной комиссии (ВАК).

Логика построения научных публикаций журнала, как правило, организована таким образом, что в тексте присутствуют выражения типа: «в работе исследуется **задача**», «работа посвящена **проблеме**», или близкие варианты. Далее следуют **методы** исследования, **результаты**, **примеры** и т.д. В контексте специальностей ВАК ПрО журнала включает описание множества задач, соответствующих выбранным специальностям. С другой стороны, тематические рубрики журнала также отражают позицию редакции к соответствию специальностям конкретных задач, отраженных в статьях. Таким образом, ПрО журнала должна включать описание множества задач, соответствующих, с точки зрения редакции, как специальностям ВАК, так и предметным рубрикам журнала. При таком взгляде на ПрО журнала вершинами ГЗ журнала могут рассматриваться тематические рубрики. Рубрики могут быть связаны с несколькими специальностям ВАК и включать описания задач, отраженных в статьях, отнесенных редакцией к этим рубрикам. Перемещение по таким узлам графа будет означать навигацию по ПрО журнала, в том числе по рубрикам и специальностям ВАК, относящимся к рубрикам. Эта навигация может быть применена как для журналов, посвященных одной научной дисциплине, так и для многодисциплинарных журналов, и отличается от навигации по библиографическим данным, которая характерна для большинства научных библиографических баз данных.

Тематическая кластеризация в исследовании реализована, как **распределение статей по рубрикам** и **распределение по специальностям ВАК**. Эти две задачи включают этапы подготовки данных, обработку данных с помощью методов МО и дальнейший анализ данных для их интеграции в цифровую библиотеку путем доработки онтологии библиотеки и построения на ее основе ГЗ ПрО журнала. На рисунке 1 представлен пример ГЗ ПрО журнала после интеграции и достраивании онтологии LibMeta. На первом уровне представлены общие понятия системного тезауруса, на втором описываются понятия ПрО, которые определяют типы узлов

в ГЗ ПрО, на третьем уровне задаются отношения между ними. На четвертом уровне — представлен уровень данных или ГЗ. На рисунке 1 приведена информация в ГЗ о публикации и показаны некоторые ее связи с типами узлов и конкретные данные (специальность, рубрика журнала, название статьи и т.д.).

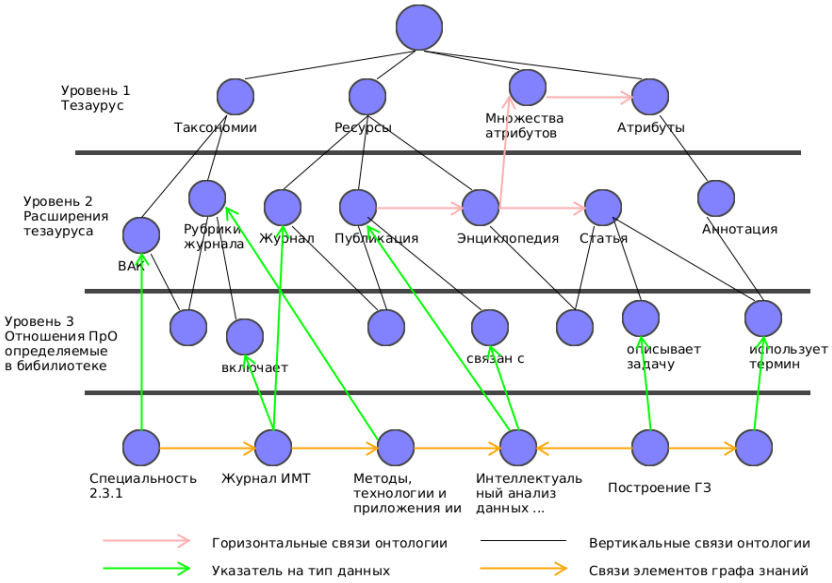


Рис. 1. Пример ГЗ журнала в семантической библиотеке LibMeta

2. Распределение по рубрикам и тематическим кластерам.

2.1. Разведочный анализ данных журнала. Предварительный анализ контента журнала и начальных рубрик выявил проблемы дублирования и широкий тематический охват исследований, опубликованных в статьях.

Для выполнения качественной кластеризации данных необходимо привлечение эксперта ПрО для оценки результата автоматической обработки с помощью алгоритмов *векторизации* [7, 8, 9], *LDA* (Latent Dirichlet Allocation [21]) и *k-means* [22]. Выявление рубрик и распределение публикаций по новым рубрикам позволит далее связать их явно со специальностями ВАК. Надо отметить, что особенность междисциплинарных журналов заключается именно в наличии статей из *различных научных дисциплин*, объединенных одной общей идеей, декламируемой издателями, которая позволяет публиковать их совместно в одном журнале. Это приводит к тому, что в журнале может обнаружиться очень *много тематических рубрик*, в том числе как достаточно близких, так и различных. В то же время все *статьи должны подчиняться специальностями ВАК*, которые в свою очередь также *допускают довольно широкий тематический диапазон*. Однако, читателю важно понимать какому разделу (рубрике) принадлежит публикация, так как с этим связан выбор статьи для ознакомления, и, в конечном счете, скорость и удобство этого выбора. Значит, в журнале должно быть некоторое разумное, *оптимальное* с точки зрения издателей и читателей количество рубрик, которое отражает, в то же время, тематику статей.

Для определения *оптимального* количества тематических рубрик необходимо подготовить исходные массивы данных и выбрать параметры их автоматической обработки с помощью алгоритмов ИИ и провести разведочный анализ данных журнала.

2.1.1. Подготовка данных. Данные публикаций изначально были представлены в слабоструктурированном виде, в процессе предобработки из них были выделены основные метаданные, такие, как автор, название, аннотация и заголовок и т.д.

Предобработка текстов (очистка) включала работу по токенизации, лемматизации (данные лемматизируются, биграммы добавляются, убираются часто встречающиеся слова, и

слова, которые встречаются меньше 5 раз, убираются имена собственные и т.д.), удалении стоп-слов и разбиении текста на разделы.

Далее с помощью статистических векторных методов, основанных на подсчете *TF-IDF* (Term Frequency-Inverse Document Frequency [23]), были выделены ключевые слова, в частности, ключевые слова, которые были предложены для индексации отдельных документов, где авторские ключевые слова не распознались или отсутствовали. Ниже приведен пример (Пример 1) извлечения из статьи ключевых слов *при предобработке* и *авторских* ключевых слов.

