УДК 004.89

ПОДХОДЫ К ИЗВЛЕЧЕНИЮ ЗНАНИЙ В НАУЧНЫХ ПРЕДМЕТНЫХ ОБЛАСТЯХ[1]

Тучкова Наталия Павловна[*], Атаева Ольга Муратовна[**]
[*]natalia_tuchkova@mail.ru, к.ф.-м.н., с.н.с., ORCID [0000-0001-6518-5817]
[**]oli@ultimeta.ru, к.т.н., н.с., ORCID [0000-0003-0367-5575]
Вычислительный центр им. А.А. Дородницына
Федерального исследовательского центра «Информатика и управление»
Российской академии наук,
г. Москва, 119333, ул. Вавилова, 40

Аннотация. Изучается проблема извлечения знаний из разнородных оцифрованных данных. Дается обзор метрических характеристик, применимых к сравнению предметных областей по формальным признакам. Обсуждаются методы искусственного интеллекта, используемые для классификации информационных ресурсов по предметным областям, и области их применения. Приложения иллюстрируются на примере математических предметных областей.

Ключевые слова: структурирование данных, тезаурус предметной области, метрики, тезаурус ОДУ.

Цитирование: Тучкова Н. П., Атаева О. М Подходы к извлечению знаний в научных предметных областях // Информационные и математические технологии в науке и управлении. 2020. № 2 (18). С. 5-18. DOI: 10.38028/ESI.2020.18.2.001

1. Введение. Проблемы структурирования данных изучаются со времени создания первых энциклопедий, реферативных журналов, справочников и библиотек. Задачи возникали в связи с процессами *сохранения и извлечения знаний.* Их решение было всегда прерогативой специалистов предметных областей, экспертов, людей, энциклопедических знаний. Изучалось и изучается сейчас, что именно из многочисленных и разнородных данных сохранить, чтобы извлечь знания. В результате многовекового огромного труда человечество получало знания и сохраняло их в печатной форме. Для научных печатных изданий было *найдено решение в виде реферативных журналов, предметных, авторских указателей, тематических рубрикаторов и классификаторов* (ВИНИТИ, Черный, Гиляревский [11], [4] и др., Институт научной информации США, Гарфилд (Garfield E.) [18] и др.). Эти решения получили развитие в цифровую эпоху.

Цифровой век полностью перевернул представления об информации и знаниях, привел к переосмыслению публикационной и библиотечной деятельности [16]. Современное состояние проблемы представления знаний характеризуется тем, что информация и знания «перемещаются» в интернет. Естественно, что появились проблемы, связанные с представлением информации в цифровом виде и последующим ее извлечением. Текст, изданный в бумажном виде, извлекался в неизменном виде, «что написано пером, то не

[1] Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований, проект № 18-00-00297комфи

вырубишь топором». Теперь появилась возможность получения «обработанной» и структурированной информации.

Структурирование исходной информации стало самостоятельной проблемой информационных технологий. Задачи структурирования данных перешли в область интеллектуального анализа данных. Особое значение в извлечении полезной информации из накопленных цифровых ресурсов получило использование алгоритмов искусственного интеллекта, в том числе алгоритмов машинного обучения. Успехи эвристического подхода в программировании и использование алгоритмов искусственного интеллекта демонстрируют прорыв в области обучающих систем. Тем не менее, обозначились пределы применения этих методов - это оценка результата, поскольку она основывается на предварительной классификации. Классификация, по-прежнему, нуждается в экспертных знаниях, то есть, экспертные знания должны присутствовать в цифровом образе предметных областей для отражения научного подхода в их представлении и сохранения фундаментальных знаний в цифровую эпоху.

Сочетание искусственного и естественного интеллекта, как путь прохождения знаний от человека-эксперта в цифровое пространство и обратно к человеку-потребителю знаний основан на структурировании исходных данных.

2. Задачи, относящиеся к искусственному интеллекту. Один из авторов термина «искусственный интеллект» Джон Маккарти (John McCarthy) в 1956 г. сформулировал проблемы эвристического программирования, как обеспечение того, чтобы компьютерные программы решали *действительно сложные задачи*, а именно: *поиск, распознавание образов, обучение, планирование и индукция* [23]. В процессе реализации этой идеи были разработаны языки программирования, использующие доказательства утверждений LISP, PROLOG, SMALLTALK, РЕФАЛ и др.

В толковом словаре [1] известные отечественные авторы определили «искусственный интеллект», как «свойство интеллектуальных систем выполнять *творческие функции*, которые традиционно считаются прерогативой человека». Далее дается следующее определение *системы интеллектуальной* (С.И.): «техническая или программная система, способная решать задачи, традиционно считающиеся творческими, принадлежащие конкретной предметной области, знания о которой, хранятся в памяти С.И. Структура С.И. включает три основных блока - *базу знаний, решатель и интеллектуальный интерфейс*».

Ассоциация разработчиков искусственного интеллекта AAAI (Association for the Advancement of Artificial Intelligence), которую Джон Маккарти возглавлял в 1983-84 г., декларирует определение искусственного интеллекта как «*научное понимание механизмов, лежащих в основе мышления и разумного поведения, и их воплощение в машинах*» [20].

