Научная визуализация, 2021, том 13, номер 1, страницы 138 - 161, DOI: 10.26583/sv.13.1.10

Методы визуального графоаналитического представления и поиска научно-технических текстов

Н.В. Максимов[1], О.Л. Голицына[2], К.В. Монанков[3], А.С. Гаврилкина[4]

Национальный исследовательский ядерный университет «МИФИ»

[1] ORCID: 0000-0002-8191-1521, nv-maks@yandex.ru
[2] ORCID: 0000-0002-3848-4755, olgolitsina@yandex.ru
[3] ORCID: 0000-0002-9267-3987, kmonankov@yandex.ru
[4] ORCID: 0000-0003-2167-1287, asgavrilkina@yandex.ru

Аннотация

Предложена технология построения и визуализации семантического образа полного текста документа, представляемого онтологией как системой трех систем: функциональной, понятийной и терминологической. Объектам и связям функциональной системы соответствуют извлеченные из текста имена сущностей и отношений, объектам понятийной системы – дескрипторы тезаурусов предметных областей. Проблема вариативного представления сущностей на знаковом уровне решается с применением правил формирования словосочетаний разной длины. 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, включающий вершины и дуги разных типов, что позволяет формализовать операции над онтологиями. Построение отображения элементов множеств онтологии на элементы графа таким образом, что элементы разных множеств разных систем различимы, узнаваемы и изображаются по-разному, позволяет реализовать принцип соответствия графического образа семантике визуализируемых данных.

Исходя из типологии поисковых задач предложены метафоры визуализации графа онтологии: метафора «поиска пути», характеризующаяся построением направленной цепочки фактов, и метафора «анализа окрестности», для которой характерно исследование окружения (контекста) факта.

Разработана технология и программное обеспечение для построения и вариантной визуализации графа онтологии.

Приведены примеры использования предложенных моделей для информационного поиска по текстам документов.

Ключевые слова: семантический поиск, обработка текста, графовые представления онтологий, визуализация графов онтологий, метафора визуализации.

Введение

В статье рассматриваются вопросы интерактивного использования графовых форм онтологических представлений текстов в задачах информационной поддержки средствами информационно-поисковых систем (ИПС) документального типа одного из сложнейших видов деятельности человека – *научного исследования* – процесса выработки новых научных знаний, в результате которого происходит установление новых фактов и последующее их обобщение.

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 ни один документ не содержит в явной форме полного описания решения, и перед человеком встает задача формирования образа решения из фрагментов доступных ему решений схожих задач. Иными словами, решение задачи основной деятельности требует предварительного формирования образа решения. Такая ситуация особенно свойственна научным или инженерным задачам, задачам проектирования и т.д.

Рассматривая использование автоматизированных документальных ИПС как деятельность, замещающую основную деятельность человека, задача синтеза нового знания может быть представлена, как задача формирования образа решения в результате (и путем) построения[1] единого текста из фрагментов текстов релевантных документов. Такой текст (в форме тезисов, пояснительной записки, научной статьи и т.д.) представляет образ решения задачи основной деятельности[2].

Знания, как объект деятельности человека, достаточно адекватно представляются онтологическими средствами, поскольку такие средства, согласно [1, 2], отражают не только имманентные и ситуативные связи предметной обрасти (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.

Такой подход вполне соответствует следующей экспликации Ч.С. Пирсом процесса познания [3]: «Эффективное рассуждение – живой процесс, обучение которому разрушает дисциплинарные барьеры. Логика имеет дело не с формами мысли или слова, а с общенаучными принципами, превращающими рассуждение в самоконтролируемый процесс, эффективный для достижения цели научного исследования. Рассуждение не может быть сведено к сугубо символическим преобразованиям, но включает наблюдение над диаграмматическими иконическими репрезентациями. Это наблюдение приводит нас к постановке эксперимента над графом. А именно сначала мы дублируем некоторые части графа, затем стираем

[1] Такое построение производится в соответствии с некоторой целью, методологической схемой и критериями оценки результата.

[2] При этом само решение человек формулирует в своем сознании.

