

Модель механизма документального информационного поиска на графах знаний

Н.В. Максимов, О.Л. Голицына, К.В. Монанков

Национальный исследовательский ядерный университет «МИФИ»

nv-maks@yandex.ru, olgolitsina@yandex.ru, kmonankov@yandex.ru

Аннотация

В статье рассматривается технология представления полного текста документа графом онтологии, как системой трех систем: функциональной, понятийной и терминологической. Функциональные отношения классифицируются в соответствии с таксономией функциональных отношений и используются для построения аспектных проекций онтологий. В качестве модели данных онтологии используется помеченный ориентированный граф, что позволяет формализовать операции над онтологиями. Таким образом, поисковый образ документа дополняется совокупностью извлекаемых из текста взаимосвязанных сущностей, что обеспечивает глубокое семантическое индексирование.

Представление семантического пространства документов графом онтологии позволяет свести задачу семантического поиска к теоретико-множественным моделям и использовать для отбора формально релевантных документов традиционные механизмы поиска. А совместное применение различных традиционных механизмов поиска, механизмов количественного анализа и механизмов поиска на графах обеспечивает взаимодополнительность подходов, позволяет повысить восприятие и глубину понимания проблемы и улучшить результаты поиска.

Ключевые слова: семантический поиск, обработка текста, графовые представления онтологий, механизм поиска, аспектная проекция

Библиографическая ссылка: Максимов Н.В., Голицына О.Л., Монанков К.В. Модель механизма документального информационного поиска на графах знаний // Компьютерная лингвистика и вычислительные онтологии. Выпуск 5 (Труды XXIV Международной объединенной научной конференции «Интернет и современное общество», IMS-2021, Санкт-Петербург, 24 – 26 июня 2021 г. Сборник научных статей). — СПб.: Университет ИТМО, 2021. С. 39-50. DOI: 10.17586/2541-9781-2021-5-39-50

Введение

Традиционно задача информационного поиска формулируется, как задача формирования выборки документов некоторой коллекции в соответствии с информационной потребностью, а основными показателями эффективности процесса поиска являются точность, полнота, оперативность. Однако в действительности человеку, как потребителю информации, нужна не выборка релевантных документов и даже не их полные тексты, а решение задачи его основной деятельности, *описание* которого полностью или частично может содержаться в найденных документах.

Зачастую (по крайней мере, для новых задач и подходов) ни один документ не содержит в явной форме полного и адекватного описания решения. Но, как известно, любое новое решение основывается и использует уже известные знания и факты.

Такая ситуация особенно свойственна научным или инженерным задачам, задачам проектирования и т.д. То есть человек может (и должен) идти путем формирования образа решения из фрагментов доступных ему решений, например, схожих задач. Такой прообраз решения может быть «собран» из фрагментов соответствующих текстов, находимых в информационных ресурсах. Это означает, что поисковая система должна отбирать документы (и более того – фрагменты документов) не только по правилу присутствия ключевых понятий, но и с учетом контекста, отражающего ситуативные связи.

Рассматривая использование автоматизированных документальных информационно-поисковых систем (ИПС) как деятельность, замещающую основную деятельность человека, задача синтеза нового знания может быть представлена, как задача формирования образа решения путем построения некоторого единого текста из соответствующих существу предполагаемого решения фрагментов текстов релевантных документов, найденных традиционными методами поиска. Построенный текст будет представлять некоторый (очевидно, неполный и несогласованный) прообраз – квазиреферат решения задачи основной деятельности.

Такая ИПС строится на методах и технологиях глубокого семантического индексирования, позволяющих извлекать смыслы из текста и представлять их в хорошо структурированной схематичной форме, удобной для манипулирования (в том числе укрупнения и детализации) – мульти-метаграфовых представлений онтологии, автоматически формируемых путем извлечения из текста и типизации сущностей и связей. К получаемым графам, помимо теоретико-множественных операций (объединение, пересечение), могут применяться операции масштабирования и аспектной проекции, которые позволяют, вычлняя соответствующие некоторому аспекту рассмотрения подграфы, компактно визуализировать содержание, а также выявлять скрытые связи в предметной области (ПрО). Для семантического индексирования и графовых операций используются построенные на решетке базовых категорий таксономии сущностей и отношений, позволяющие соотносить сходные по смыслу понятия и связи, которые в текстах вследствие свойств естественного языка могут быть представлены различающимися вербальными конструкциями.

Знания, как объект деятельности человека, достаточно адекватно представляются онтологическими средствами, поскольку такие средства, согласно [1; 2], отражают не только имманентные и ситуативные связи ПрО, но также и отношения между понятиями и категориями инструмента познания и, в том числе, языка. В этом смысле онтологии могут быть «полигоном с контурными картами», на котором пользователь реализует траекторию как информационного, так и предметного поиска: понимание целесообразности использования ключевых понятий обеспечивается за счет визуализируемых контекстов этих понятий и построения путей между ними – возможных смыслов.

