УДК 004.65 + 005 + 001.5

ФОРМИРОВАНИЕ РАСШИРЕННЫХ ПОИСКОВЫХ ЗАПРОСОВ НА ОСНОВЕ ТЕЗАУРУСА ПРЕДМЕТНОЙ ОБЛАСТИ В ОНТОЛОГИИ ЗНАНИЙ СЕМАНТИЧЕСКОЙ БИБЛИОТЕКИ

О. М. Атаева[1], В. А. Серебряков[2], Н. П. Тучкова[3]

[1,2, 3]*Вычислительный центр им. А.А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Аннотация

Обсуждены возможности расширения поискового запроса при наличии тезауруса предметной области. Роль контекста, задаваемого связями терминов тезауруса, заключается как в уточнении запроса, так и в увеличении масштабов выборки по запросу. Особое значение процесс расширения запроса имеет для научных предметных областей, где поиск опирается на специальную терминологию. В этом случае необходимо использовать тезаурусы предметных областей, чтобы минимизировать появление информационного шума. Предлагаемый подход позволяет учитывать особенности применения аналогичной терминологии в различных предметных областях. Примеры использования тезауруса отдельных разделов уравнений математической физики и смежных областей демонстрируют эффективность выбранного подхода исследований. Благодаря связям с понятиями информационных ресурсов других областей знаний, расширение информационного запроса захватывает поисковые поля отдаленных предметных областей и различных типов данных, текстов, символьных, звуковых и видеоархивов. Исследования показали, что расширение запроса на основе семантики контекста улучшает качество поиска научных публикаций в цифровой информации и повышает эффективность научных междисциплинарных исследований.

Ключевые слова*: сравнение научных текстов, семантический поиск, тезаурус для онтологии знаний*, *информационный запрос с помощью тезауруса, семантические библиотеки.*

ВВЕДЕНИЕ

Исследования в области использования терминов тезауруса в поисковом запросе в контексте повышения эффективности информационных систем ведутся в различных коллективах. Известны многочисленные разработки для приложений в автоматическом реферировании, системах перевода специализированных текстов, обучающих системах и др. Важность и актуальность этих исследований определяются возрастающим потоком цифровых данных и разнообразием их типов, необходимостью работы с научной информацией, особенностями предметных областей (ПО) в междисциплинарных исследованиях. Особый круг задач рассматривается в проблеме *расширения поискового запроса*. Средства расширения запроса позволяют *уточнять запрос* с помощью подсказок пользователю, сужая поле поиска с помощью дескрипторов тезаурусов, и использовать имеющиеся связи терминов (синонимов, аббревиатур и т. д.), *увеличивая поле поиска* и получая тем самым дополнительный информационный шум. Эти два процесса находятся в противоречии, но в итоге приводят к получению пертинентного результата, то есть удовлетворяющего информационный запрос пользователя. Разработки в этом направлении ведутся довольно давно, и многие информационные системы допускают расширение запроса. В работе [1] приведены результаты, свидетельствующие, что *привлечение синонимов* из базы WordNet, *не связанных с контекстом*, не улучшают качество информационного запроса. И только привлечение технологии *прописывания «вручную» семантических связей* позволяет расширить запрос до полезного информационного поля, но, естественно, что таким образом не удастся охватить сколько-нибудь значительное количество связей. В итоге возникает необходимость сформулировать *задачу автоматического учета семантических связей*, что возможно *при наличии тезауруса, соответствующего тематике*. Особенную трудность уточнения и расширения информационного запроса представляет *процесс поиска научной информации*, поскольку основу для поиска составляет использование специальной терминологии и связей, задаваемых логикой ПО. Сложность составляет также *иерархическая система представления научных данных*, когда появляется *проблема установления горизонтальных связей между понятиями* [2]. На примере ПО задач математической физики и смежных областей предла-

гается показать, как расширение запроса на основе тезауруса LibMeta может улучшать результаты поиска.

1. ОСОБЕННОСТИ МЕТОДА РАСШИРЕНИЯ ИНФОРМАЦИОННОГО ЗАПРОСА

Расширение информационного запроса (Query expansion[1]) предполагает *переформулирование исходного запроса* с целью улучшения результата поиска. Этот процесс непосредственно связан с *пониманием* предмета поиска как со стороны пользователя (уровень компетентности в некоторой ПО), так и со стороны информационно-поисковой системы (наличие информационных и функциональных средств расширения и уточнения запроса).

Расширение запроса включает такие *методы*, как:

- поиск и использование синонимов для слов из запроса, а также поиск новых синонимов;
- поиск и использование семантических связей с другими словами; это могут быть, например, антонимы (противоположные по смыслу), меронимы (части слов), гипонимы (видовые понятия), гиперонимы (родовые понятия);
- поиск и использование всех различных морфологических форм слов из поискового запроса;
- фиксация ошибок правописания и автоматический поиск исправленной или предложенной словоформы;
- переназначение смысловой нагрузки слов в оригинальном запросе.