некоторые его части, т.е. скрываем от наблюдения некоторую часть утверждения. Мы наблюдаем результат этого эксперимента – это и есть наше дедуктивное заключение».

И этот подход – интерактивная визуализация и преобразование семантических графов как онтологических образов документальной информации, – позволяет практически реализовать давно сформулированное требование: «…документальная ИПС должна быть организована таким образом, чтобы *человек* мог как бы *исследовать* поисковый массив, изменяя формулировку поискового предписания в зависимости от промежуточных результатов поиска» [4].

Как видно, идеи сформулированы давно (особенно относительно темпов развития сферы ИТ), но их воплощение потребовало не только увеличения мощностей, емкостей ИТ и использования эффективных средств визуализации, но и выявления фундаментальных положений, связывающих понятия таких методологически не столь близких областей, как информатика, лингвистика, деятельность, психология.

В статье в контексте основных положений информационного поиска будут рассмотрены:

- модели, средства и технологии построения семантических поисковых образов по полным текстам документов, обеспечивающих возможности формального анализа и синтеза графовых конструкций, отвечающих когнитивным ситуациям;
- модели и средства вариантной визуализации, обеспечивающие возможности снижения размера графа до операбельного уровня и представления в соответствии с выбранной когнитивной метафорой;
- когнитивно-подобные средства поиска на полных текстах, обеспечивающие за счет операций смыслового масштабирования и построения аспектных проекций контролируемое манипулирование компонентами графа и соответствующими фрагментами текста.

Приведенные в статье модели и средства вариантной визуализации графа онтологии опираются на базовые принципы построения визуальных моделей [5-7]:

- принцип соответствия решаемой пользователем задаче, который, в частности, отвечает положениям соответствия графического образа семантики визуализируемых данных индивидуальным особенностям восприятия;
- принцип обеспечения динамизма отображения;
- принцип минимизации временных затрат на анализ данных, в том числе за счет выбора оптимальных алгоритмов;
- принцип целостности (системности) представления;
- принцип независимости уровней – физического, логического и представления.

1 Построение семантического поискового образа полного текста документа

1.1 Основные положения информационного поиска

1. Понятие поиска информации всегда, так или иначе, связывается с процессом, имеющим неопределенность исхода, и, если это управляемый поисковый процесс, – с выбором, который, в свою очередь, строится на основе сопоставления данных, полученных извне, с наличными знаниями. Неопределенность (неполнота и неточность) выбора обусловлена последовательными преобразованиями (*понимание – выражение - формализация*) в связанных посредством ИПС цепочках *«знания – информация - документ – поисковый образ документа»* и *«проблемная ситуация – задача – вопрос - поисковый образ запроса»,* каждое из которых привносит свою неопределенность.

2. Вследствие принципов организации вычислительной среды поиск на физическом уровне сводится к операции (или их последовательности для композиционного запроса) полного или частичного сравнения заданного термина (точнее подстроки) с терминами (индексами) базы данных (БД). То есть необходимо понимать (и принимать как должное), что поисковые механизмы ИПС не имеют средств додумывания, угадывания или интерпретации введенного термина. Отметим, что такого рода инструменты, и в частности, технологии расширения запроса (с использованием тезаурусов, лингво-процессоров, статистических связей и т.п.) относятся к уровню логики, поскольку определяются и зависят от особенностей понятийно-знаковой системы ПрО. Но главное, что их использование в автоматическом режиме на практике скорее ухудшает интегральные показатели эффективности поиска. Несколько более высокая эффективность может быть достигнута в режиме интерактивного использования лексикографических справочников для предметно-ориентированного подбора лексики. Но опять же, это будет выбор из устойчивого и предопределенного множества, и, вероятно, не содержащего новую или альтернативную лексику.

3. Общей основой информационного документального поиска является координатное индексирование – способ выражения основного смыслового содержания документа или запроса в виде совокупности ключевых слов (терминов, термов), причем изначально считается, что ключевые слова не связаны между собой, а отдельному термину и документу соответствует точка в n-мерном семантическом пространстве, что, собственно, полностью соответствует двоичной форме представления информации в вычислительной среде.