При этом интерактивная визуализация семантического графа онтологии документа позволяет использовать граф как инструмент навигации по материалу документа, т.к. дает возможность оперировать контекстно-определенными подграфами и переходить от вершин графа к фрагментам текста. В пределе граф онтологии, сформированный по текстам на естественном языке, может служить в качестве инструмента построения ряда образов (альтернатив и дополнений), которые в совокупности позволяют решить прагматическую задачу пользователя. А исследование (анализ и синтез) графа онтологии как семиотического объекта (точнее – системы), содержащего множество взаимосвязанных фактов, отражающих смысл исходного текста в рамках языка и концептуальных схем ПрО, позволяет проверять совокупность фактов на непротиворечивость, а также находить неявные (непосредственно не отраженные в исходном тексте) факты и связи.

Но следует учитывать, что *поисковое* взаимодействие человека с машиной имеет свои особенности. Пользователь осуществляет узнавание и распознавание объектов решаемой практической задачи в некотором контексте.

Узнавание заключается в отождествлении находимых объектов с наличным знанием «в целом», а распознавание - в выявлении отдельных «полезных» свойств. В том случае, когда решается некоторая практическая задача, отображаемая семантическая сеть должна быть направленной (от исходных положений к «ответу») и представлять собой в идеале алгоритм этого решения. В случае задач информационно-аналитического характера (предварительные исследования, поиск гипотез или возможных путей решения) обычно проводится анализ состава элементов, структуры (блочность, взаимосвязь), содержания (характер сущностей и связей), формы (упорядочение и вид представления).

Кроме того, физически (пространственно) процесс последовательного восприятия/понимания (выбор элемента, его идентификация и связывание) реализуется с использованием некоторой (привычной или специальной) схемы. Например, «слева-направо» в случае продолжения, расширения области; «сверху вниз» или «вглубь» – в случае уточнения или детализации.

В статье рассматривается технология формирования семантического образа полного текста, как системы трех взаимосвязанных систем – функциональной, понятийной и терминологической, на которых задана операция сопоставления элементов различных систем.

В качестве модели данных функциональной системы используется помеченный ориентированный граф, обладающий свойством мультиграфа, и на котором могут быть построены метаграф и гиперграф, что позволяет формализовать и реализовать операции над онтологиями.

Наибольший интерес представляет операция «аспектная проекция», которая позволяет «осмысленно» сократить размер графа, и подготовить фрагменты текста истинно релевантные потребности пользователя.

Интерактивная визуализация графа онтологии и механизмы поиска на графах, такие как «поиск пути» и «анализ семантической окрестности», обеспечивают навигацию по текстам документов и понятиям. А сведение задачи семантического поиска к теоретико-множественным моделям позволяет использовать для отбора формально релевантных документов традиционные механизмы поиска.

1. Даталогическая модель семантического образа

Согласно [3], факт в философии науки – это особого рода предложение, фиксирующее эмпирическое знание, утверждение или условие, которое может быть верифицировано, а смысл факта находится за пределами самого факта и определяет его место в некоторой целостности.

Исходя из этого определения, выделяют [4] следующие типы информационных компонентов:

- элементарный факт – образ, фиксирующий некоторое состояние отдельного взаимодействия пары сущностей, где в роли сущности выступает понятие, объект, субъект и т.п., а взаимодействие представлено связью (отношением);
- ситуативный факт – элементарный факт, в котором обе сущности (или одна из них) доопределены обстоятельствами участия сущности во взаимодействии – конкретной ситуацией; таким образом, формируется новая именованная сущность, включающая совокупность элементарных фактов;
- завершённый факт (высказывание, утверждение, описание) – сеть элементарных и/или ситуативных фактов, образующая целостность, соотносимую с информационным запросом, и таким образом формирующая смысл.

Факт может быть зафиксирован и существовать в разных формах, в том числе, в виде текста, как набора знаков некоторого языка, в частности, естественного.

Так же, как и осмысленный текст, граф онтологии, построенный по этому тексту, может быть рассмотрен, как набор фактов, в совокупности выражающий некоторый

смысл. В таком случае элементарному факту в графе онтологии соответствует триплет «сущность – отношение – сущность», а ситуативному факту – триплет, в котором одна или обе сущности представлены совокупностью элементарных триплетов, составляющих семантическую окрестность атомарной сущности. Такая метасущность отражает некоторую ситуацию, может быть поименована и снабжена характеристическими атрибутами. Отношение же в рамках ситуативного факта будет иметь характер метаотношения, т.к. соединяет не атомарные сущности, а целые ситуации. Завершенный факт представлен в графе онтологии некоторой целостной конструкцией триплетов, с одной стороны, реконструирующей намерения создателя исходного текста, а с другой – соответствующей контексту решаемой пользователем задачи основной деятельности.