Пример 1.

Тучкова Н. П.', 'Атаева О. М. Подходы к извлечению знаний в научных предметных областях
Ключевые слова: ['тезаурус', 'предметный', 'предметный область', 'знание', 'извлечение знаний', 'область', 'извлечение', 'публикация', 'данные', 'интеллект', 'искусственный', 'метрика', 'искусственный интеллект', 'наукометрический', 'онтология тезаурус', 'цифровой', 'структурирование', 'научный', 'сохранение извлечение', 'наукометрический показатель']
Авторские ключевые слова из статьи: ['структурирование данных', 'тезаурус предметной области', 'метрики', 'тезаурус оду']

2.1.2. Распределение по рубрикам. Начальное разбиение контента журнала за весь период публикаций состояло из 53 рубрик, которые содержали тематические пересечения и дублирование. После предварительной обработки и очистки был произведен анализ текстов и их разбиение на кластеры.

В процессе исследования данных было построено *векторное представление* документов (контента журнала) с помощью статистической меры *TF-IDF* и на основе этого представления было проведено разбиение на кластеры с помощью алгоритма *LDA*. Анализ результатов подтвердил предположение о том, что изначальная рубрикация была избыточна. Было выявлено, что документы часто не попадали в одну рубрику, а распределялись по 5 и более рубрикам.

2.1.3. Оценка разбиения на кластеры. Следующий шаг в исследованиях был направлен на определение примерного «оптимального» количества тематических кластеров. Была применена метрика согласованности тем (Topic coherence) как показатель связности при тематическом моделировании для того чтобы оценить, насколько интерпретируемы темы. В этом случае *темы* представлены как первые *N* слов с наибольшей вероятностью принадлежащих к некой конкретной теме. То есть, показатель связности свидетельствует от том, насколько эти слова *похожи друг на друга* для разного количества тем.

В диапазоне от 5 до 7 расстояния между кластерами перестают сильно изменяться, как видно на рисунке 2.

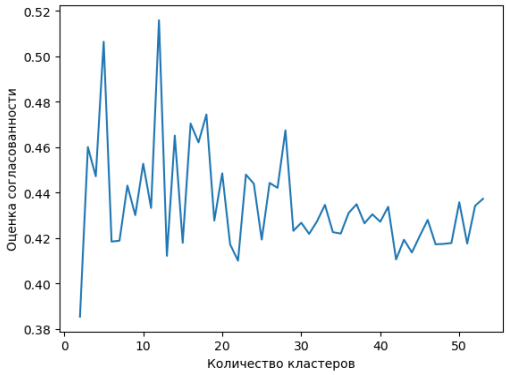


Рис. 2. Кривая оценки согласованности тематик

2.1.3. Метод k-means. Следующий эксперимент был проведен с использованием алгоритма *k-means* с использованием эмбеддингов на основе нейросетевой модели *SciRus-tiny* [19], обученной на научных текстах. Модель имеет небольшое количество параметров и невысокие требования к вычислительным ресурсам.

Для нахождения оптимального количества кластеров была использована функция оценки расстояния между кластерами на основе косинусного расстояния (рисунок 3). Эксперимент показывает, что использование эмбеддингов научных текстов на нейросетевых моделях на данных журнала не сильно сказалось на результате. Для интерпретации получившихся результатов каждый кластер более детально рассмотрен экспертами. Оптимальное количество кластеров было определено равным 12-и.

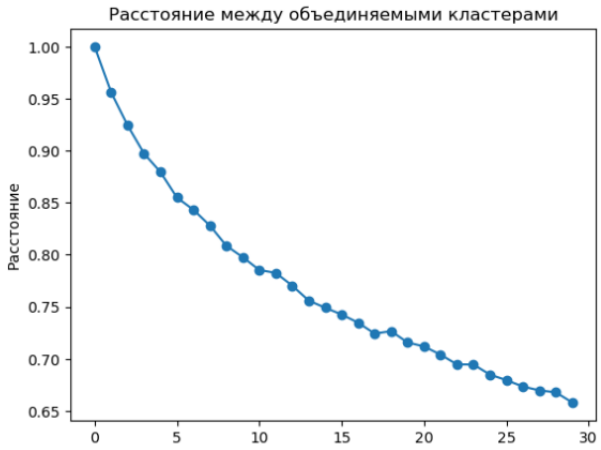


Рис. 3. Оценка расстояния между кластерами на основе косинусного расстояния

Гиперпараметры моделей подбирались на основе анализа и исследования степени соответствия построенной кластерной структуры тематической модели, при этом использована метрика *Silhouette Coefficient*.

На рисунке 4 значение кластеров 5 и 10, выглядят *лучше*, в том смысле, что оценка силуэта кластера больше среднего значения для каждого кластера. Для значений 15 и 20 разбиение *не оптимально,* так как присутствуют кластеры с оценкой меньше среднего, есть большие колебания в размере графика. На рисунке 5 приведена визуализация разбиения на кластеры для k = 5, 10, 15.