Последнее, а именно, «переназначение смысловой нагрузки», может оказать негативное влияние на результат поиска, если новая смысловая нагрузка не связана семантически с исходной ПО поискового запроса, что может привести к увеличению поискового шума.

В связи с расширением поискового запроса обсуждались также «термины расширения» (term expansion) и вопросы «улучшения запроса» (query enhancement). В работе [3] отмечено, что историю исследований расширения информационного запроса можно отследить, начиная с 1965 года, когда в работе [4] было дано формализованное описание релевантности результатов поискового запроса на основе векторной модели обратной связи (известное, как ал-

[1] https://en.wikipedia.org/wiki/Query_expansion

горитм Роккио — Rocchio algorithm). Более ранние исследования в области оценки веса связанных и не связанных терминов при расширении запроса принадлежат Спарку Джонсу (Spärck Jones) [6] и Ван Ризербергу (van Rijsbergen) [7]. Идея обратной связи по релевантности (Relevance Feedback — RF) заключается в привлечении пользователя к процессу поиска, чтобы улучшить итоговый список результатов. В частности, пользователь сообщает системе о релевантности документов в первоначальном списке результатов. Алгоритм Роккио — классический алгоритм для реализации метода RF. Он добавляет модель обратной связи по релевантности в модель векторного пространства [5]. Автоматическая генерация тезаурусов обсуждалась в работах Кью и Фрая (Qui and Frei) [8] и Шютце (Schütze) [9]. Использование локальных и глобальных методов расширения запросов исследовано в работах Крофта (Croft) и соавторов [10].

Эти работы заложили основу для дальнейших исследований в области расширения информационного запроса для текстовых документов в эпоху, когда еще не было достаточно инструментов для обработки символьной информации. В настоящее время получили развитие программные средства, позволяющие учитывать в базах данных формулы [11], стало возможно использовать формульную запись для расширения поискового запроса.

Предлагается подход, основанный на *учете смежных областей,* благодаря ассоциативным связям терминов тезауруса. Ранее была предложена *технология пополнения тезауруса адресата*, тезауруса по обыкновенным дифференциальным уравнениям и уравнениям смешанного типа [12, 13]. Это технология — для осведомленного пользователя. Если пользователь недостаточно знаком с ПО, то любая информация может быть (или не быть) *пополнением тезауруса адресата*. Расширение поискового запроса для такого пользователя может служить полезной подсказкой в процессе поиска. Рассматриваются варианты онтологии *одной ПО* и онтологий *различных ПО*.

1.1. Расширение запроса для одной предметной области

В качестве примеров продемонстрируем процесс поиска математических публикаций. Эти примеры характерны тем, что в тезаурусе математических ПО термины довольно часто сопровождаются формулами, символьным представлением терминов. На примере рис. 1 показано увеличение полей поиска науч-

ных публикаций за счет *связей терминологии смежных областей и переформулирования запроса*.

Известно, что одни и те же явления, встречающиеся в естественных науках, поддаются моделированию в различных областях знаний, при этом могут использоваться идентичные (аналогичные) символьные выражения. Например, «волновое уравнение» используется при моделировании различных технических процессов. Запись волнового уравнения практически везде одна и та же.

Рис. 1. Схема расширения запроса для смежных предметных областей

По цепочке связей тезауруса уравнений математической физики можно легко перейти к поиску из массивов литературы по одной ПО к другой. В примере на рис. 1 показано, как от формулировки запроса по теме «волновое уравнение» осуществляется переход к формулировке по теме «околозвуковая газовая

динамика». Другой пример: «уравнение Трикоми» из раздела «уравнений смешанного типа» также имеет многочисленные приложения — от описания задач «магнитогидродинамических течений» до задач «околозвуковой газовой динамики» из одной более общей ПО «уравнений математической физики» (УрМатФиз). Этих примеров можно найти неограниченное количество, поскольку УрМатФиз, как предметная область, появилась для моделирования физических и технических процессов, т. е. имеет множество приложений и смежных областей. Их информационные образы в поисковых системах могут быть охвачены, благодаря возможностям расширенных запросов.

1.2. Расширение запроса для различных предметных областей

Особое значение имеет процесс расширения запроса при интеграции большого объема информации из различных ПО и распределенных источников. Имея онтологии ПО, можно организовать поиск с расширением запроса в различных направлениях, задаваемых цепочками семантических связей.

В рамках разрабатываемой технологии на основе LibMeta [14] проводится интеграция данных из различных областей знаний, представленных в виде предметных онтологий. В частности, энциклопедические данные, интегрированные в систему, позволяют использовать ассоциативные связи терминов для расширения информационного запроса вплоть до обращения не только к смежным областям знаний. Схематично история и технология расширения информационно-поискового запроса при наличии онтологий различных ПО отражена на рис. 2.