Онтологический подход¹ позволяет представить семантику описанного в документе отдельного решения системой понятий и отношений, т.е. при поиске можно будет использовать завершенные смысловые конструкции. Граф онтологии при этом будет представлять технологическое пространство «точек входа» в информационный массив, обеспечивая возможность непосредственного перехода от вершин графа к фрагментам текста документа.

Представление онтологии в виде графов позволяет формализовать операции над онтологиями на основе теоретико-графовых аксиом. Согласно [1; 5], основными операциями являются: бинарные – объединения, пересечения, проекции и унарная – масштабирования онтологий.

В качестве модели данных функциональной системы (представляющей собственно описываемый объект/процесс) используется помеченный (для вершин и дуг которого определены свойства A_f) ориентированный граф $G(V, E) = \langle V, E \rangle$, где V – множество вершин, а E – множество дуг. Множество вершин и множество дуг в совокупности соответствуют множеству элементарных фактов, представляющих соответствующие сущности, связанные типизированными отношениями. Так как одна и та же пара атомарных сущностей может участвовать в нескольких элементарных фактах, в графе $G(V, E)$ для двух вершин может существовать более одной дуги, т.е. $G(V, E)$ обладает свойством мультиграфа.

На множествах V и E определяются (и могут быть динамически построены):

1) Метаграф, формально задаваемый как $MG = \langle V, MV, E, ME \rangle$, где V – множество вершин, MV – множество метавершин, E – множество дуг, ME – множество метадуг. Каждая метавершина соответствует метасущности ситуативного факта и представляет собой граф $mv_i = \langle V_i, E_i \rangle$, где $V_i \subset V$, $E_i \subset E$, а метадуга – метаотношению.

Для графовых форм, отображающих семантику текстов (и познания), наличие метавершин вполне конструктивно и естественно. Метавершина соответствует (по своему имени) сущности (понятию, узлу, композиции и т.п.) и выступает в качестве атомарного

¹ В [1] онтология, как семиотически целостное образование, определяется с позиций Общей теории систем как совокупность трех взаимосвязанных систем $O = \langle S_f, S_c, S_t, \equiv \rangle$, где S_f – функциональная система (объекты и связи действительности), определенная как $S_f = \langle M_f, A_f, R_f, Z_f \rangle$, где M_f – множество объектов (сущностей), A_f – множество характеристических свойств, R_f – множество функциональных отношений, представленных типизированными ситуативными связями, характерными для ПрО, Z_f – закон композиции, т.е. правила и схемы упорядочения объектов (например, мерономия ПрО); S_c – понятийная система, определенная как $S_c = \langle M_c, A_c, R_c, Z_c \rangle$, где M_c – множество понятий ПрО, A_c – множество признаков систематизации понятий (таксономия ПрО), R_c – множество отношений (прежде всего, парадигматических), Z_c – закон композиции (схема представления); S_t – терминологическая система, определенная как $S_t = \langle M_t, A_t, R_t, Z_t \rangle$, где M_t – множество терминов, A_t – множество свойств, R_t – множество отношений эквивалентности и включения, а также лингвистических отношений, Z_t – закон композиции (грамматика); \equiv – операция сопоставления элементов различных систем на уровне знаков, обеспечивающая их тождество в функциональной, понятийной и терминологической системах.

семантического эквивалента смысла, определяемого неатомарной конструкцией (выражением).

Метаграф также обладает свойством мультиграфа, т.к. ситуативные факты могут, например, различаться только метаотношениями.

2) Гиперграф, когда на множестве вершин $VUMV$ формируется множество гиперребер W , при этом в основе правил задания гиперребра лежат множества A_f и Z_f :

$$W = \{w_1, w_2, \dots, w_n\}, w_i = V_i \cup MV_i, \text{ где } V_i \subset V, MV_i \subset MV; V_i \neq \emptyset \vee MV_i \neq \emptyset$$

Смысловая направленность является важным фактором адекватности восприятия. Исходя из того, что схема восприятия информации предопределяется характером задачи и ее когнитивным состоянием, при визуализации графа можно использовать разные варианты укладки вершин (закон композиции), которые будут определять характер «навигации» по графу. В частности, в [6] используются следующие алгоритмы укладки вершин на плоскости:

- укладка вершин в соответствии с некоторой схемой действий, например, в соответствии с функциональной моделью IDEF0;
- построение пути между двумя указанными сущностями, обеспечивающего представление цепочки фактов;
- укладка вершин в порядке употребления сущностей в тексте;
- укладка вершин в соответствии с упорядоченностью по значимости, где в качестве значимости используется длина пути или суммарный вес вершин;
- укладка вершин методом решения гравитационной задачи n -тел.