Авторитетные отечественные специалисты опираются также на определение искусственного интеллекта, как «одного из направлений информатики, целью которого является разработка программно-аппаратных средств, позволяющих пользователю-непрограммисту ставить и решать свои, традиционно считающиеся *интеллектуальными, задачи*, общаясь с ЭВМ на ограниченном подмножестве естественного языка» [3].

Из приведенных определений видно, что ключевое слово «интеллект», трактуется в сообществе специалистов, начиная с 50-х годов XX в., как *программа*, а слово *искусственный* является в этом контексте общепризнанным уточнением, что не

противоречит сути *компьютерной программы,* которая, собственно, и выполняет *интеллектуальные задачи, переданные ей естественным интеллектом программиста*.

Искусственный интеллект и машинное обучение в последнее время были весьма успешными во многих практических приложениях (например, распознавание речи, распознавание лиц, автономное вождение, системы восстановления, классификация изображений, обработка естественного языка, автоматизированная диагностика и др.) [19]. С распространением цифровизации, когда количество накопленной оцифрованной информации позволяет уже только машинную обработку, применение методов искусственного интеллекта стало необходимостью для анализа научных данных, в том числе библиографических.

Методы интеллектуального анализа данных включают использование различных метрик на основе статистических данных, эвристические и итерационные алгоритмы, алгоритмы машинного обучения, семантический анализ, алгоритмы распознавания – целый арсенал математических алгоритмов.

3. Ресурсы и метрики. Особенность цифровых данных заключается в том, что они изначально *неизбежно как-то структурированы.* Систематизация цифровых данных происходит практически одновременно с накоплением (иногда присутствуют стихийные, сиюминутные решения). Интеграция и обновление данных составляют самостоятельные задачи, так как это связано с дублированием данных и установлением принадлежности (авторства) информации. Следствием перечисленных особенностей является *затруднение получения достоверной, актуальной информации* и *извлечение из нее знаний* для удовлетворения информационной потребности пользователя. Возникают вопросы выбора поисковых систем, баз данных, критериев поиска и пр. В целом – это проблемы *сохранения и извлечения* истинных и достоверных сведений о природе и человеке. В то же время известно, что рост научных публикаций сохраняет экспоненциальный характер и для научной работы надо охватывать этот поток данных. Поэтому одним из приоритетных направлений накопления и структурирования цифровых данных продолжает оставаться *улучшение поиска с использованием метрик.*

Для оценки данных и поиска в информационных системах используются:

- *наукометрические характеристики* в разделах коллекций научных публикаций для оценки публикационной активности [10];
- *метрики квалиметрии жизни* в коллекциях статистических, медицинских и географических данных для оценки качества жизни [24];
- *метрики рекламного потенциала* в социальных сетях для оценки потребностей пользователей и маркетинга [2];
- *метрики, связанные с профессиональным профилем,* в специальных информационных бизнес-ресурсах.

В качестве одного из примеров современного подхода сохранения и извлечения знаний можно отметить проект Techopedia[2], где накапливается и систематизируется массив научных и коммерческих публикаций и проектов по различным разделам информационных технологий. В этом проекте в основе систематизации данных используются словари предметных областей.

[2] https://www.techopedia.com/definition/1181/data-mining

Отметим, что «пространства знаний», открытые и закрытые энциклопедии [13], построенные на онтологиях и тезаурусах предметных областей, *практически не используют метрики* посещений и цитирования.

На рис. 1 приведены некоторые примеры популярных наукометрических показателей в коллекциях (базах данных) публикаций. Основные показатели: цитируемость публикации (выше - лучше), Н-индекс автора (больше - лучше), квартиль журнала (Q1 – лучше, чем Q4).

- https://www.researchgate.net/ для каждой публикации рассчитываются показатели: Research Interest, Citations, Recommendations, Reads
- https://www.elibrary.ru/kand_ras_2019.asp для кандидатов на выборы в РАН 2019 рассчитаны показатели: число публикаций, цитирований, входящих в ядро РИНЦ, индекс Хирша по ядру РИНЦ
- https://istina.msu.ru/home/reports/ Разделы Мои отчеты, Моя страница
- http://www.mathnet.ru/ раздел Персоналии, Статистика Публикацийи Просмотров в MathSciNet, в zbMATH, в Web of Science, в Scopus
- https://publons.com/researcher Web of Science, статистика и графика
- https://www.scopus.com/Scopus статистика и графика
- https://www.scimagojr.com/ Scimago Journal & Country Rank

Рис. 1. Примеры наукометрических показателей в базах данных публикаций

С точки зрения сохранения и извлечения знаний совершенно неважно, из каких источников поступает информация, имеет значение только то, насколько научное экспертное сообщество вовлечено в процесс оценки качества получаемых знаний. В этом смысле особенно интересны метрики в науке и в приложениях, которые используют системы искусственного интеллекта [27].