Создание предметных тезаурусов и внедрение этих знаний в виде онтологий ПО позволяет предоставлять пользователям информационных систем *расширять поисковые запросы*. Таким образом, используя связи смежных областей, можно обеспечить переход к *поиску в различных типах* цифровых ресурсов из *различных ПО*. Этот подход реализуется для поиска среди объектов информационной системы, но и за ее пределами, в доступных для интеграции базах данных, где встречаются термины из расширенного запроса.

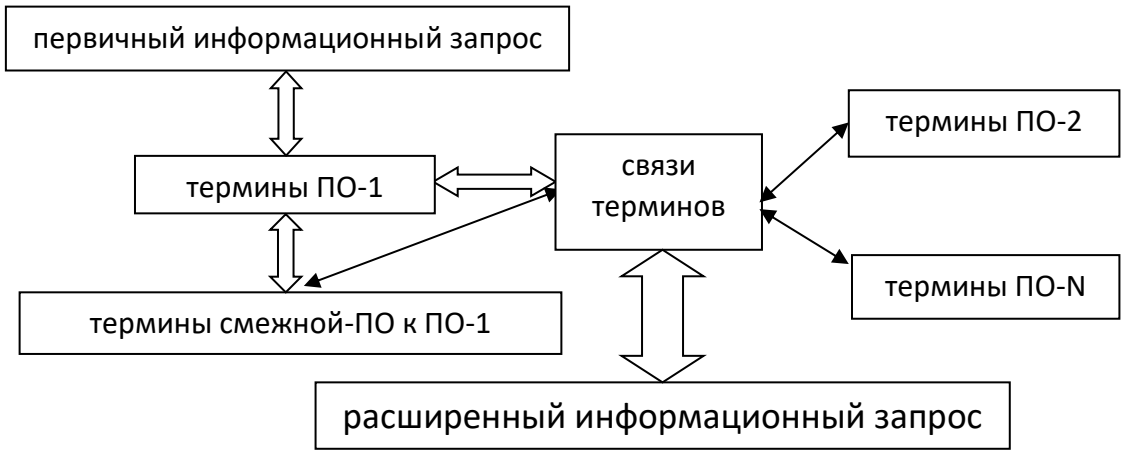

Рис. 2. Схема формирования расширенного информационного запроса для поиска в различных предметных областях

2. ПРЕИМУЩЕСТВА ИНТЕГРАЦИИ ДАННЫХ В КОНТЕКСТЕ РАСШИРЕНИЯ ЗАПРОСА

На первый взгляд, преимущества очевидны: чем больше охват запроса, тем больше информации в качестве результата получит пользователь интегрированной информационной системы. Тем не менее, известно также, что расширение запроса приводит к увеличению информационного шума, что никак нельзя отнести к преимуществам при поиске. Сочетание этих двух особенностей должно принимать некое «оптимальное» значений для того, чтобы услуга расширения поискового запроса составляла полезное свойство информационной системы.

Оптимальные свойства интегрированной системы, обеспечивающие *эффективность расширения информационного запроса*, реализуются, благодаря следующим особенностям:

- структуре данных;
- функциональным свойствам;
- возможности "настройки" на ПО пользователя.

2.1. Структура данных LibMeta

Онтология описывает ресурсы ПО и их взаимосвязи. Для каждой ПО LibMeta набор ресурсов может отличаться как по формату, так и по набору самих ресурсов.

Для описания библиотеки в LibMeta используются смысловой контент *конкретной* ПО и понятия, общие для *любой* из них, то есть предлагается набор по-

нятий, формирующих описание контента библиотеки, достаточно универсальный для включения в систему конкретной ПО. Такой подход позволяет реализовать средства интеграции данных в рамках библиотеки, адаптируемые под условия любой ПО с учетом ее специфики. Это позволяет решать одну из основных проблем интеграции данных из различных источников, а именно, согласование разнородной цифровой информации.

Понятия онтологии в системе LibMeta можно условно разделить по функциональному предназначению для следующих целей:

- описание контента ПО;
- формирование тезауруса любой ПО,
- описание тематических коллекций,

описание задачи интеграции контента библиотеки с данными из внешних источников.

Между этими группами понятий определены семантически значимые связи.

2.2. Функциональные особенности системы LibMeta

Семантическая библиотека LibMeta представляет собой информационную систему, в рамках которой задается описание ПО с терминологической поддержкой и возможностью интеграции данных из разных источников данных, удовлетворяющим требованиям, предъявляемым к источникам данных в LOD (Linked open data[2] [15]). Соответствие требованиям может быть *неполным,* и это означает, что, возможно, требование, касающееся *связанности данных* с другими источниками, может не выполняться, но с помощью LibMeta появляется возможность достаточно просто выполнить его. Для этого от пользователя – эксперта в ПО – не требуется специальных технических знаний об используемом для этого стеке технологий LOD.