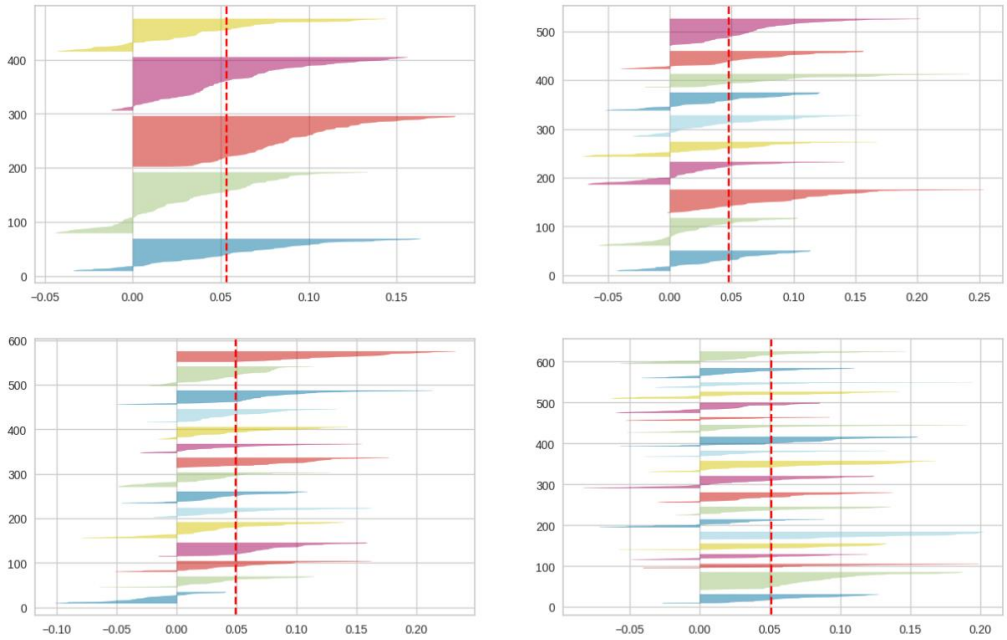
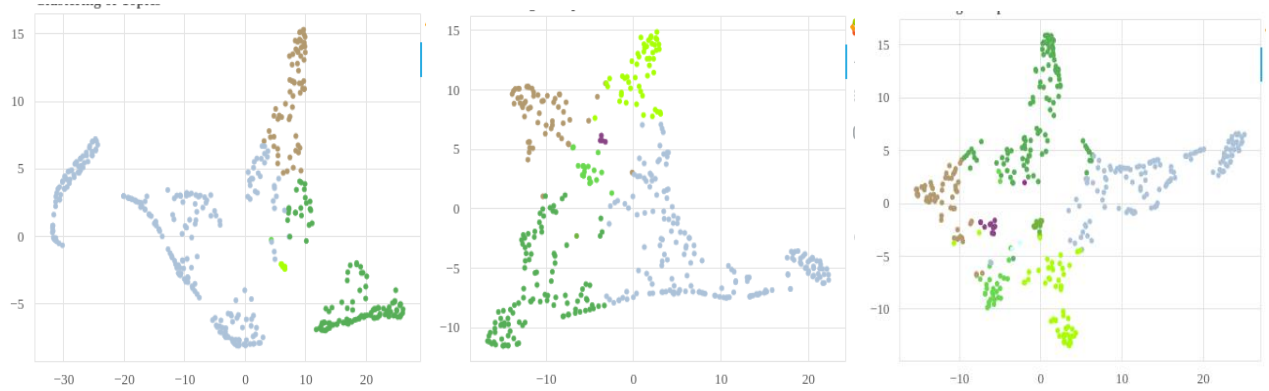


Рис. 4. Оценка расстояния между кластерами на основе косинусного расстояния

Полученные результаты были предоставлены для анализа экспертам и редакционной коллегии журнала для подтверждения и коррекции. Каждый тематический кластер представляется в виде следующих множеств данных: набора ключевых слов и словосочетаний, которые

определяются с помощью статистических методов, а также отдельного набора авторских ключевых слов из статей, которые были отнесены к этому кластеру и списка самих статей.



(а) (б) (в)

Рис. 5. Распределение статей по темам (*k-means*): а) k = 5, б) k =с10, в) k = 15

2.2. Достраивание онтологии. На основе проведенной работы и оценки экспертов были выделены 7 тематических рубрик, которые добавлены в онтологию и связаны с классификаторами (ВАК, УДК), тезаурусом, энциклопедиями, публикациями, ключевыми словами, что иллюстрируется на рисунке 6.

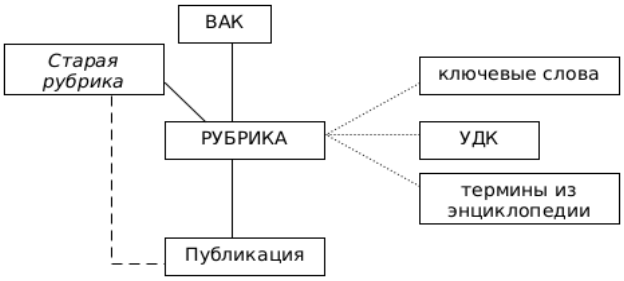


Рис. 6. Схема связей публикации с рубриками журнала и метаданными

Пунктирной линией обозначены связи предыдущего разбиения публикаций на начальные «старые» рубрики. После создания новых рубрик, каждая из них благодаря связям публикаций со «старыми» рубриками, была наполнена ключевыми словами, связями УДК, терминами энциклопедии.

3. Распределение по специальностям ВАК.

3.1. Подготовка данных. Для этого этапа необходимо проанализировать средствами *NLP* [24, 25] не только контент журнала, но и содержание паспорта ВАК. В разделах, которые содержат описание специальности ВАК, прослеживается четкая структура предложений, где выделяются области (сферы) исследования, цели, предмета исследования, средства исследования. Это позволяет проанализировать содержание паспорта ВАК с помощью лингвистических правил.

3.2. Методы. Разбор предложений на структуры был выполнен с помощью лингвистических правил [20], где было выделено: «что делается», «какими методами», «в какой области» и «какой результат должен получиться». Для этого была проанализирована структура предложения, определены связанные слова с применением частеречных меток (POS-тэгирование, part-of-speech tagging), и т.д. Схематически этот процесс разбора содержания паспорта специальности иллюстрируется на рисунке 7.

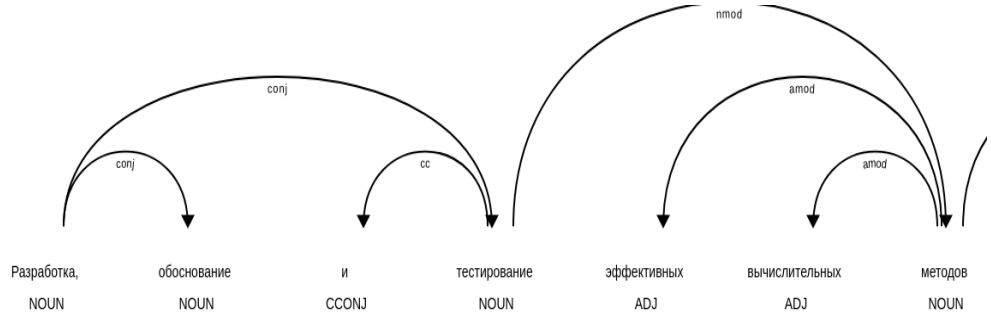


Рис. 7. Связи в предложениях описания специальностей ВАК на основе лингвистических правил

3.3. Анализ результатов. Разбор предложений на структуры позволил увидеть и сформировать *для описания специальности ВАК* словари общей лексики, которые включают в себя, например, описание для всех специальностей: задачи, методы, решения.