В целом использование различных вариантов укладки вершин во время решения пользовательской задачи позволяет взглянуть на ее возможные решения с разных сторон, что за счет взаимодополнительности подходов позволяет повысить уровень восприятия и глубину понимания проблемы.

Наличие в функциональной системе онтологии, помимо множеств сущностей и функциональных отношений, множества характеристических свойств и закона композиции позволяет группировать сущности не только в динамике, например, по принципу соответствия синтезируемой цепочки фактов, но и в статике – например, по принципу обладания общим свойством, по лексикографическому включению и т.п.

Сформированный таким способом граф, по существу, является рефлексивным образом существующего решения/состояния проблемной ситуации, но поскольку рефлексия – это отображение содержания текста (представленного концептуальным графом) на проблемную ситуацию субъекта, то такой граф представляет и некоторое видение будущего.

Таким образом, вследствие того, что в графической форме нивелируются различия между смыслом содержания и спецификой представляющих его лингвистических конструкций, то интерактивно выполняемые операции над графами (теоретико-множественные и «поисковые») позволяют «соединить» логические (абстрактные) операции и конкретные схемы. Это облегчает восприятие, понимание и оценку содержания, в том числе и потому, что в фокусе внимания будет компактная целостная картина.

2. Построение аспектного представления графа онтологии

Аспектное представление, как одна из форм заверщенного факта, представляющего некоторый смысловой срез ПрО, в рамках функциональной системы онтологии реализуется в виде подграфа. В основе построения аспектного представления лежит операция проекции, которая в [1] сводится к операции пересечения исходной – $O = \langle S_f, S_c, S_t, \Xi \rangle$ и аспектной – $O_i = \langle S_f^i, S_c, S_t, \Xi \rangle$ онтологий: $O_{proj} = O \cap O_i$. Таким образом, для каждого аспектного представления должна быть задана (на уровне функциональной системы) своя аспектная онтология.

В общем случае используется таксономия аспектов, которая (являясь объектом, открытым для расширения и модификации) задает множество возможных аспектов, связанных с классами отношений, характерными для этой точки зрения. Множество аспектов определяется в соответствии с моделью деятельности и задается на таксономии функциональных отношений, классы отношений которой связаны с лингвистическими конструкциями в тексте [7], дополненной множеством структурно-лингвистических отношений. Последние позволяют учесть связи, обусловленные языком (синонимия, парадигматика), а также «конструкционные» связи, характерные для задания свойств (наименование, размерность, величина параметра).

Таким образом, задание аспекта в рамках таксономии сводится к формированию функциональной системы $S_f^i = \langle \emptyset, A_f^i, R_f^i, Z_f^i \rangle$ с непустым множеством R_f^i и, возможно, непустыми множествами A_f^i и/или Z_f^i . Более детально аспект может быть задан характерным для него множеством опорных понятий (имен сущностей), т.е. $M_f^i \neq \emptyset$.

В статье будет приведен пример использования аспекта «эффективность» для построения аспектного представления графа онтологии. Аспект «эффективность» позволяет оценить возможный или реальный эффект от выполнения действий над объектом деятельности и задается следующим образом:

- M_f^e включает следующие опорные понятия: «результат», «цель», «следствие», «последствие», «достижение», «итог», «успех», «эффект», «исход», «конец»;
- A_f^e включает следующие роли объекта: «результат», «источник»;
- R_f^e включает отношения следующих классов: «быть источником», «быть результатом», «быть условием», «сравнение», «тождество», «создание», «уничтожение», «исчезновение», «воздействие», «вычисление», «разложение», «преобразование», «трансформация», «присоединение», «включение», а также классы, для которых класс «конкретное» является родительским.

3. Модель механизма поиска на графах знаний

Согласно [8], механизм поиска – это «совокупность реализованных в ИПС моделей и алгоритмов процесса формирования выдачи документов в ответ на поисковый запрос». При этом на множестве моделей механизмов поиска может быть построена «замкнутая система моделей, полная с точки зрения множества операционных объектов и их состояний, в которых в качестве запроса рассматриваются множества терминов и множества документов, а также логические выражения, построенные над этими множествами» [8], и включающая следующие модели:

- модель механизма поиска по совпадению терминов;
- модель механизма поиска по логическому выражению;
- модель механизма поиска документов-аналогов;
- модель механизма эвристического поиска;
- модель механизма поиска с использованием «двойной обратной связи по релевантности».