Большинство современных подходов извлечения знаний и их структуризации заключается в оптимизации метрик, однако, если измерение становится целью, то оно перестает быть измерением. Это замечание о том влиянии, которое оказывает целеполагание, основанное на достижении некоторых метрических показателей. В каждом измерении есть определенная потребность. Получение количественной характеристики становится целью для процесса или явления, когда качественная характеристика неясна и субъективна, что в полной мере можно отнести к наукометрическим показателям.

4. Извлечение знаний. Термин «знание» обрел новое, по сравнению с доцифровым, значение, непосредственно связанное с понятием искусственного интеллекта. В классическом отечественном учебнике по искусственному интеллекту [3] есть следующие определения данных и знаний: «данные – это отдельные факты, характеризующие объекты, процессы и явления предметной области, а также их свойства»; «знания – это закономерности предметной области (принципы, связи, законы), полученные в результате практической деятельности и профессионального опыта, позволяющие специалистам ставить и решать задачи в этой области»; «знания – это хорошо структурированные данные».

Извлечение знаний человеком – процесс исключительно субъективный и зависит от уровня образования и целеполагания. Не случайно классики информационных технологий

ввели специальный термин «тезаурус адресата» [12], чтобы обозначить характер информационных потребностей пользователя в информационной среде. Извлечение знаний может трактоваться как «общение эксперта с источником знаний» [3]. Исходя из приведенных определений, информационные ресурсы, производящие знания, реализуют функции создания *структур предметных областей*.

Появление в 2003 году онтологий, как технологии учета семантических связей [26], стало поводом для многочисленных проектов представления данных. Это позволило решить множество проблем искусственного интеллекта, повлияло на решение проблем поиска, дублирования данных, верификацию, установления авторства и др. [21]. Конечно, остались и неразрешенные проблемы, например, идентификация объектов и авторов, хотя много было достигнуто в этом направлении.

На основе структурированных данных и извлеченных из них знаний можно сформировать инструменты сравнения текстов для идентификации авторов и предметных областей, используя различные меры, такие, как:

- меру сходства статей (для разных авторов),
- меру пересечения предметных областей,
- меру близости публикации той или иной предметной области,
- меры связанности терминов и т.д.

В результате можно оценить, получен новый результат или переписан старый, т.е. применить перечисленные инструменты именно для:

- сохранения приоритетов в науке;
- доказательства достоверности результатов и фактов;
- структурирования знаний и предметных областей;
- сохранения и извлечения знаний.

Также, имея терминологическое описание предметной области в виде тезауруса, можно «научить» тезаурус автоматически расширяться на основе полученных новых связей [18], [19].

5. Роль и задачи машинного обучения. Способы извлечения знаний в самом общем виде можно сгруппировать как: визуальные (зрительные), основанные на когнитивной графике графы, карты и пр. [25]; основанные на анализе семантических связей [5]; эмпирические, основанные на экспертном знании предметной области [6]; основанные на интеллектуальном анализе данных [15]. Все перечисленные способы задействованы в современных информационных системах. Учет семантических связей стал возможен только на современном этапе развития технологий онтологического моделирования, а интеллектуальный анализ – с развитием методов машинного обучения.

Основной проблемой при разработке тезаурусов и онтологий предметной области изначально является проблема приобретения знаний. Эффективным способом оказывается совмещение сбора знаний в виде онтологий и тезаурусов с методами машинного обучения. С их помощью решаются как задачи извлечения знаний, так и задачи их обновления. Предоставляются различные методы решения этих проблем.

Для повышения эффективности поиска знаний используются различные алгоритмы с построением метрической конфигурации пространства и использованием различных мер. Для извлечения знаний используются так называемые интеллектуальные алгоритмы (*Data mining часто понимают как синоним data extraction)* (см., например:

http://www.machinelearning.ru). Среди них методы выделения ключевых слов: частотный алгоритм (*Rapid Automatic Keyword Extraction algorithm*) (см., например: https://medium.com/datadriveninvestor/rake-rapid-automatic-keyword-extraction-algorithm-f4ec17b2886c), метод латентных семантик: *Latent semantic analysis, LSA, latent semantic indexing LSI – учет контента ключевых слов* (см., например: https://medium.com/acing-ai/what-is-latent-semantic-analysis-lsa-4d3e2d18417a*),* векторные методы, основанные на контекстном представлении документа (*Doc2vec*) (см., например: https://radimrehurek.com/gensim/models/doc2vec.html) и т.д. На результатах работ этих алгоритмов выводятся основные характеристики пространства знаний отдельной предметной области.

С помощью методов машинного обучения решаются задачи:

- извлечение метаданных из данных;
- слияние и отображение онтологий и тезаурусов путем анализа концептов;
- поддержка онтологий на основе анализа данных экземпляров;
- улучшение работы семантических приложений путем наблюдений за пользователями.

Для этого могут использоваться:

- кластеризация;
- инкрементное обновление онтологии и тезауруса;
- разработка и улучшение больших онтологий естественного языка
- «обучение» онтологий и тезаурусов на информационных ресурсах предметной области

5.1. Подготовка данных для извлечения знаний о предметной области. Применение алгоритмов машинного обучения к коллекциям научных публикаций облегчается тем, что изначально научный текст уже структурирован. Тем не менее, создание структуры из набора данных требует предварительной обработки даже такой «хорошей» информации, как научная публикация.