Для конкретных специальностей эта общая лексика уточняется, например, для специальности 2.3.1 рассматриваются решения и методы для *задач оптимизации, управления, принятия решений, обработки информации, искусственного интеллекта*.

На рисунке 8 приведено описание паспорта ВАК специальности 1.2.2. В каждом предложении соответствующими цветами выделены: область (серый), задача (желтый), предмет (розовый) и инструмент исследования (фиолетовый).

Паспорт научной специальности 1.2.2.

«Математическое моделирование, численные методы и комплексы программ»

1.Разработка новых математических методов моделирования объектов и явлений

2. Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий.

3. Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента

4. Разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели.

5. Разработка новых математических методов и алгоритмов валидации математических моделей объектов на основе данных натурного эксперимента или на основе анализа математических моделей.

Рис. 8. Пример анализа паспорта ВАК для специальности 1.2.2.

3.4. Достраивание онтологии. В онтологию LibMeta добавлен тезаурус для поддержки описания специальностей ВАК, модель которого включает соответствующие связи и атрибуты, и загружена информация о специальностях, соответствующих контенту журнала (рисунок 9).

На рисунке 9 представлен фрагмент информационной модели специальности ВАК, которая включает свойства «Область исследования», «Предмет исследования», «Инструмент исследования». Анализ паспортов ВАК позволил выделить эти свойства, как общие для всех рассматриваемых специальностей. Некоторые конкретные значения этих свойств из специальности 1.2.2 приведены на рисунке 9.

В результате анализа публикации и определения ее специальности для каждой публикации получаем соответствующее описание по разделам специальности ВАК. Более того, формальные описания специальностей наполняются в тезаурусе LibMeta конкретными ключевыми словами из публикаций журнала. Это позволяет автоматически связывать новые публикации со специальностями для дальнейшей оценки экспертами, рецензентами и авторами.

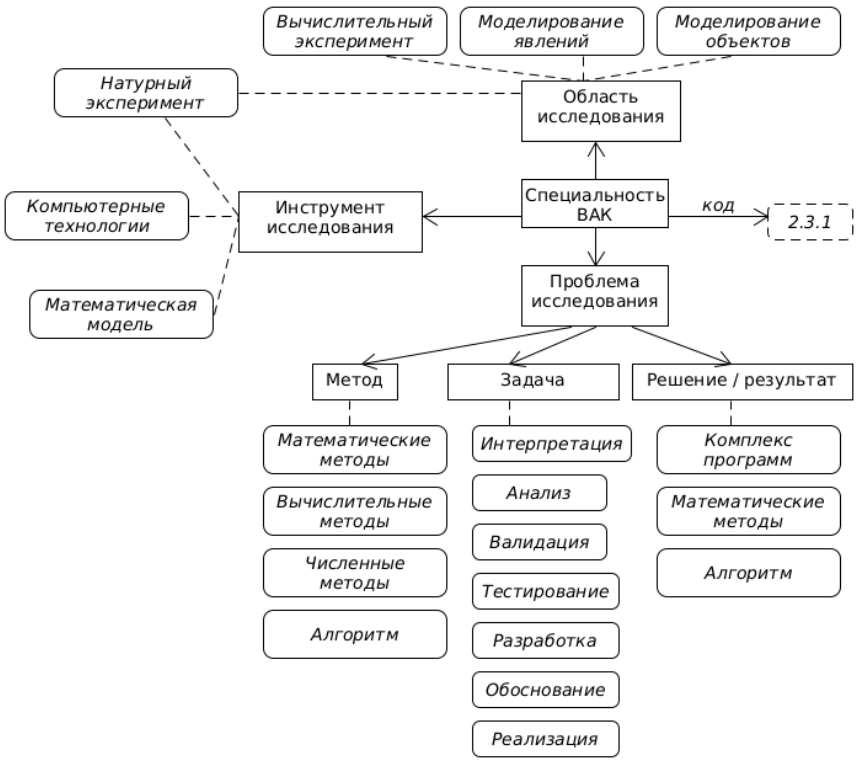


Рис. 9. Схема специальности ВАК 1.2.2 для журнала

4. Методология. Исходные данные представляют собой статьи журнала, посвященные различным информационным и математическим технологиям в науке и управлении, то есть междисциплинарным исследованиям. Большинство журналов, относящихся к нескольким специальностям ВАК, естественным образом захватывают несколько дисциплин. В работе предложен подход к тематической кластеризации на примере одного журнала[1], но используются методы, не ограничивающие его общность, что позволяет говорить об общей методологии в задаче тематической кластеризации таких массивов данных. Результаты предложенных исследований носят прикладной характер, поскольку позволяют организовать навигацию по контенту журнала и использованием графа знаний.

Основная идея, которая была заложена в предлагаемом подходе, – это использовать технологии, которые позволяют работать с «сырым» материалом русскоязычных научных публикаций, выделяя из текстов «задачи», «методы», «решения», «результаты» и помещать их в семантическую библиотеку. Для этого выделяются тематические кластеры и материал разбивается по рубрикам, узлам графа. Множеству рубрик соответствуют ключевые слова, входящие в специальности ВАК. Специальности ВАК – это тоже узлы графа. Специальности ВАК, кроме собственных терминов, приобретают множество ключевых слов из статей, входящих в рубрики. Таким образом, статьи необходимо подготовить, а именно, *выделить задачи, методы их исследования и результаты* с помощью средств обработки естественного языка и консультации экспертов. Такая разметка возможна только совместно с экспертами, если не присутствует в самом тексте статей.

ПрО журнала, таким образом, представлена терминологией, включая описание задач, которым посвящены публикации, при этом вершинами графа ПрО будут *рубрики* журнала и *специальности* ВАК. Перемещаясь по вершинам графа, получаем множества задач, методов,

[1] Рассматривался журнал «Информационные и математические технологии в науке и управлении». В Приложении (после статей этого выпуска) приведен список принятых, после выполненного анализа, рубрик журнала и соответствующих им специальностей ВАК

результатов и т.д., которые относятся к статьям в рубрике, и далее, публикации, в которых изучается та или иная задача междисциплинарной ПрО. Полученный информационный ресурс отличается от того, как журнал присутствует в интернете на основе библиографических данных. Результат (рисунки 10, 11) позволяет быстро *посмотреть, нужна ли пользователю конкретная статья или нет, не скачивая, не читая целиком.* Применение результатов предполагает создание рекомендательной системы для информационного сопровождения пользователя, автора, редактора, рецензента и др.