Перечислим основную функциональность системы:

- создание/просмотр/редактирование информационных ресурсов и их структуры;
- создание/просмотр/редактирование информационных объектов и их структуры;

[2] https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/

- подключение источников данных;
- загрузка данных из подключенных источников данных, в дальнейшем становящихся частью контента библиотеки;
- создание/просмотр/редактирование структуры тезауруса поддерживаемой ПО;
- создание/просмотр/редактирование понятий тезауруса
- атрибутный/семантический/полнотекстовый поиск и навигация по доступным информационным объектам системы;
- атрибутный/семантический/полнотекстовый поиск по источникам данных;

создание/просмотр/редактирование коллекций информационных объектов.

2.3. Настройка на предметную область пользователя в LibMeta

Адаптация данных ПО трактуется как «настройка» источников, в которой можно выделить несколько основных этапов:

- *Подключение источника данных* S_i. Каждый источник данных характеризуется соответствующим уникальным *URL*-адресом и некоторым набором параметров, необходимых для доступа к данным. Проводится предварительный анализ доступной из источника информации, в частности, определяются *типы его ресурсов и их свойства*, участвующие в интеграции. Результатом этого первого этапа становится определение той части схемы источника S_i, по которой будут извлекаться данные.
- Определение *типов ресурсов библиотеки LibMeta*, соответствующих *типам ресурсов источников*. Для каждого ресурса источника, определенного его схемой, извлеченной на этом этапе, ставится в соответствие ресурс библиотеки LibMeta. Результатом этого этапа становится *установление связи* между ресурсом библиотеки и ресурсом источника с помощью соответствующей операции, которая декларирует, что существуют экземпляры этих ресурсов, соответствующие одному и тому же объекту *реального* мира. На базе определенных (выявленных) связей на следующем этапе проходит отображение атрибутов.
- Для каждого ресурса LibMeta определяется отображение атрибутов на соответствующие им свойства ресурса источника данных. В первую очередь строится отображение для идентифицирующих атрибутов, являющихся обяза-

тельными, затем для остальных. *Для каждой такой пары определяются тип связи и набор операций*.

Благодаря такому построению отображения получаем *набор правил*, по которым можно представить каждый *найденный объект в источнике* в рамках понятий библиотеки LibMeta и, соответственно, позволить сохранить его *метаданные в локальном хранилище* по требованию пользователя либо просто *сохранить связь* между найденным объектом в источнике и объектом в библиотеке.

2.4. Интеграция данных различных предметных областей

Формальная модель процесса интеграции данных из различных ПО может быть представлена следующим образом.

Исходя из основных понятий LibMeta, модель контента библиотеки G представляет собой:

- множество ресурсов $R = \{r_j\}$,
- множество атрибутов $A = \{a_i\}$,
- набор атрибутов $N(r) \subset A$, то есть $r_j(a_1,.., a_n)$, $a_n \in N(r)$, определенный для каждого ресурса.

В каждый набор атрибутов входят *идентифицирующие атрибуты*, $I(r) \subset N(r) \subset A$, используемые для однозначной идентификации информационных объектов этого ресурса.

Формально подсистема интеграции I_T представляется тройкой $<G,\{S_i\},\{M_i\}>$, где G – предварительно определенная модель контента, состоящая из множества ресурсов R и их описаний в виде набора атрибутов $N(r)$, S_i – схема i-го источника, подключенного к системе, M_i – отображение i-го источника, $1 \le i \le n$, где n – количество источников данных.

Использование источника данных может происходить по двум сценариям:

1. *в режиме проставления связей* с объектами, имеющимися в библиотеке,
2. *в режиме атрибутного поиска* по источнику данных в рамках заданного отображения.

При этом сохранение данных об объектах из источников может быть выполнено двумя способами:

1. *связывание* – этот способ идентичен по смыслу проставлению связи «смотри также» и означает, что на одном конце содержится более полная и обширная информация по ресурсу;
2. *идентификация* – этот способ идентичен по смыслу проставлению связи «такой же как» и означает, что на одном конце содержится точно такой же по качеству информации объект, как и с другой.

В связи с гибкостью модели контента библиотеки предполагается возможным сценарий создания дополнительных типов ресурсов для подключаемых источников, информацию из которых можно использовать как значения некоторых атрибутов основных ресурсов.

Схему ресурса библиотеки *G,* как источника данных *S,* так и контента, можно представить в виде графа (рис. 3), который включает *объекты* и *отношения*.

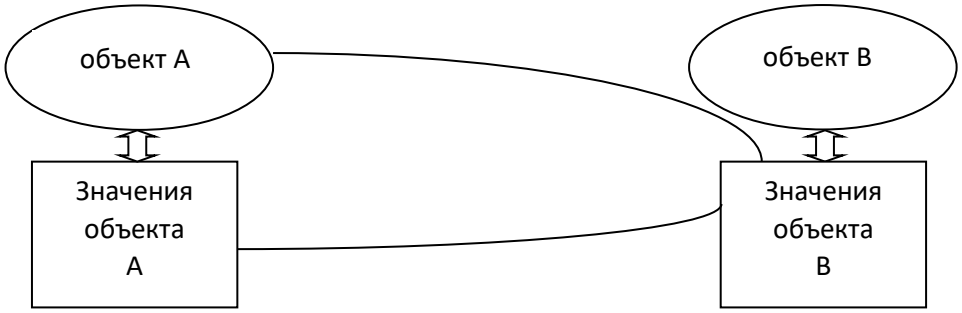

Рис. 3. На схеме линии представляют примеры связей «значение – значение» и «объект – значение» при отображении ресурсов из источников