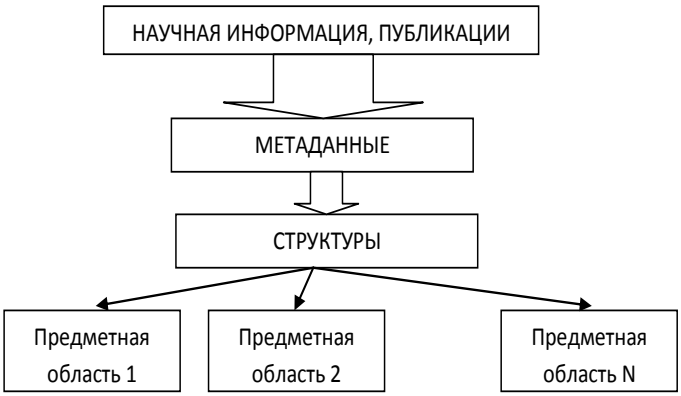

Рис. 2. Схема подготовки данных для обобщенной модели предметных областей

На рис. 2 представлена схематически процедура анализа массива публикаций, для последующего индексирования принадлежности к определенной предметной области. Эта последовательность реализована в цифровой библиотеке LibMeta [13] для формирования пространства знаний для некоторого множества предметных областей.

На рис. 3 показана структура связей, реализованная в классической математической энциклопедии [22]. Понятия этой библиотеки для раздела «обыкновенные

дифференциальные уравнения» (ОДУ) [7] связаны семантически с объектами LibMeta, что проиллюстрировано на рис. 4.

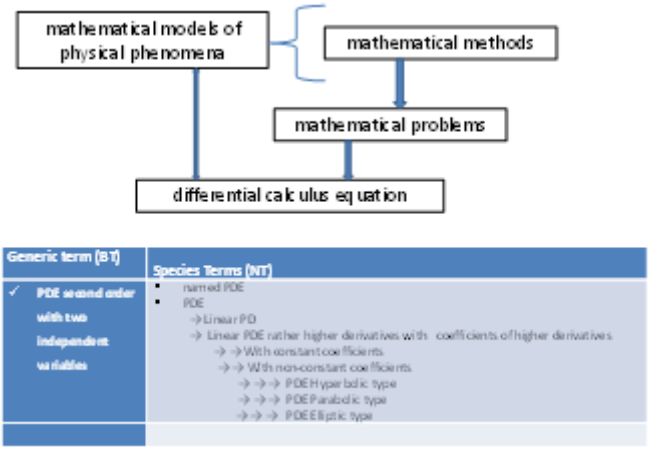

Рис. 3. Пример организации семантических связей в математической энциклопедии

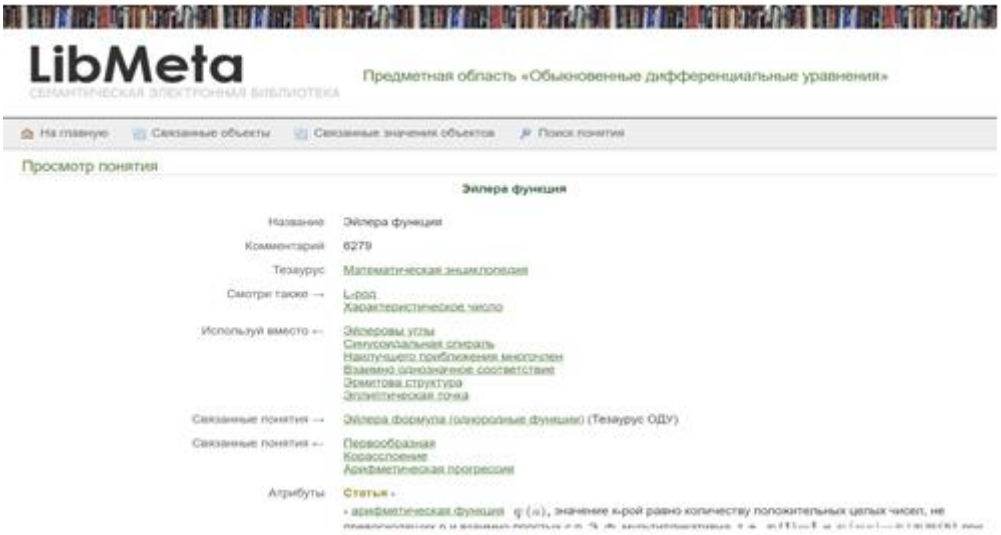

Рис. 4. Связи понятий тезауруса ОДУ LibMeta с понятиями математической энциклопедии

На рис. 5 показано, как устроены структуры данных на уровне связей между терминами предметной области и ее публикациями в семантической библиотеке LibMeta и коммерческом ресурсе Techopedia. Для сравнения выбран проект Techopedia, как наиболее развитый в разделе структурирования знаний на основе публикаций предметных областей информационных технологий. За основу представления данных предметной области в Techopedia выбран словарь (Dictionary), а в библиотеке LibMeta – тезаурус, более сложная структура, реализованная в онтологии предметных областей.