4. В совокупной человеко-машинной информационной системе «*источник информации – ИПС – потребитель информации*» ИПС по существу (в итоге – физически) исполняет роль коммутатора. Источник (документ) будет передан потребителю, если его поисковый образ будет отвечать критерию формальной релевантности, используемому данной системой. Но и поисковый образ документа, и поисковый образ запроса – это описательные выражения – образы знания (состоявшегося и искомого), построенные с помощью языка, допускающего вывод результата сравнения путем вычисления значения некоторой меры близости и соотнесения его с пороговой величиной. Наиболее распространенным и адекватным языком является информационно-поисковой язык (ИПЯ) дескрипторного типа, словарный состав которого представлен некоторым множеством термов, а грамматика отражает способ построения поискового образа путем координации (взаимосвязи) термов [4].

При этом, в общем случае, термами могут быть как имена сущностей, так и имена отношений, а координация может реализовываться позиционным (словосочетания, фразы) или ключевым (использованием реляторов – операторов-связок либо отношений) способом.

Соответственно, координация может иметь разную глубину выражения смысла:

- — термины комбинаторным путем задают характеристические свойства описываемого объекта (обязательность/необязательность, взаимозаменяемость, ассоциативность и т.п.), что хорошо соответствует булевой алгебре, когда смысл определяется предикатом в виде логической формулы на множестве терминов;
- — термины поискового образа представляются «в контексте», позволяющем конкретизировать смысл;
- — термины поискового образа представляются «в последовательности», позволяющей представить (сформировать образ) смысл и ход решения практической задачи.

5. Процесс поиска, в целом, построен по простой схеме «запрос-ответ» и включает три основных технологических операции: формирование запроса пользователем,

формирование выдачи системой, оценка релевантности выдачи пользователем[3]. Однако необходимо учитывать следующие особенности. Во-первых, процесс поиска будет итеративным и не одноактным. Во-вторых, при адаптации выражения запроса должны быть семантически сопряжены объекты трех пространств: ментального, операционного/интерфейсного, машинного. При этом определяющим будет именно формирование выражения запроса, которое по существу сводится к выбору терминов и, возможно, их связыванию. Такой выбор производит пользователь, обращаясь либо к своему сознанию (знание лексики ПрО), либо, если система интерактивная, выбирая из технологических объектов, генерируемых системой. Такими технологическими объектами могут быть линейные упорядоченные структуры (словари БД, словники и т.д.) или более сложные семантические структуры (тезаурусы, семантические сети, когнитивные карты и т.д.). Очевидно, что навигация по таким сложным и объемным структурам будет не менее сложной, чем по содержанию самих документов. Поэтому для эффективного использования таких структур необходимо иметь средства их упорядочения, предметно-определенного членения, а также навигации и управления отображением.

6. 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.

1.2 Даталогическая модель семантического образа

Согласно [8], 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 можно выделить 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;
- завершенный факт (высказывание, утверждение, описание) – сеть элементарных и/или ситуативных фактов, образующая целостность,

[3] Следует обратить внимание на то, что конечной целью поиска является не только решение задачи с помощью найденной информации, но и подтверждение полноты итоговой выдачи (уверенности в том, что нет других, альтернативных путей и решений). И это отдельная ветвь технологии.

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 подход 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.

В [1] онтология, как семиотически целостное образование, определяется с позиций общей теории систем как совокупность трех взаимосвязанных систем $O =< S_f, S_c, S_t \equiv >$, где

S_f – функциональная система (объекты и связи действительности), определенная как $S_f =< M_f, A_f, R_f, Z_f >$, 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 $S_c =< M_c, A_c, R_c, Z_c >$, 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 $S_t =< M_t, A_t, R_t, Z_t >$, 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.

Представление онтологии на структурном уровне в виде графов позволяет формализовать операции над онтологиями на основе теоретико-графовых аксиом. Основными операциями при этом являются: бинарные – объединения, пересечения, проекции и унарная – масштабирования онтологий [1, 9].