Каждый объект может быть связан отношением с другим объектом, значения которого представлены простыми типами данных (*строки, числа, даты*) или отношениями с другими объектами, значения которых соответствуют некоторым ресурсам. При этом для *отображения ресурса* мы можем использовать *его представление* Z_s, то есть выбрать не полный набор его атрибутов и отношений с другими объектами для отображения на схему *G*. При этом *представление* Z_s должно обязательно включать в себя набор атрибутов, значения которых позволяют однозначно идентифицировать объект в системе.

Благодаря описанному подходу, структура тезауруса может гибко настраиваться для произвольных ПО. В этом смысле понятия тезауруса могут иметь *мультидисциплинарный характер и*, в силу указанных связей, содержать *указание на смежную область науки* или *явную ссылку на понятия тезауруса смежной ПО*. Также для каждого понятия могут указываться, например, соответствующий код УДК и/или любого другого рубрикатора науки и использоваться наряду с другими как *средство расширения запроса*. Это позволяет *уточнять семантику связанных ресурсов* и использовать ключевые слова экземпляров ресурсов и термины тезауруса как ключевые слова соответствующих рубрик используемых рубрикаторов. Помимо *основных понятий в тезаурусе можно ввести дополнительные категории*, поддерживающие возможность сохранения дополнительной информации. Для этого в тезаурусе могут вводиться связи с ресурсом библиотеки, а именно, в структуре понятия тезауруса могут быть предусмотрены дополнительные соответствующие атрибуты.

2.5. Пример эффективного расширения информационного запроса в LibMeta

Для поддержки поиска по формулам в системе было введено понятие *Формула,* которое позволяет хранить оригинальную строку формулы из того источника, откуда она получена. Строка может быть в формате Content MathML, Presentation MathML, LaTeX. При необходимости количество типов представления формулы в различных нотациях легко расширяется. Понятие *Формулы* связано отношениями с *информационными объектами,* составляющими контент семантической библиотеки и *понятиями* тезауруса. Таким образом, мы всегда можем построить сеть связей формулы как с понятиями тезауруса, так и с различными информационными объектами системы. Каждая формула может быть дополнена ключевыми словами. Ключевые слова могут проставляться как экспертом системы, так и добавляться автоматически, поступая вместе с формулой из ее источника, а также дополняясь ключевыми словами связанных объектов.

Рассмотрим механизм использования парадигматических связей на примере ПО «задачи математической физики для уравнений смешанного типа. Для понятия тезауруса «уравнения Трикоми» покажем преимущества уточнения запроса при использовании формул. Наиболее распространенная запись для уравнения Трикоми – это $ u_{xx} + u_{yy} = 0 $, остальные составляют, «с точки зре-

ния тезауруса», *формулы-синонимы* (так же, как и все записи, аналогичные приведенной с точностью до обозначений). Этой формулой индексирована, в частности, работа [16].

Попытаемся найти публикации, которые также посвящены этой тематике. Делаем поисковый запрос, содержащий выражение: u_{xx}+u_{yy}, которое является частью описания понятия тезауруса «уравнения Трикоми», и получаем список публикаций, связанных с задачей Трикоми, хотя сам термин в поисковом запросе не использовался. Поиск производится по данным, извлеченным из присоединенных источников. При этом при формировании запроса учитываются структура понятия тезауруса предметной области, ключевые слова, привязанные к этим понятиям, или осуществляется навигация по связям тезауруса для дальнейшего уточнения запроса.