Также в [8] вводится понятие механизм количественного анализа, как механизм постобработки результатов поиска, позволяющий выявлять характеристические свойства документальных потоков. Так, например, на множестве документов, отобранных в результате информационного поиска по различным направлениям, могут быть построены временные ряды потоков публикаций и применены методы исследования взаимосвязи ПрО [9].

С целью привнести в поисковый образ более глубокую семантику, поисковый образ дополняется сущностями, связанными отношениями, что позволяет ввести понятие *механизмов поиска на графах знаний*, к которым относятся:

- механизм поиска пути на графе онтологии;

- механизм анализа семантической окрестности сущности;
- механизм формирования аспектной проекции графа онтологии.

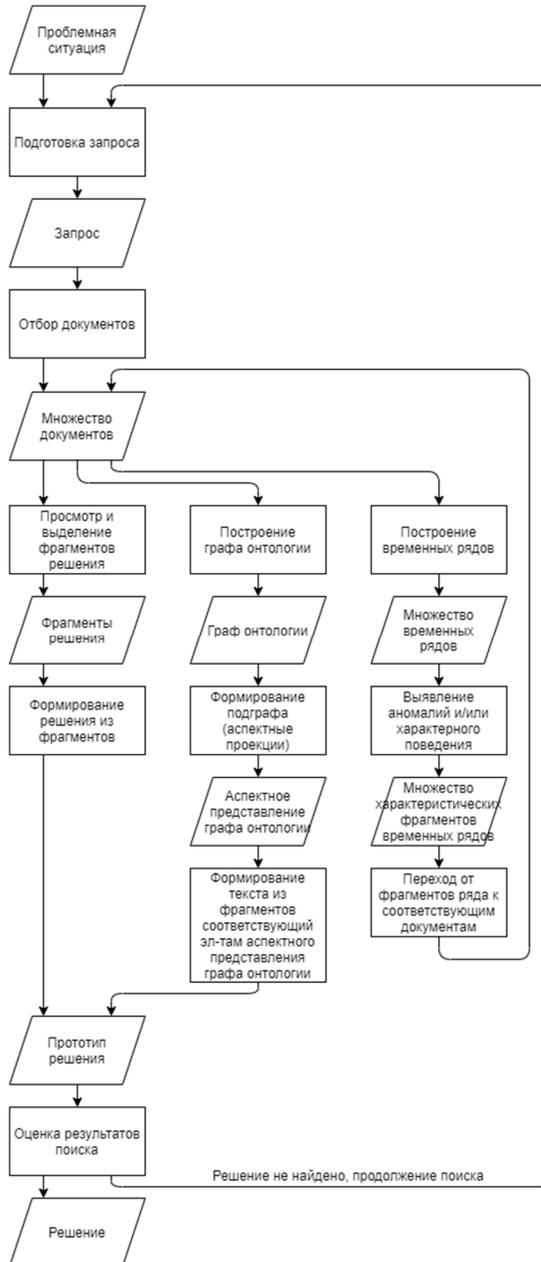


Рис. 1. Обобщенная схема информационного поиска

Для сведения механизмов поиска на графах знаний к теоретико-множественной модели документ информационного массива рассматривается как совокупность элементарных фактов, каждый из которых представлен триплетом «сущность» – «отношение» – «сущность». Сущности и отношения выражаются терминами из универсального словаря *D*. При этом множество отношений классифицируется в соответствии с классификацией

отношений [10], а каждой сущности присваивается роль в тексте исходного документа (из пространства ролей R), которая определяется классом связанного отношения.

Таким образом, исходный информационный массив, состоящий из N_0 документов, можно представить матрицей «термин_роль – документ» размерности $D \times R$ на N_0 . Строки такой матрицы соответствуют терминам информационного массива, сцепленным с кодом роли $r_i \in R$, столбцы – документам, а элемент матрицы характеризует вхождение термина в документ в роли r_i .

Такое представление семантического пространства позволяет свести задачу семантического поиска к теоретико-множественным моделям и использовать для отбора формально релевантных документов традиционные механизмы поиска.

Определение фактической релевантности остается на стороне пользователя. Граф онтологии документа позволяет представить документ в форме, обеспечивающей навигацию по документу. Применение операций поиск пути на графе, анализ окрестности и переход от элементов графа к фрагментам исходного текста позволяют осуществлять управляемый отбор фрагментов текста.

Результат применения модели информационного поиска на графах онтологий позволяет упорядочить множество фрагментов исходного текста в соответствии с цепочками вершин графа.

Таким образом, представление семантического пространства документов графом онтологии позволяет свести задачу семантического поиска к теоретико-множественным моделям и использовать для отбора формально релевантных документов традиционные механизмы поиска.