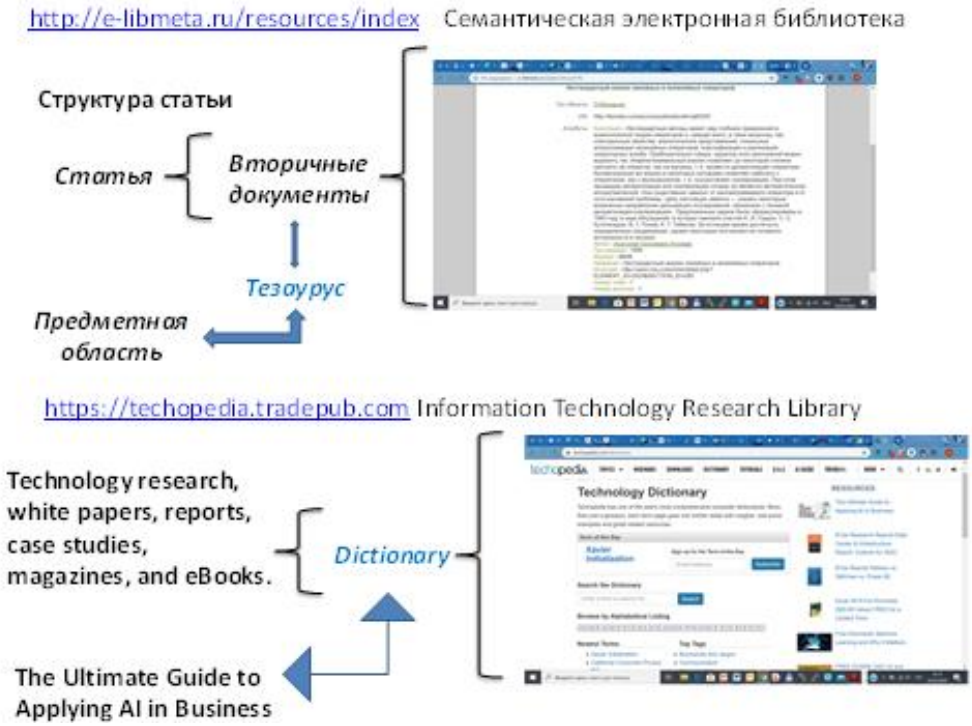

Рис. 5. Сравнение структур данных ресурсов LibMeta и Techopedia

5.2. Учет экспертного мнения. Проблема учета экспертного мнения решается с применением понятия тезауруса адресата-индивидуума, введенного в информатику Ю.А. Шрейдером [12]. В библиотеке LibMeta реализована технология составления *тезауруса предметной области автора*. Это позволяет учесть специфику авторского подхода, принадлежность определенной школе, вклад автора в предметную область, а не только публикационную активность, задаваемую наукометрическими показателями. Таким образом, экспертная позиция автора отражается через его публикации и связи с другими понятиями предметной области.

Состав данных для тезауруса адресата:

- частотный словарь индивидуума;
- варианты сочетаний терминов;
- контексты частотных терминов;
- специальные обозначения и формулы.

5.3. Структурирование предметной области. Описание предметной области основывается на терминологическом словаре со связями – тезаурусе [8, 14]. Обновление тезауруса происходит благодаря связям с вновь поступившими публикациями. Расширение тезауруса предметной области допускается путем установления связей с авторскими словарями и тезаурусами предметных областей авторов.

Семантическая поддержка понятий предметной области ОДУ, которая отвечает за классификацию ресурсов, таких как: события, теоремы, персоны, публикации, формулы, включает в себя тезаурус ОДУ, классификатор MSC, классификатор УДК. На основе тезауруса ОДУ и связанных с ним классификаторов были выявлены связи между классификаторами MSC, УДК, понятиями математической энциклопедии и формулами, составлен список ключевых слов связанных с понятиями тезауруса и формулами.

На рис. 6 приведен фрагмент тезауруса, реализованного в библиотеке LibMeta, согласно подходу о структурированном представлении данных предметной области.

Рис. 6. Тезаурус ОДУ, пример реализации.

Аналогичная структура данных, основанная на терминологическом описании предметных областей, прошла апробацию для других предметных областей, в частности отдельно реализован проект для математической энциклопедии.

6. Заключение и выводы. Сохранение данных и знаний в цифровом виде в современных масштабах вызвало необходимость нового подхода к анализу больших данных. Эти подходы основаны на создании и анализе семантических структур в виде онтологий и алгоритмов машинного обучения. Тем не менее, характеристики профессиональной научной деятельности предлагается оценивать на анализе метрик, которые не относятся к характеристикам знаний. Для автоматизации процесса оценки знаний необходимо внести механизмы экспертных оценок специалистов предметных областей, совместить экспертные системы с системами интеллектуального анализа данных, создавать интегрированные системы.