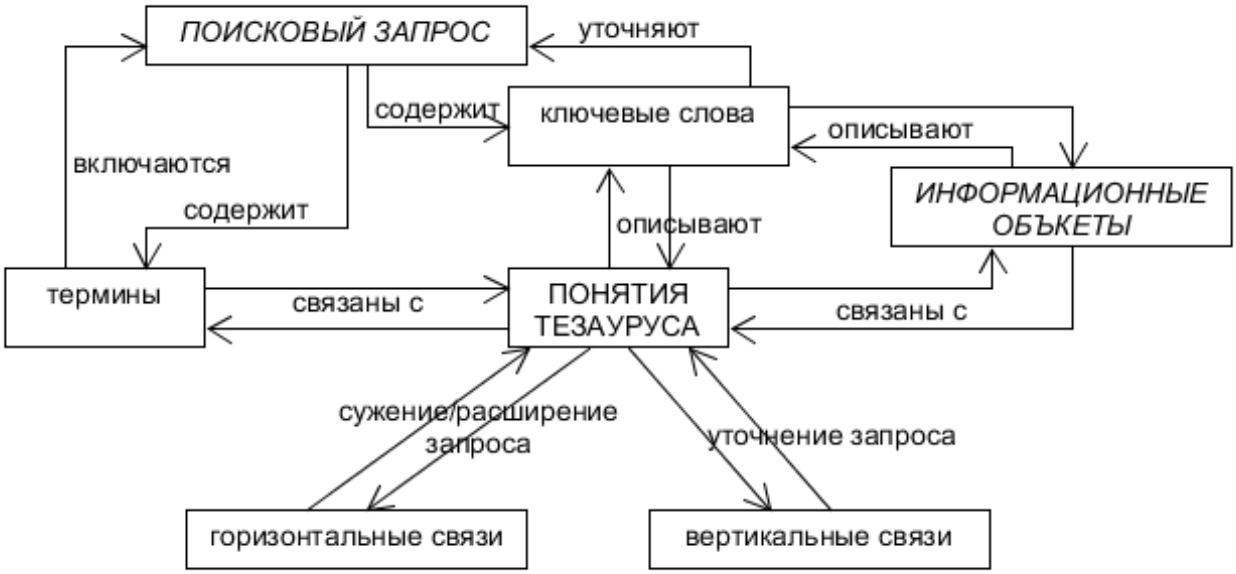

Рис. 4. На схеме представлены примеры формирования уточняющих запросов на основе понятий тезауруса и их связей

На рис. 4 схематически приведены пути формирования уточняющих запросов с использованием понятий тезауруса и его основных связей в системе LibMeta. При этом любой путь от «Поискового запроса» до «Информационного объекта» может оказаться достаточным для получения необходимого ответа на запрос. Ключевые слова могут использоваться не только в общепринятом

понимании, но, как было описано выше, в качестве них могут выступать, например, формулы.

В частности, ещё одно возможное расширение запроса основано на использовании соответствия между разделами MSC и других рубрикаторов, например, УДК. В случае успеха достаточно связать класс, соответствующий разделам MSC, с новым классом для разделов УДК, чтобы категоризовать понятия тезауруса и связанные с ними ресурсы по-новому. Подобного рода работа была проделана на основе использования понятий Математической энциклопедии в качестве тезауруса системы. Это позволило использовать внушительный объём знаний, содержащийся в Математической энциклопедии [17], и связи между ними для представления широкому кругу пользователей-любителей и экспертов в области математики, что особенно ценно при отсутствии в открытом доступе аналогичных ресурсов на русском языке.

В силу специфики данных нам не удалось найти источник данных в LOD для осуществления демонстрации возможностей LibMeta в области интеграции данных с такими источниками. Поэтому мы смоделировали эту ситуацию и выбрали в качестве гипотетического источника из LOD свой локальный источник данных. Наполнением этого источника является массив данных об авторах и публикациях из MathNet[3], который накопился у нас в рамках совместной работы. Этот массив хранится в виде RDF-троек[4].

В результате поиска, в том числе получаем список авторов, которые работают в этой области (например, Ю.М. Крикунов, Г.Л. Алфимов, Richard H., Cushman, Larry M. Bates и др.), а также возможность отследить семантическую сеть связей этой формулы в рамках тезауруса, а также в публикациях, представленных в библиотеке, и их авторов.

При необходимости можно расширять раздел ссылок соответствующей статьи-формулы указателя и расширить описание тезауруса. Ссылки на найденные публикации и авторов можно включить в статьи указателя, где встречаются формулы-синонимы, поиск по которым в данном случае не проводился. Так, через связи реализуется процесс пополнения тезауруса для ПО. В результате в

[3] http://www.mathnet.ru

[4] https://www.w3.org/RDF/

библиотеке LibMeta появятся новые данные о публикациях, и пользователь библиотеки при запросе получит новый список публикаций по теме «уравнения Трикоми». Делая запрос по этой теме, пользователь также получит полную информацию о семантических связях формулы, которая будет включать ссылки на формулы-синонимы, что особенно важно для специалиста.

ЗАКЛЮЧЕНИЕ

Развитие онтологического представления научных предметных областей способствует повышению эффективности поисковых запросов и научных исследований в целом. Учет ассоциативных связей терминов тезауруса позволяет делать выборку не только по смежным областям, но по цифровым массивам различных областей знаний, не увеличивая при этом поисковый шум. Эти выводы вполне ожидаемы, а проблемы, обсуждаемые в работе, актуальны с точки зрения объединения онтологий отдельных областей знаний. Сама эта проблема слияния онтологий представляет собой нетривиальную задачу, как с технологической, так и методологической точек зрения. Этот процесс может привести к качественному возрастанию времени обработки запроса и методологическим противоречиям, характерным для различных научных школ и направлений науки. В приведенных примерах в основном используются данные математических предметных областей, как характерные для расширения запроса за счет использования формул в смежных областях, что, естественно, не ограничивает расширение запроса на другие предметные области, интегрированные в LibMeta. В проекте реализованы связи с любыми источниками, удовлетворяющими требованиям LOD, и идут информационное наполнение и тестирование связей с лингвистической базой данных и математической энциклопедией. Исследования в данном направлении составляют предмет дальнейшей работы.