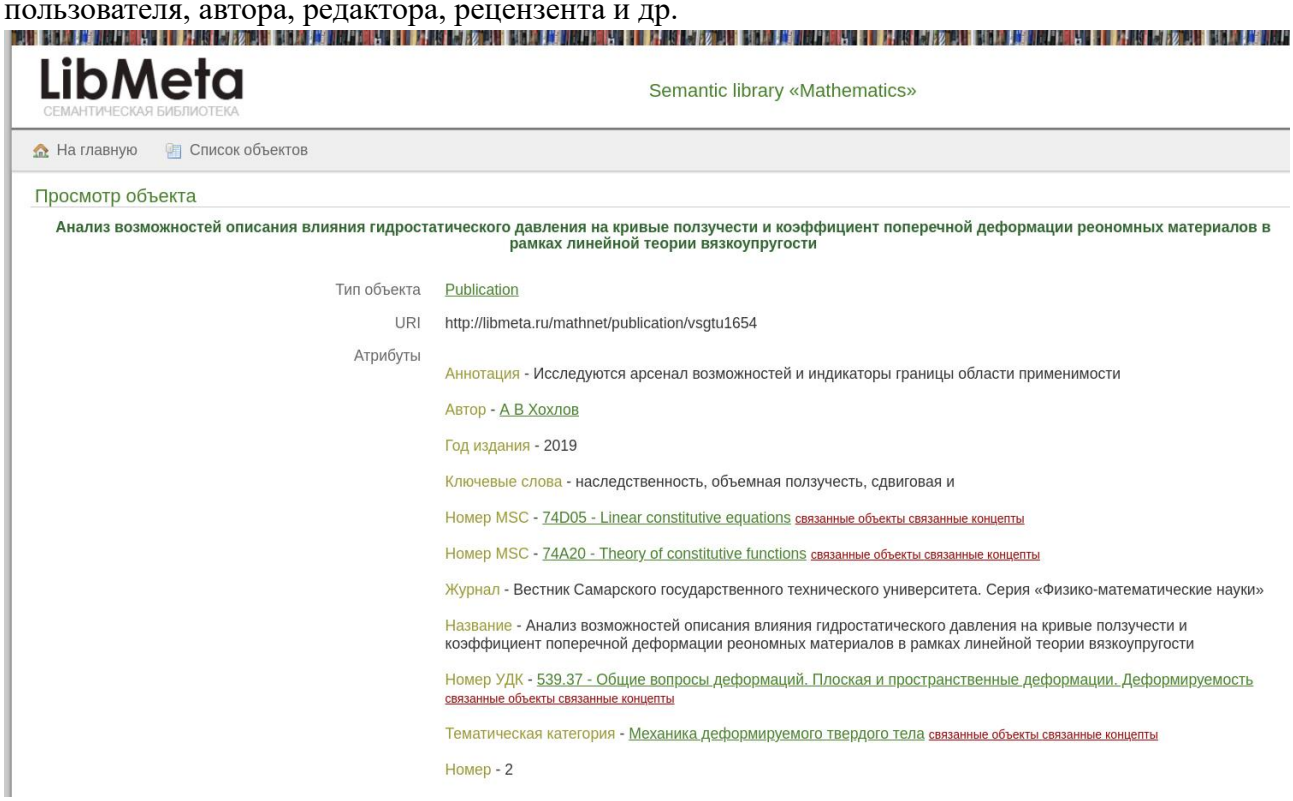


Рис. 10. Скрин 1. Публикация и тематические разделы

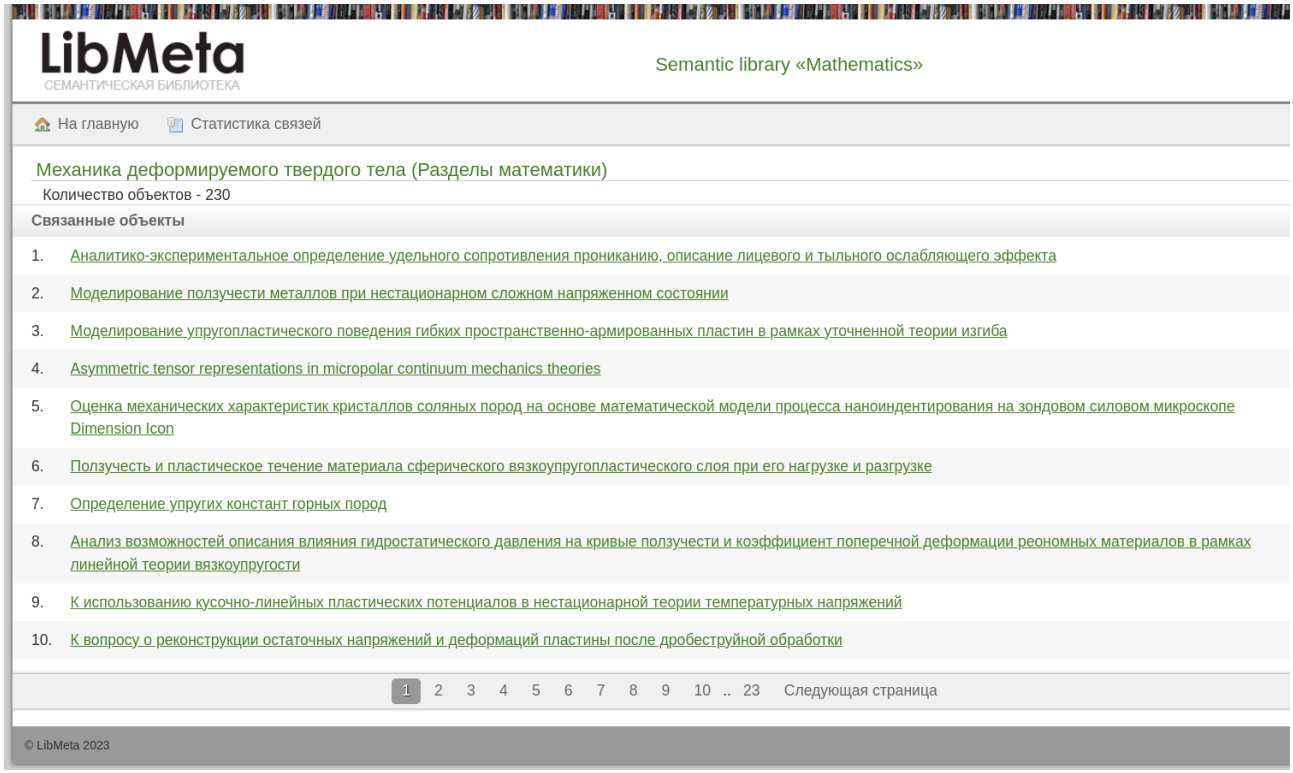


Рис. 11. Скрин 2. Рубрика журнала и ее публикации

В процессе исследований было получено следующее:

- анализ паспортов специальностей позволил выявить наличие общей лексики для различных рубрик журнала, а именно, описание *методов, решений, задач*, *результатов*;
- анализ контента журнала позволяет *наполнить эти термины «смыслом»*, характерным для каждой специальности конкретно в журнале;
- для тематик журнала были выявлены общие *типы задач*, которые в дальнейшем помогут более глубоко анализировать текст.