Обобщенная схема информационного поиска, включающая применение механизмов поиска на графах онтологий и механизмов количественного анализа (в частности, на временных рядах), приведена на рис. 1.

4. Пример поиска на графах знаний

Рассмотрим пример поиска на графах знаний по теме «проект «Прорыв» с применением операции аспектная проекции для построения аспектного представления в аспекте «эффективность».

В результате информационного поиска по теме «Проект «Прорыв» найдено 7 релевантных документов, полные тексты которых объединены в единый текст. С применением сервиса визуального онтологического анализа научно-технических текстов [11] по единому тексту построен граф онтологии, размер которого не позволяет визуализировать его в воспринимаемой форме (граф насчитывает тысячи вершин).

К графу применена операция «аспектная проекция» в аспекте «эффективность», в результате чего сформировано аспектное представление – подграф, представленный на рис. 2, в который включены элементы графа, удовлетворяющие условиям задания аспекта, а именно: дуги, соответствующие отношениям из R^c ; а также дуги, относящиеся к структурно-лингвистическим отношениям «контекст», «равно», «включает сущность», «обстоятельство употребления»; и, соответственно, вершины, инцидентные этим дугам.

В подграфе выполнен поиск вершин по выражению «Прорыв». Далее, при визуальном исследовании окрестности найденной вершины (рис. 2, правый нижний угол), можно обнаружить цепочку между вершинами «прорыв» и «достижение нового качества ядерной энергетики» через промежуточную вершину «двухкомпонентной ядерной энергетики». Наличие такой цепочки позволяет сделать предположение², что возможным эффектом реализации проекта «Прорыв» является

² Необходимо отметить, что разработанные технологии направлены на решение информационно-поисковых задач и не относятся к классу экспертных систем. Система предоставляет пользователю материалы для выбора и анализа, в частности, информацию о возможном существовании связи

достижение нового качества ядерной энергетики за счет двухкомпонентной атомной энергетики.

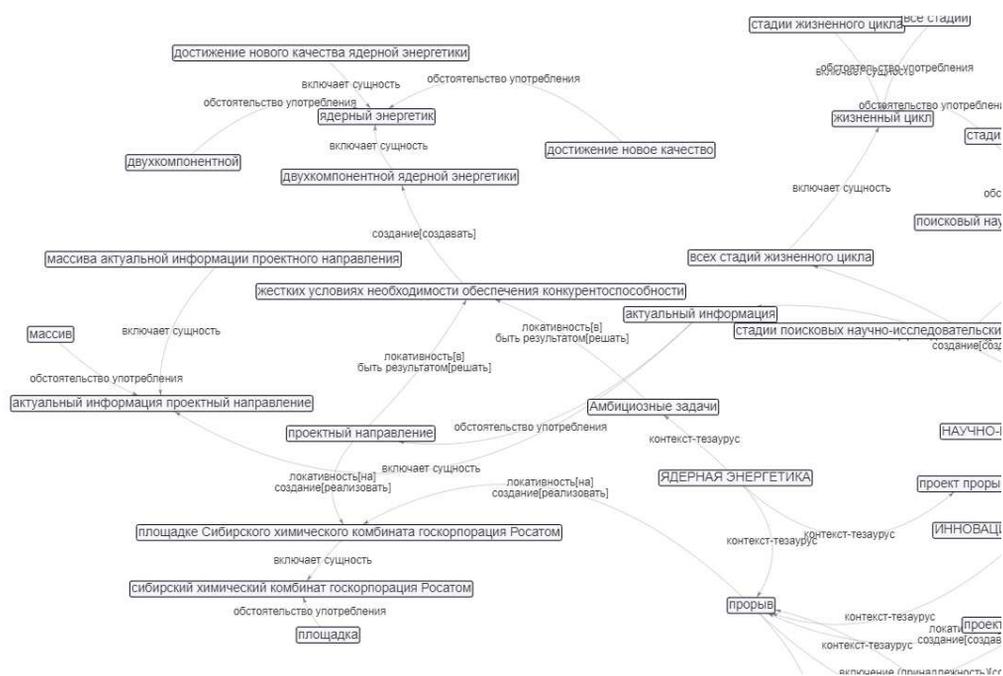


Рис. 2. Аспектное представление графа онтологии в аспекте «эффективность»³

Переход от вершин рассматриваемой цепочки к соответствующим фрагментам исходного текста позволяет подтвердить это предположение, так как исходный текст содержит следующие фрагменты⁴: «Реализуемый Госкорпорацией «Росатом» проект «Прорыв» нацелен на **достижение нового качества ядерной энергетики**, разработку, создание и промышленную реализацию замкнутого ядерного топливного цикла (ЗЯТЦ) на базе реакторов на быстрых нейтронах, развивающих крупномасштабную ядерную энергетику. ... **Амбициозные задачи**, стоящие перед **проектным направлением Прорыв** - замыкание ядерного топливного цикла на базе реакторов и топлива нового поколения, исключение аварий на АЭС, требующих эвакуации населения, решаются в **жестких условиях необходимости обеспечения конкурентоспособности создаваемой двухкомпонентной ядерной энергетики**».