Благодарности

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проекты № 17-07-00217а, 18-00-00297комфи, 17-07-00214.

СПИСОК ЛИТЕРАТУРЫ

1. *Voorhees E.M.* Query expansion using lexical-semantic relations. In SIGIR 94. ACM 1994. P. 61–69.

2. *Golden P., Shaw R., Buckland M.* Decentralized coordination of controlled vocabularies // Proceedings of the American Society for Information Science and Technology. Annual Meeting, October 31 – November 4, 2014, Seattle, WA, USA. 2014 DOI: 10.1002/meet.2014.14505101146 77th ASIS&T

3. *Vechtomova O.* Query Expansion for Information Retrieval. In: LIU L., ÖZSU M.T. (eds.) Encyclopedia of Database Systems. Springer, Boston, MA. 2009 DOI: 10.1007/978-0-387-39940-9_947

4. *Salton G.* The SMART retrieval system (Chapter 14). Prentice-Hall, Englewood Cliffs NJ. (Reprinted from Rocchio J.J. (1965). Relevance feedback in information retrieval. In Scientific Report ISR-9, Harvard University), 1971.

5. *Маннинг К.Д., Рагбхаван П., Шютце Г.* Введение в информационный поиск. Издательский дом Вильямс. 528 с. ISBN 978-5-8459-1623-5.

6. *Spärck Jones K.* Automatic keyword classification for information retrieval. Butterworths, London, 1971.

7. *van Rijsbergen C.J.* A theoretical basis for the use of co-occurrence data in information retrieval // J. Doc. 1977. V. 33. No 2. P. 106–119.

8. *Qui Y., Frei H.* Concept based query expansion. SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval Pittsburgh, Pennsylvania, USA June 27 – July 01, 1993. ACM New York, NY, USA. P. 160–169. ISBN 0-89791-605-0. DOI:10.1145/160688.160713.

9. *Schütze H.* Automatic Word Sense Discrimination // Computational Linguistics, March 1998 – Special Issue on Word Sense Disambiguation. 1998. V. 24. No 1. P. 97–123. https://www.aclweb.org/anthology/J98-1004.pdf

10. *Larkey L.S., Croft W.B.* Combining classifiers in text categorization // SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval Zurich, Switzerland. August 18–22. 1996. 1996. P. 289–297. ISBN:0-89791-792-8 DOI: 10.1145/243199.243276.

11. Zentralblatt MATH https://zbmath.org

12. *Муромский А.А., Тучкова Н.П.* Об онтологии адресата в математической предметной области // Электронные библиотеки. 2018. Т. 21. № 6. С. 506–533.

13. *Моисеев Е.И., Муромский А.А., Тучкова Н.П.* О тезаурусе предметной области смешанные уравнения математической физики // CEUR Workshop Proceedings. 2018. V. 2260. P. 395–405. DOI: 10.20948/abrau-2018-43

14. *Атаева О.М., Серебряков В.А., Тучкова Н.П.* Подходы к организации математических знаний при формировании предметных тезаурусов различных разделов математики // CEUR Workshop Proceedings. 2018. V. 2260. P. 42–54. ISSN:1613-0073. DOI: 10.20948/abrau-2018-66.

15. *Bizer C., Heath T., Berners-Lee T.* Linked Data – The Story So Far // International Journal on Semantic Web and Information Systems. 2009. V. 5. No 3. URL: https://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf. DOI:10.4018/jswis.2009081901.

16. *Моисеев Е.И., Лихоманенко Т.Н.* Собственные функции задачи Трикоми с наклонной линией изменения типа // Дифференциальные уравнения. 2016. Т. 52, № 10, С. 1375–1382.

17. *Виноградов И.М. (ред.).* Математическая энциклопедия: В 5-ти т. Сов. энцикл., 1979.

CREATION OF QUERY EXPANSION BASED ON THE SUBJECT DOMAIN THESAURUS IN THE ONTOLOGY OF KNOWLEDGE OF THE SEMANTIC LIBRARY

O.M. Ataeva[1], V.A. Serebriakov[2], N.P. Tuchkova[3]

[1,2,3]*Dorodnicyn Computing Centre FRC CSC RAS, Moscow*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Abstract

Possibilities of query expansion with subject area thesaurus are discussed**.** The role of the context defined by thesaurus term links is both to refine the query and to increase the size of the sample on the query. Of particular importance is the process of expanding the query for scientific subject areas where the search based on special terminology. In this case, thesauruses of subject areas must be used to minimize the occurrence of information noise. The proposed approach takes into account the application of similar terminology in various subject areas. Examples of the use of thesaurus of separate sections of equations of mathematical physics and related fields demonstrate the effectiveness of the chosen approach of research. By linking to concepts of information resources of other areas of knowledge, the extension of the information query captures search fields of remote subject areas and various types of data, texts, symbolic, audio and video archives. Research shows that expanding the query based on context semantics improves the search quality of scientific publications in digital information and increases the effectiveness of scientific interdisciplinary research.