Последовательно были реализованы следующие этапы исследований.

Установлены и использованы четыре меры смыслового соответствия:

- эквивалентность рубрик по содержанию,
- включение рубрики ВАК в сопоставленную рубрику журнала,
- включение сопоставленной рубрики журнала в рубрику ВАК,
- значительное пересечение объёмов рубрик журнала.

Решены следующие задачи:

- анализ текстов статей и паспортов ВАК,
- извлечение сущностей и связей,
- поиск УДК,
- поиск схожих
  - персон,
  - публикаций,
- статистика,
- классификация по тематикам,
  - по специальностям,
  - по рубрикам,
- обзоры и сравнения статей.

На рисунке 12 представлен узел графа знаний «Публикация». Пунктиром выделены «стандартные» метаданные публикации. Сплошными стрелками указываются дополнительные связи, которые появились благодаря анализу рубрик и специальностей ВАК.

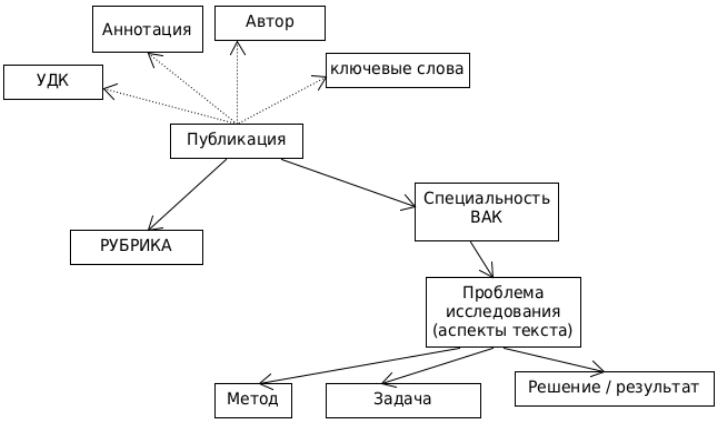


Рис.12. Узел ГЗ «Публикация» с новыми связями

Предложенная в статье методика построения графа знаний включает этапы подготовки данных, выбора методов, анализа результатов, достраивания онтологии. Целью этих этапов является выделение, структуризация и связывание информации по предметной области из сырых текстов, фиксируя ее в виде графа знаний предметной области.

Заключение. Навигация по различным узлам ГЗ позволяет осуществлять поиск, который отличается от классического, реализуемого в библиографических базах данных.

Применение методов обработки естественного языка к графам знаний журнала применимо для насыщения и развития ПрО журнала.

Получаем в перспективе:

- статистику для рецензирования и других целей;
- соответствие терминов ПрО, ВАК, рубрикам;
- подбор УДК и других классификаторов;

и другие элементы анализа статей журнала

Интеллектуальный анализ – единственный инструмент установления соответствия рубрик, создания онтологии разметки и ГЗ.

ПрО журнала, таким образом, будет представлена *задачами, методами, результатами и т.д.*, которым посвящены публикации, при этом вершинами графа ПрО будут рубрики и специальности ВАК.

Перемещаясь по вершинам графа, получаем множества публикаций, которые относятся к рубрике, и в которых изучается та или иная задача, относящаяся к определенной специальности ВАК. Полученный информационный ресурс отличается от того, как журнал представлен в интернете на основе библиографических данных.

Результат позволяет пользователю оценить*, нужна ли конкретная статья или нет, не скачивая, не читая целиком.*

Благодарности. Авторы выражают благодарность редакторскому коллективу журнала «Информационные и математические технологии в науке и управлении» за внимание к предмету исследования статьи, обсуждение и экспертный анализ результатов.

Работа представлена в рамках выполнения темы НИР «Математические методы анализа данных и прогнозирования» ФИЦ ИУ РАН, а также в рамках проекта ИСЭМ СО РАН, № темы FNEU-2021-0007, рег. № АААА-А21-121012090007-7.

Список источников

1. Tryon C. Cluster analysis. London. Ann Arbor Edwards Bros, 1939, 139 p.
2. Oyewole G.J., Thopil G.A. Data clustering: application and trends. Artif Intell Rev, 2023, 56, pp.6439–6475, DOI:10.1007/s10462-022-10325-y
3. Pitafi S., Anwar T., Sharif Z. A taxonomy of machine learning clustering algorithms, challenges, and future realms. Appl. Sci., 2023, 13, 3529, DOI: 10.3390/app13063529.
4. Xu Q., Gu H, Ji S. Text clustering based on pre-trained models and autoencoders. Front. Comput. Neurosci, 2024, 17, 1334436, DOI:10.3389/fncom.2023.1334436.
5. Alkaissi H., McFarlane S.I. Artificial Hallucinations in ChatGPT: Implications in scientific writing. Cureus, 2023, 15(2): e35179, DOI:10.7759/cureus.35179.
6. Probierz B., Kozak J., Hrabia A., Clustering of scientific articles using natural language processing. Procedia computer science, 2022, v. 207, 2022, pp. 3449-3458, DOI:10.1016/j.procs.2022.09.403.
7. Salton G., Wong A., Yang C.S. A vector space model for automatic indexing. 1975. Communications of the ACM, vol. 18, pp. 613-620, DOI:10.1145/361219.361220.
8. Mikolov T., Sutskever I., Chen K., et al. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, ArXiv, 2013, DOI:10.48550/arXiv.1310.4546
9. Mikolov T. Efficient estimation of word representations in vector space. 1st Int. Conf. Learn. Represent. ICLR 2013, Work. Track Proc., 2013, p. 1–12.
10. Turian J., Ratinov L., Bengio Y. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual meeting of the association for computational linguistics, ACL ’10. Association for computational linguistics, 2010.
11. Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов / К.В. Воронцов // Доклады РАН, 2014. – Т. 456. – № 3. – С. 268–271.
12. ruSciBench – бенчмарк для оценки эмбеддингов научных текстов. – URL: https://habr.com/ru/articles/781032/.
13. Huggingface. Availeble at: https://huggingface.co/mlsa-iai-msu-lab/sci-rus-tiny.
14. Alkaissi H., McFarlane S.I. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus, 2023, 15(2): e35179, DOI:10.7759/cureus.35179.