История интегрированных систем начиналась с простых диалоговых систем «компьютер-эксперт» и продолжается на современном цифровом этапе, сохраняя те же тенденции, но на основе новых технологий – структурирование данных с использованием семантических связей, обучающие и обучающиеся программы, визуализация данных и когнитивная графика. Это позволяет организовать в цифровом виде энциклопедические данные и реализовать модели предметных областей. Вопрос метрических оценок таких ресурсов переходит в плоскость получения объемов информации и связей. Метрики по-прежнему несут количественные характеристики и не дают качественных оценок накопленных знаний. Способность информационной системы вырабатывать правильное решение и оценивать его эффективность является ключом к успеху в ее эксплуатации и только в этом контексте метрики могут составлять полезное свойство системы.

Авторские разработки, изложенные в статье, соответствуют тенденциям развития информационных технологий. В частности, используя совокупность предложенных методов, можно сформировать технологию обработки информации для *вновь поступающих данных* в семантическую библиотеку, которая позволяет извлекать структурированные данные (для

научных приложений). Дальнейшие разработки направлены на получение качественных оценок публикаций в рамках предметных областей. Использование онтологий и тезаурусов предметных областей направлено на внедрение методов сравнения и получения качественных оценок при анализе знаний извлекаемых из ее ресурсов.

СПИСОК ЛИТЕРАТУРЫ

1. Аверкин А. Н., Гаазе-Рапопорт М. Г., Поспелов Д. А. Толковый словарь по искусственному интеллекту. М.: Радио и связь. 1992. 256 с. http://www.raai.org/library/tolk/aivoc.html#L208. (доступно 15.08.2020)
2. Бродовская Е. В. Цифровые граждане, цифровое общество и цифровая гражданственность // Власть. 2019. Т. 27. № 4. С. 65-69. DOI: https://doi.org/10.31171/vlast.v27i4.6587. (доступно 15.08.2020)
3. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб.: Питер, 2000. 384 с.
4. Гиляревский Р.С., Шапкин А.В., Белоозеров В.Н. Рубрикатор как инструмент информационной навигации. СПб.: Профессия. 2008. 352 с.
5. Ильин Н., Киселев С., Рябышкин В., Танков С. Технологии извлечения знаний из текста // Открытые системы. СУБД. 2006 № 06 https://www.osp.ru/os/2006/06/2700556/. (доступно 15.08.2020)
6. Курмашин А**.** Полтора века - от таблицы Менделеева к Периодической системе // Наука и жизнь. 2019. № 9. С. 71-80.
7. Моисеев Е.И., Муромский А.А., Тучкова Н.П. Тезаурус информационно-поисковый по предметной области: обыкновенные дифференциальные уравнения. М.:МАКС Пресс. 2005. 116 с.
8. Муромский А.А., Тучкова Н.П. Представление математических понятий в онтологии научных знаний // Онтология проектирования. 2019. Т. 9. № 1(31). С. 50- 69. DOI: 10.18287/2223-9537-2019-9-1-50-69. (доступно 15.08.2020)
9. Тучкова Н.П. Роль и возможности специализированных тезаурусов в когнитивных технологиях // Информационные и математические технологии в науке и управлении. 2019. № 1 (13). С. 5-15. DOI: 10.25729/2413-0133-2019-1-01
10. Цыганов А. В. Краткое описание наукометрических показателей, основанных на цитируемости // Управление большими системами. Специальный выпуск 44: «Наукометрия и экспертиза в управлении наукой». 2013. С. 8-13. http://ubs.mtas.ru/archive/search_results_new.php?publication_id=19061. (доступно 15.08.2020)
11. Черный А.И. Введение в теорию информационного поиска. М.: Наука, 1975. 238с.
12. Шрейдер Ю. А. Тезаурусы в информатике и теоретической семантике // Научно-техническая информация. Сер. 2. 1971. № 3. С. 21-24.
13. Ataeva O.M., Sererbryakov V.A., Tuchkova N.P. Query Expansion Method Application for Searching in Mathematical Subject Domains // CEUR Workshop Proceedings, M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen (Aachen, Germany). Vol. 2543, pp. 38-48.