между объектами, а не предоставляет решение - достоверные причинно-следственные связи, и не выполняет логические выводы.

³ Приведен снимок экрана, на котором изображен граф онтологии, автоматически построенной по полному тексту при помощи программы «Сервис визуального онтологического анализа научно-технических текстов» [11]. Вершина графа соответствует сущности и помечается ее именем – понятием, именем объекта, единицей измерения и т.п., дуга соответствует отношению и помечается именем класса отношения, рядом с которым в квадратных скобках указывается нормализованная лингвистическая конструкция, по которой был определен класс. При рисовании графа использовался алгоритм укладки вершин методом решения гравитационной задачи n-тел. Подробнее о визуализации графов онтологий изложено в [4].

⁴ Полу жирным выделены слова и словосочетания, соответствующие элементам графа.

Заключение

Технология построения семантического образа полного текста документа позволяет представить семантику описанного в документе отдельного решения системой понятий и связей и использовать при поиске завершенные смысловые конструкции. Технологической основой такого подхода является графовое представление семантических образов ПрО, используемого в качестве упорядоченного множества точек входа в информационный ресурс, обеспечивающих задание направленности поискового процесса и его контроль за счет явно формируемого и схематично представляемого пользователю контекста ситуации.

Использование графового представления онтологии документа в процессе информационного поиска позволяет выявлять и синтезировать семантические пути и/или окрестности, а применение аспектных проекций позволяет сократить размерность графа в заданном аспекте.

Совместное применение различных механизмов поиска, механизмов количественного анализа и механизмов поиска на графах обеспечивает взаимодополнительность подходов, позволяет повысить восприятие и глубину понимания проблемы и улучшить результаты поиска, переходя на уровень релевантных фрагментов.

Работа выполнена при поддержке Министерства науки и высшего образования РФ (проект государственного задания № 0723-2020-0036).

Литература

- [1] Голицына О.Л., Максимов Н.В., Окропишина О.В., Строгонов В.И. Онтологический подход к идентификации информации в задачах документального поиска // Научно-техническая информация. Сер. 2. 2012. № 5. С. 1–10.
- [2] Максимов Н.В. Методологические основы онтологического моделирования документальной информации // Научно-техническая информация. Сер. 2. 2018. № 3. С. 6–22.
- [3] Новая философская энциклопедия. В четырех томах. / Ин-т философии РАН. Научно-ред. совет: В.С. Степин, А.А. Гусейнов, Г.Ю. Семигин. М., Мысль, 2010.
- [4] Максимов Н.В., Голицына О.Л., Монанков К.В., Гаврилкина А.С. Методы визуального графоаналитического представления и поиска научно-технических текстов // Научная визуализация. 2021. Т. 13, № 1. С. 138–161.
- [5] Голицына О.Л., Максимов Н.В., Окропишина О.В., Строгонов В.И. Онтологический подход к идентификации информации в задачах документального поиска: практическое применение // Научно-техническая информация. Сер. 2. 2013. №3. С. 1–8.
- [6] Максимов Н. В. Документальная информационно-аналитическая система xIRBIS (редакция 6.0): программа для ЭВМ. / Максимов Н.В., Голицына О.Л., Монанков К.В., Гаврилкина А.С. // Свидетельство о гос. регистрации №2020661683 от 29.09.2020.
- [7] Максимов Н.В., Голицына О.Л., Монанков К.В., Лебедев А.А., Баль Н.А., Кюрчева С.Г. Средства семантического поиска, основанные на онтологических представлениях документальной информации // Научно-техническая информация. Сер. 2. 2019. №7. С. 8–19.
- [8] Голицына О.Л., Максимов Н.В. Модели информационного поиска в контексте поисковых задач // Научно-техническая информация. Сер. 2. 2011. № 2. С. 1-12.
- [9] Горбун Е.С., Максимов Н.В., Монанков К.В., Низаметдинов Ш.У. Модель и средства интерактивного анализа динамики и связей потоков публикаций научной информации // Научная визуализация. 2015. Т. 7(5). – С. 12–25.