Keywords*: comparison of scientific texts, semantic search, thesaurus for the ontology of knowledge, information query using the thesaurus, LibMeta*

REFERENCES

1. *Voorhees E.M.* Query expansion using lexical-semantic relations. In SIGIR 94. ACM 1994. P. 61–69.

2. *Golden P., Shaw R., Buckland M.* Decentralized coordination of controlled vocabularies // Proceedings of the American Society for Information Science and

Technology. Annual Meeting, October 31 – November 4, 2014, Seattle, WA, USA. 2014 DOI: 10.1002/meet.2014.14505101146 77th ASIS&T

3. *Vechtomova O.* Query Expansion for Information Retrieval. In: LIU L., ÖZSU M.T. (eds.) Encyclopedia of Database Systems. Springer, Boston, MA. 2009 DOI: 10.1007/978-0-387-39940-9_947

4. *Salton G.* The SMART retrieval system (Chapter 14). Prentice-Hall, Englewood Cliffs NJ. (Reprinted from Rocchio J.J. (1965). Relevance feedback in information retrieval. In Scientific Report ISR-9, Harvard University), 1971.

5. *Manning C.D., Raghavan P., Schütze H.* Introduction to Information Retrieval, Cambridge University Press. 2008. 544 p. ISBN: 0521865719. Online edition (c) 2009 Cambridge UP. URL: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf.

6. *Spärck Jones K.* Automatic keyword classification for information retrieval. Butterworths, London, 1971.

7. *van Rijsbergen C.J.* A theoretical basis for the use of co-occurrence data in information retrieval // J. Doc. 1977. V. 33. No 2. P. 106–119.

8. *Qui Y., Frei H.* Concept based query expansion. SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval Pittsburgh, Pennsylvania, USA June 27 – July 01, 1993. ACM New York, NY, USA. P. 160–169. ISBN 0-89791-605-0. DOI:10.1145/160688.160713.

9. *Schütze H.* Automatic Word Sense Discrimination // Computational Linguistics, March 1998 – Special Issue on Word Sense Disambiguation. 1998. V. 24. No 1. P. 97–123. https://www.aclweb.org/anthology/J98-1004.pdf

10. *Larkey L.S., Croft W.B.* Combining classifiers in text categorization // SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval Zurich, Switzerland. August 18–22. 1996. 1996. P. 289–297. ISBN:0-89791-792-8 DOI: 10.1145/243199.243276.

11. Zentralblatt MATH https://zbmath.org

12. *Muromskij A.A., Tuchkova N.P.* Ob ontologii adresata v matematicheskoj predmetnoj oblasti // Elektronnye biblioteki. 2018. T. 21. No 6. S. 506–533.

13. *Moiseev E.I., Muromskij A.A., Tuchkova N.P.* O tezauruse predmetnoj oblasti smeshannye uravneniya matematicheskoj fiziki // CEUR Workshop Proceedings. 2018. Vol. 2260. P. 395–405. DOI: 10.20948/abrau-2018-43.

14. *Ataeva O.M., Serebryakov V.A., Tuchkova N.P.* Podhody k organizacii matematicheskih znanij pri formirovanii predmetnyh tezaurusov razlichnyh razdelov matematiki // CEUR Workshop Proceedings. 2018. Vol. 2260. P. 42–54. ISSN:1613-0073. DOI: 10.20948/abrau-2018-66.

15. *Bizer C., Heath T., Berners-Lee T.* Linked Data – The Story So Far // International Journal on Semantic Web and Information Systems. 2009. V. 5. No 3. URL: https://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf. DOI:10.4018/jswis.2009081901.

16. *Moiseev E.I., Lihomanenko T.N.* Sobstvennye funkcii zadachi Trikomi s naklonnoj liniej izmeneniya tipa // Differencial'nye uravneniya. 2016. T. 52, № 10, S. 1375–1382.

17. *Vinogradov I.M.* Matematicheskaya entsiklopediya [Mathematical Encyclopedia] //Moscow, Sovetskaya entsiklopediya Publ. 1979.

СВЕДЕНИЯ ОБ АВТОРАХ

АТАЕВА Ольга Муратовна – научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, специалист в области системного программирования и баз данных.

Olga Muratovna ATAEVA – researcher of the of Dorodnicyn computing center FRC SCS RAS, expert in the field of system programming and databases.

email: oli@ultimeta.ru

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 РАН, «Научный портал РАН».

Vladimir Alekseevich SEREBRIAKOV – 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 RAS, Scientific portal RAS.

email: serebr@ultimeta.ru

ТУЧКОВА Наталия Павловна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат физ.-мат. наук, окончила ВМиК МГУ им. М.В. Ломоносова. Специалист в области алгоритмических языков и информационных технологий.

Natalia Pavlovna TUCHKOVA – 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.

email: natalia_tuchkova@mail.ru

Материал поступил в редакцию 15 ноября 2019 года