14. Ataeva O.M., Sererbryakov V.A., Tuchkova N.P. Mathematical Physics Branches: Identifying Mixed Type Equations // Lobachevskij Journal of Mathematics. 2019. Vol. 40. № 7. Pp. 876–886. DOI: 10.1134/S1995080219070047. (доступно 15.08.2020)
15. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018. Hamburg, Germany. August 27–30. 2018. Proceedings In book: Machine Learning and Knowledge Extraction 10.1007/978-3-319-99740-7_1. (доступно 15.08.2020)
16. Eberhart G. Redefining the Library in the Digital Age, https://www.britannica.com/topic/Redefining-the-Library-in-the-Digital-Age-1369994. (доступно 15.08.2020)
17. Encyclopedia of Mathematics. https://encyclopediaofmath.org/wiki/Main_Page. (доступно 15.08.2020)
18. Garfield E. Citation Indexes for Science // Science. 1955. Vol. 122. № 3159. Pp. 108–111.
19. Goebel R. et al. Explainable AI: the new 42. In: Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2018. Lecture Notes in Computer Science. Vol. 11015. Springer, Cham. 2018. https://link.springer.com/chapter/10.1007/978-3-319-99740-7_21. (доступно 15.08.2020)
20. Information about AI from the News, Publications, and Conferences https://www.sciencedaily.com/news/computers_math/artificial_intelligence/. (доступно 15.08.2020)
21. Leydesdorff L., Carley S., Rafols I. Global maps of science based on the new Web-of-Science categories // Scientometrics. 2013. Vol. 94 Issue 2. Pp. 589-593. https://doi.org/10.1007/s11192-012-0784-8. (доступно 15.08.2020)
22. Mathematical encyclopedia. Chief Editor. I.M.Vinogradov. M.: Soviet encyclopedia. 1979. 1104 p.)
23. McCarthy J. Artificial intelligence, logic and formalizing common sense. 1990 http://jmc.stanford.edu/articles/ailogic/ailogic.pdf.
24. Quality of Life Index. NUMBEO https://www.numbeo.com/quality-of-life/indices_explained.jsp. (доступно 15.08.2020)
25. Scientific Visualization: The Visual Extraction of Knowledge from Data. Eds: Georges-Pierre Bonneau, Thomas Ertl, Gregory M. Nielson. Springer. Mathematics and Visualization (Series). 2006 https://link.springer.com/content/pdf/10.1007%2F3-540-30790-7.pdf. (доступно 15.08.2020)
26. Semantic Web https://www.w3.org/standards/semanticweb/. (доступно 15.08.2020)
27. The problem with metrics is a big problem for AI. https://www.fast.ai/2019/09/24/metrics/. (доступно 15.08.2020)
28. Tuchkova N.P. Intellectual Contribution of Specialized Thesauruses to Cognitive Technologies // VIth International Workshop 'Critical Infrastructures: Contingency Management, Intelligent Agent-Based, Cloud Computing and Cyber Security' (IWCI 2019). Advances in Intelligent Systems Research. 2019. Vol. 169. https://www.atlantis-press.com/proceedings/iwci-19/125917328. DOI: 10.2991/iwci-19.2019.36. (доступно 15.08.2020)

UDK 004.89

APPROACHES TO KNOWLEDGE EXTRACTION IN SCIENTIFIC SUBJECT DOMAINS

Natalia P. Tuchkova[*], Olga M. Ataeva[**]

[*]natalia_tuchkova@mail.ru, PhD., senior researcher, ORCID [0000-0001-6518-5817]

[**]oli@ultimeta.ru, PhD., researcher, ORCID [0000-0003-0367-5575]

Dorodnicyn Computing Center of Federal Research Center "Informatics and Control" of Russian Academy of Science,
Vavilov st. 40, 119333 Moscow, Russia

Abstract. The studying is focusing to the problem of extracting knowledge from heterogeneous digital data. An overview of metric characteristics applicable to the comparison of subject areas by formal criteria is given. Methods of artificial intelligence used for the classification of information resources by subject domains and areas of application are presented. Applications are illustrated with mathematical subject domains.

Keywords: data structuring, domain thesaurus, metrics, ODE thesaurus.

References

1. Averkin A. N., Gaaze-Rapoport M. G., Pospelov D. A. Tolkovyj slovar' po iskusstvennomu intellektu. M.:Radio i svyaz' [Explanatory dictionary of artificial Intelligence]. 1992. 256 p. Available at: http://www.raai.org/library/tolk/aivoc.html#L208. (in Russian) (accessed 15.08.2020)
2. Brodovskaya E. V. Cifrovye grazhdane, cifrovoe obshchestvo i cifrovaya grazhdanstvennost' [Digital citizens, digital society and digital citizenship] // Vlast. 2019. V. 27. № 4. Pp. 65-69. DOI: https://doi.org/10.31171/vlast.v27i4.6587. (in Russian) (accessed 15.08.2020)
3. Gavrilova T.A., Horoshevskij V.F. Bazy znanij intellektual'nyh sistem.[Knowledge bases of intelligent systems] SPb.: Piter. 2000. 384 p. (in Russian)
4. Gilyarevskij R.S., SHapkin A.V., Beloozerov V.N. Rubrikator kak instrument informacionnoj navigacii.[Rubricator as a tool for information navigation] Spb.: Proffesija. 2008. 352 p. (in Russian)
5. Il'in N., Kiselev S., Ryabyshkin V., Tankov S. Tekhnologii izvlecheniya znanij iz teksta // Otkrytye sistemy. [Technologies for extracting knowledge from text] SUBD. 2006. № 06 Available at: https://www.osp.ru/os/2006/06/2700556/. (in Russian) (accessed 15.08.2020)
6. Kurmashin A. Poltora veka - ot tablicy Mendeleeva k Periodicheskoj sisteme [A century and a half - from the periodic table to the Periodic table] // Nauka i zhizn'. 2019. № 9. Pp. 71-80. (in Russian)
7. Moiseev E.I., Muromskij A.A., Tuchkova N.P. Tezaurus informacionno-poiskovyj po predmetnoj oblasti: obyknovennye differencial'nye uravneniya. [Information retrieval thesaurus in the subject area: ordinary differential equations] M.:MAKS Press. 2005. 116 p. (in Russian)
8. Muromskij A.A., Tuchkova N.P. Predstavlenie matematicheskih ponyatij v ontologii nauchnyh znanij [Representation of mathematical concepts in the ontology of scientific

knowledge]// Ontologiya proektirovaniya. 2019. Vol. 9. № 1(31). Pp. 50-69. DOI: 10.18287/2223-9537-2019-9-1-50-69. (in Russian)