15. Cadeddu A., Chessa A., De Leo V. Vincenzo, et al. A comparative analysis of knowledge injection strategies for large language models in the domain. Engineering applications of Artificial Intelligence, 2024, vol.133, part B. 108166, DOI:10.1016/j.engappai.2024.108166.
16. Peng C., Xia F., Naseriparsa M. Osborne F. Knowledge graphs: opportunities and challenges. Available at: https://arxiv.org/pdf/2303.13948.
17. Ataeva O., Serebryakov V., Tuchkova N. Ontological approach to a knowledge graph construction in a semantic library. Lobachevskii journal of mathematics, 2023, vol. 44, no. 6, pp. 2229–2239, DOI:10.1134/S1995080223060471.
18. Ataeva O., Kornet Yu.N., Serebryakov V., Tuchkova N., Approach to creating a thesaurus and a knowledge graph of an applied subject area. Lobachevskii journal of mathematics, 2023, vol. 44, no. 7, pp. 2577–2586.
19. Blashfield R.K. The growth of cluster analysis: Tryon, Ward, And Johnson, Multivariate behavioral research, 1980, 15:4, pp. 439-458, DOI:10.1207/s15327906mbr1504_4.
20. Blei D.M. Probabilistic topic models. Communications of the ACM, 2012, vol. 55, no. 4, pp. 77–84.
21. Blei D.M., A.Y.Ng, Jordan M.I. Latent dirichlet allocation. Journal of machine learning research, 2003, 3, pp. 993-1022.
22. Ikotun A.M., Ezugwu A.E., Abualigah L., et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information sciences, 2023, vol. 622, pp. 178-210, DOI:10.1016/j.ins.2022.11.139, available at: https://www.sciencedirect.com/science/article/pii/S0020025522014633 (accessed: 10/13/2024)
23. Sparck J.K. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 1972, vol. 28, no. 1, pp. 11-21, DOI:10.1108/eb026526.
24. Chowdhary K.R. Natural language processing. Fundamentals of artificial intelligence, 2020, pp. 603-649.
25. Young T. et al. Recent trends in deep learning based natural language processing. IEEE Computational intelligence magazine, 2018, vol. 13, no. 3, pp. 55-75.

Атаева Ольга Муратовна. *Старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат техн. наук, специалист в области системного программирования и баз данных. ORCID: 0000-0003-0367-5575, oli.ataeva@gmail.com, 119333, Россия, Москва, ул. Вавилова, д.40*

Массель Людмила Васильевна. *Специалист в области информационных технологий и искусственного интеллекта, главный научный сотрудник ИСЭМ СО РАН, зав. отделом «Системы искусственного интеллекта в энергетике, главный редактор журнала «Информационные и математические технологии в науке и управлении», доктор техн. наук, профессор, заслуженный деятель науки РФ, ORCID: 0000-0002-9088-9012, massel@isem.irk.ru, 664033, Россия, Иркутск, Лермонтова, 130*

Серебряков Владимир Алексеевич. *Специалист в области теории формальных языков и её приложений, доктор физ.-мат. наук, профессор, зав. отделом Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН. Руководитель и участник разработки ряда известных программных проектов, в частности ИСИР и ИСИР РАН, Научный портал РАН, ORCID: 0000-0003-1423-621X, serebrvas@gmail.com, 119333, Россия, Москва, ул. Вавилова, д.40*

Тучкова Наталия Павловна. *Старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат физ.-мат. наук, специалист в области алгоритмических языков и информационных технологий, ORCID: 0000-0001-5357-9640, tuchkova.nataly@gmail.com, 119333, Россия, Москва, ул. Вавилова, д.40*

UDC 004.8+519.217 + 519.876.2
DOI:10.25729/ESI.2024.35.3.001

Data mining when constructing a knowledge graph of a multidisciplinary journal

Olga M. Ataeva[1], Ludmila V. Massel[2], Vladimir A. Serebryakov[1], Natalia P. Tuchkova[1]

[1]FRC CSC RAS, Russia, Moscow, *oli.ataeva@gmail.com*

[2]Melentiev Energy Systems Institute SB RAS, Russia, Irkutsk

Abstract. The paper explores the thematic diversity of the interdisciplinary journal. The purpose of the research is to build a knowledge graph of the journal for the thematic presentation and systematization of the electronic

archive and new publications of the journal. The initial data are journal articles devoted to various information and mathematical technologies in science and management, that is, interdisciplinary research. The systematization of texts using vector analysis methods is proposed. In the process of thematic analysis of the content of the journal, a division into headings is proposed, links of headings and articles with the corresponding descriptions of the specialties of the Higher Attestation Commission are established. To analyze the topic, an exploratory analysis of the source texts is used, then data mining methods are used. The results of the division are provided to the experts of the journal, after which a decision is made on the formation of a thematic heading and the inclusion of the specialties of the Higher Attestation Commission in it. The journal articles are integrated into the LibMeta semantic library, which is why the library's ontology is being completed and the journal's ontology is being formed, and the journal's knowledge graph is being built on this basis. A procedure for navigating through the content of the journal using the knowledge graph in the LibMeta semantic library is proposed, which can become the basis for information support of scientific research and the creation of a digital assistant in an interdisciplinary subject area. Examples are given for specific journal content, but the proposed technology can be extended to other journals, since most journals belonging to several specialties of the Higher Attestation Commission naturally capture several disciplines.

Keywords*:* knowledge graph, semantic library, ontology completion, clustering of scientific articles, text summarization

Acknowledgements: The work is presented within the framework of the research topic "Mathematical Methods of Data Analysis and Forecasting" of the FRC ICS RAS, as well as within the framework of the ISEM SB RAS project, topic No. FNEU-2021-0007, reg. No. AAAA-A21-121012090007-7.