- [10]Максимов Н.В., Гаврилкина А.С., Андропова В.В., Тазиева И.А. Систематизация и идентификация семантических отношений в онтологиях научно-технических предметных областей // Научно-техническая информация. Сер. 2. 2018. №11. С. 32–42.
- [11]Максимов Н. В. Опытный образец сервиса визуального онтологического анализа научно-технических текстов: программа для ЭВМ. / Максимов Н.В., Голицына О.Л., Монанков К.В., Гаврилкина А.С. // Свидетельство о гос. регистрации № 2021610648 от 15.01.2021.

Model of the Mechanism of Documentary Information Retrieval on Knowledge Graphs

N. Maksimov, O. Golitsina, K. Monankov

National Research Nuclear University MEPhI

The technology of representing the full text of a document by an ontology graph as a system of three systems is discussed: functional, conceptual and terminological. Functional relationships are classified according to the taxonomy of functional relationships and are used to construct aspect projections of ontologies. As the data model of the ontology, a labeled directed graph is used, which makes it possible to formalize operations on ontologies. Thus, the search image of a document is supplemented with a set of interrelated entities extracted from the text, which provides deep semantic indexing.

Representation of the semantic space of documents by an ontology graph allows reducing the problem of semantic search to set-theoretic models and using classical search mechanisms to search formally relevant documents. And the combined use of various classical search mechanisms, quantitative analysis mechanisms and graph search mechanisms provides complementarity of approaches, increases the perception and depth of understanding of the problem, and improves search results.

Keywords: semantic search, text processing, graph representations of ontologies, search mechanism, aspect projection

Reference for citation: Maksimov N., Golitsina O., Monankov K. Model of the Mechanism of Documentary Information Retrieval on Knowledge Graphs // Computer Linguistics and Computing Ontologies. Vol. 5 (Proceedings of the XXIV International Joint Scientific Conference «Internet and Modern Society», IMS-2021, St. Petersburg, June 24-26, 2021). - St. Petersburg: ITMO University, 2021. P. 39 – 50. DOI: 10.17586/2541-9781-2021-5-39-50

Reference

- [1] Golitsyna, O. L., Maksimov, N. V., Okropishina, O. V., & Strogonov, V. I. (2012). The ontological approach to the identification of information in tasks of document retrieval. *Automatic Documentation and Mathematical Linguistics*, 46(3), 125-132.
- [2] Maksimov, N. V. (2018). The methodological basis of ontological documentary information modeling. *Automatic Documentation and Mathematical Linguistics*, 52(2), 57-72.
- [3] Stepin, V. S., Guseynov, A. A., & Semigin, G. Y. (2010). *Novaya filosofskaya entsiklopediya. V chetyrekh tomakh [Sociology history and modernity]. / In-t filosofii RAN. Nauchno-red. M., Mysl, 4, 275-276. [in Russian].*
- [4] Maksimov, N. V., Golitsyna, O. L., Monankov, K. V., & Gavrilkina, A. S. (2021). Methods of visual graph-analytical presentation and retrieval of scientific and technical texts. *Scientific Visualization*, 13(1), 138-161.

- [5] Golitsina, O. L., Maksimov, N. V., Okropishina, O. V., & Strogonov, V. I. (2013). An ontological approach to information identification in tasks of document retrieval: A practical application. *Automatic Documentation and Mathematical Linguistics*, 47(2), 45-51.
- [6] Maksimov, N. V., Golitsyna, O. L., Monankov, K. V., & Gavrilkina, A. S. (2020). Document Information-Analytical System xIRBIS. State Registration Certificate, (2020661683). [in Russian].
- [7] Maksimov, N. V., Golitsina, O. L., Monankov, K. V., Lebedev, A. A., Bal, N. A., & Kyurcheva, S. G. (2019). Semantic Search Tools Based on Ontological Representations of Documentary Information. *Automatic Documentation and Mathematical Linguistics*, 53(4), 167-178.
- [8] Golitsina, O. L., Maksimov, N. V. (2011). Information retrieval models in the context of retrieval tasks. *Automatic Documentation and Mathematical Linguistics*, 45(1), 20-32.
- [9] Gorbun, E.S., Maksimov, N. V., Monankov, K. V., & Nizametdinov Sh.U. (2015). Model and tools for interactive analysis of dynamics and relations of scientific information publications flows. *Scientific Visualization*, 7(5), 12-25.
- [10] Maksimov, N. V., Gavrilkina, A. S., Andronova, V. V., & Tazieva, I. A. (2018). Systematization and identification of semantic relations in ontologies for scientific and technical subject areas. *Automatic Documentation and Mathematical Linguistics*, 52(6), 306-317.
- [11] Maksimov, N. V., Golitsyna, O. L., Monankov, K. V., & Gavrilkina, A. S. (2021). A Prototype of a Service for Visual Ontological Analysis of Scientific and Technical Texts. State Registration Certificate, (2021610648). [in Russian].