9. Tuchkova N.P. Rol' i vozmozhnosti specializirovannyh tezaurusov v kognitivnyh tekhnologiyah [The role and capabilities of specialized thesauri in cognitive technologies] // Informacionnye i matematicheskie tekhnologii v nauke i upravlenii. 2019. № 1(13). Pp. 5-15. DOI: 10.25729/2413-0133-2019-1-01. (in Russian)
10. Cyganov A. V. Kratkoe opisanie naukometricheskih pokazatelej, osnovannyh na citiruemosti [Brief description of scientometric indicators based on citation] // Upravlenie bol'shimi sistemami. Special'nyj vypusk 44: «Naukometriya i ekspertiza v upravlenii naukoj», 2013. Pp. 8-13. Available at: http://ubs.mtas.ru/archive/search_results_new.php?publication_id=19061 (in Russian) (accessed 15.08.2020)
11. Chernyj A.I. Vvedenie v teoriyu informacionnogo poiska. Chernyj A.I. [Vvedenie v teoriyu informacionnogo poiska]. M.: Nauka. 1975. 238 p. (in Russian)
12. Shrejder YU. A. Tezaurusy v informatike i teoreticheskoj semantike [Thesauri in informatics and theoretical semantics] // Nauchno-tekhnicheskaya informaciya. Seria. 2. 1971. № 3. Pp. 21-24. (in Russian)
13. Ataeva O.M., Sererbryakov V.A., Tuchkova N.P. Query Expansion Method Application for Searching in Mathematical Subject Domains // CEUR Workshop Proceedings, M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen (Aachen, Germany). Vol. 2543. Pp. 38-48.
14. Ataeva O.M., Sererbryakov V.A., Tuchkova N.P. Mathematical Physics Branches: Identifying Mixed Type Equations // Lobachevskij Journal of Mathematics. 2019. Vol. 40. №7. Pp. 876–886. DOI: 10.1134/S1995080219070047.
15. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings In book: Machine Learning and Knowledge Extraction 10.1007/978-3-319-99740-7_1.
16. Eberhart G. Redefining the Library in the Digital Age. Available at: https://www.britannica.com/topic/Redefining-the-Library-in-the-Digital-Age-1369994. (accessed 15.08.2020)
17. Encyclopedia of Mathematics. Available at: https://encyclopediaofmath.org/wiki/Main_Page. (accessed 15.08.2020)
18. Garfield E. Citation Indexes for Science // Science. 1955. Vol. 122. № 3159. Pp. 108–111.
19. Goebel R. et al. Explainable AI: the new 42? In: Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2018. Lecture Notes in Computer Science. Vol 11015. Springer, Cham. 2018. Available at: https://link.springer.com/chapter/10.1007/978-3-319-99740-7_21. (accessed 15.08.2020)
20. Information about AI from the News, Publications, and Conferences. Available at: https://www.sciencedaily.com/news/computers_math/artificial_intelligence/ (accessed 15.08.2020)

21. Leydesdorff L., Carley S., Rafols I. Global maps of science based on the new Web-of-Science categories // Scientometrics. 2013. Vol. 94. Issue 2. Pp. 589-593. https://doi.org/10.1007/s11192-012-0784-8.
22. Mathematical encyclopedia. Chief Editor. I.M.Vinogradov. M.: Soviet encyclopedia. 1979. 1104 p.)
23. McCarthy J. Artificial intelligence, logic and formalizing common sense. 1990. Available at: http://jmc.stanford.edu/articles/ailogic/ailogic.pdf. (accessed 15.08.2020)
24. Quality of Life Index. NUMBEO. Available at: https://www.numbeo.com/quality-of-life/indices_explained.jsp. (accessed 15.08.2020)
25. Scientific Visualization: The Visual Extraction of Knowledge from Data. Eds: Georges-Pierre Bonneau, Thomas Ertl, Gregory M. Nielson. Springer. Mathematics and Visualization (Series), 2006. Available at: https://link.springer.com/content/pdf/10.1007%2F3-540-30790-7.pdf. (accessed 15.08.2020)
26. Semantic Web. Available at: https://www.w3.org/standards/semanticweb/. (accessed 15.08.2020)
27. The problem with metrics is a big problem for AI. Available at: https://www.fast.ai/2019/09/24/metrics/ (accessed 15.08.2020)
28. Tuchkova N.P. Intellectual Contribution of Specialized Thesauruses to Cognitive Technologies // VIth International Workshop 'Critical Infrastructures: Contingency Management, Intelligent Agent-Based, Cloud Computing and Cyber Security' (IWCI 2019). Advances in Intelligent Systems Research. 2019. Vol. 169. Available at: https://www.atlantis-press.com/proceedings/iwci-19/125917328. DOI: 10.2991/iwci-19.2019.36. (accessed 15.08.2020)