References

1. Tryon C. Cluster analysis. London. Ann Arbor Edwards Bros, 1939, 139 p.
2. Oyewole G.J., Thopil G.A. Data clustering: application and trends. Artif Intell Rev, 2023, 56, pp.6439–6475, DOI:10.1007/s10462-022-10325-y
3. Pitafi S., Anwar T., Sharif Z. A taxonomy of machine learning clustering algorithms, challenges, and future realms. Appl. Sci., 2023, 13, 3529, DOI: 10.3390/app13063529.
4. Xu Q., Gu H, Ji S. Text clustering based on pre-trained models and autoencoders. Front. Comput. Neurosci, 2024, 17, 1334436, DOI:10.3389/fncom.2023.1334436.
5. Alkaissi H., McFarlane S.I. Artificial Hallucinations in ChatGPT: Implications in scientific writing. Cureus, 2023, 15(2): e35179, DOI:10.7759/cureus.35179.
6. Probierz B., Kozak J., Hrabia A., Clustering of scientific articles using natural language processing. Procedia computer science, 2022, v. 207, 2022, pp. 3449-3458, DOI:10.1016/j.procs.2022.09.403.
7. Salton G., Wong A., Yang C.S. A vector space model for automatic indexing. 1975. Communications of the ACM, vol. 18, pp. 613-620, DOI:10.1145/361219.361220.
8. Mikolov T., Sutskever I., Chen K., et al. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, ArXiv, 2013, DOI:10.48550/arXiv.1310.4546
9. Mikolov T. Efficient estimation of word representations in vector space. 1st Int. Conf. Learn. Represent. ICLR 2013, Work. Track Proc., 2013, p. 1–12.
10. Turian J., Ratinov L., Bengio Y. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual meeting of the association for computational linguistics, ACL ’10. Association for computational linguistics, 2010.
11. Vorontsov K.V. Additivnaya regulyarizatsiya tematicheskikh modeley kollektsiy tekstovykh dokumentov [Additive regularization of topic models of text document collections]. Doklady RAN [Reports of the RAS], 2014, vol. 456, no.3, pp. 268–271.
12. ruSciBench – бенчмарк для оценки эмбеддингов научных текстов. – URL: https://habr.com/ru/articles/781032/.
13. Huggingface. Availeble at: https://huggingface.co/mlsa-iai-msu-lab/sci-rus-tiny.
14. Alkaissi H., McFarlane S.I. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus, 2023, 15(2): e35179, DOI:10.7759/cureus.35179.
15. Cadeddu A., Chessa A., De Leo V. Vincenzo, et al. A comparative analysis of knowledge injection strategies for large language models in the domain. Engineering applications of Artificial Intelligence, 2024, vol.133, part B. 108166, DOI:10.1016/j.engappai.2024.108166.
16. Peng C., Xia F., Naseriparsa M. Osborne F. Knowledge graphs: opportunities and challenges. Available at: https://arxiv.org/pdf/2303.13948.
17. Ataeva O., Serebryakov V., Tuchkova N. Ontological approach to a knowledge graph construction in a semantic library. Lobachevskii journal of mathematics, 2023, vol. 44, no. 6, pp. 2229–2239, DOI:10.1134/S1995080223060471.

18. Ataeva O., Kornet Yu.N., Serebryakov V., Tuchkova N., Approach to creating a thesaurus and a knowledge graph of an applied subject area. Lobachevskii journal of mathematics, 2023, vol. 44 , no. 7, pp. 2577–2586.
19. Blashfield R.K. The growth of cluster analysis: Tryon, Ward, And Johnson, Multivariate behavioral research, 1980, 15:4, pp. 439-458, DOI:10.1207/s15327906mbr1504_4.
20. Blei D.M. Probabilistic topic models. Communications of the ACM, 2012, vol. 55, no. 4, pp. 77–84.
21. Blei D.M., A.Y.Ng, Jordan M.I. Latent dirichlet allocation. Journal of machine learning research, 2003, 3, pp. 993-1022.
22. Ikotun A.M., Ezugwu A.E., Abualigah L., et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information sciences, 2023, vol. 622, pp. 178-210, DOI:10.1016/j.ins.2022.11.139, available at: https://www.sciencedirect.com/science/article/pii/S0020025522014633 (accessed: 10/13/2024)
23. Sparck J.K. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 1972, vol. 28, no. 1, pp. 11-21, DOI:10.1108/eb026526.
24. Chowdhary K.R. Natural language processing. Fundamentals of artificial intelligence, 2020, pp. 603-649.
25. Young T. et al. Recent trends in deep learning based natural language processing. IEEE Computational intelligence magazine, 2018, vol. 13, no. 3, pp. 55-75.

Ataeva Olga Muratovna. *Senior researcher of the of Dorodnicyn computing center FRC SCS RAS, PhD, expert in the field of system programming and databases, oli.ataeva@gmail.com, ORCID: 0000-0003-0367-5575, 119333, Russia, Moscow, Vavilova, 40.*

Massel Liudmila Vasilievna. *Expert in the field of information technologies and artificial intelligence, chief researcher of the ESI SB RAS, head of the department of "Artificial Intelligence Systems in Energy sector", editor-in-chief of the journal "Information and mathematical technologies in science and management", Doctor of technical sciences, professor, honored scientist of the Russian Federation, massel@isem.irk.ru, ORCID: 0000-0002-9088-9012, 664033, Russia, Irkutsk, Lermontov, 130.*

Serebryakov Vladimir Alekseevich. *Expert in the field of theory of formal languages and its applications, doctor of sciences, professor, head of Dorodnicyn computing center FRC SCS RAS department, head and participant in the development of a number of well-known program projects, in particular ISIR and ISIR RAS, scientific portal RAS, ORCID: 0000-0003-1423-621X, serebrvas@gmail.com, 119333, Russia, Moscow, Vavilova, 40.*

Tuchkova Natalia Pavlovna. *Senior researcher of Dorodnicyn computing center FRC SCS RAS, PhD in physics with a math degree, graduated from CS Faculty of Lomonosov MSU, the expert in the field of algorithmic languages and information technologies, ORCID: 0000-0001-5357-9640, tuchkova.nataly@gmail.com, 119333, Russia, Moscow, Vavilova, 40.*

Статья поступила в редакцию 30.07.2024; одобрена после рецензирования 11.10.2024; принята к публикации 14.10.2024.

The article was submitted 07/30/2024; approved after reviewing 10/11/2024; accepted for publication 10/14/2024.