В качестве модели данных функциональной системы онтологии используется помеченный (для вершин и дуг которого определены свойства A_f) ориентированный[4]

[4] Отметим, что ориентированность графа онтологии (в первую очередь это относится к функциональной системе) определяется не только ориентированностью дуг, но и смысловой «направленностью», отражающей эволюцию смыслового образа объекта/результата. Это означает, что имя отдельной сущности или отношения в графе будет

граф $G(V, E) = <V, E>$, где V – множество вершин, а E – множество дуг. Множество вершин и множество дуг в совокупности соответствуют множеству элементарных фактов. Т.к. 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.

Согласно [10] на множествах V и E определяются (могут быть динамически построены):

1. Метаграф, формально задаваемый как $MG = <V, MV, E, ME>$, где V – множество вершин, MV – множество метавершин, E – множество дуг, ME – множество метадуг. Каждая метавершина соответствует метасущности ситуативного факта и представляет собой граф $mv_i = <V_i, E_i>$, где $V_i \subset V, E_i \subset E$, а метадуга – метаотношению. Метаграф также обладает свойством мультиграфа, т.к. ситуативные факты могут, например, различаться только метаотношениями.

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 семантического эквивалента смысла, определяемого неатомарной конструкцией (выражением).

2. Гиперграф, когда на множестве вершин $V \cup MV$ формируется множество гиперребер W, при этом в основе правил задания гиперребра лежат множества A_f и Z_f:

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

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.

1.3 Применение операции проекции для построения аспектного представления

Аспектное представление, как одна из форм завершенного 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 – $O = \langle S_f,\ S_c,\ S_t,\ \equiv\rangle$ и аспектной – $O_i = \langle S_f^i,\ S_c,\ S_t,\ \equiv\rangle$ 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[5] определяется в соответствии с моделью деятельности и задается на таксономии функциональных отношений, классы отношений которой связаны с лингвистическими конструкциями в тексте [10], дополненной множеством структурно-

---

представлено в нескольких «экземплярах». В этом случае можно утверждать, что и понятия, и отношения выступают как лингвистические переменные, конкретный смысл которых доопределяется ситуацией – хорошо определенным контекстом (это также управляемый контекст: через задание аспекта и/или параметр концептуальной глубины и/или широты), специфицируемым типами отношений и характером связанных сущностей. И, поскольку граф представляет *целенаправленный процесс*, должен быть определен порядок следования вершин, в том числе могут быть заданы исходные и конечные (целевые) вершины.

[5] Аспектные представления являются одной из методологических основ синтеза знания. В основе синтеза знаний как самоорганизующегося процесса лежит структурная особенность системы – сложная система может быть описана при помощи набора относительно независимых аспектных представлений. Причем в процессе декомпозиции не только выделяются и связываются составляющие, но и формируется *схема декомпозиции* – система характеристических признаков деления.

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 этом случае в формировании проекции будут участвовать только имена заданных сущностей.

1.4 Индексирование фактов

Основой информационного поиска являются технологии индексирования. Традиционно (и вполне обоснованно) в качестве поисковых индексов используются имена понятий (или объектов, свойств и т.д.), выделяемые из текста. Такие индексы ориентированы на использование ИПЯ дескрипторного типа, рассматривающих в качестве операционных объектов линейные поисковые образы документов. Семантическая сила таких языков ограничена наличием синонимии, полисемии и омонимии в естественном языке и отсутствием средств выражения ситуативных и имманентных связей между реальными объектами, процессами и т.п., представленными на вербальном уровне в тексте.

Еще в [11] была предложена семантическая классификация ИПЯ на основе двух классификационных уровней: парадигматическом и синтагматическом. На парадигматическом уровне в классификации представлены классы языков, в которых отсутствуют средства выражения имманентных отношений; языков, в которых имеются средства выражения части имманентных отношений, и языков со всеми имманентными отношениями рассматриваемой ПрО. На уровне синтагматическом – классы языков, в которых отсутствуют средства выражения ситуативных отношений; языков, в которых есть средства для выражения ситуативных отношений, но нет средств для их различения, и языков, в которых ситуативные отношения выражаются и различаются.

Формирование онтологии как поискового образа сетевой организации требует:

1. Задать понятийную систему онтологии с множеством парадигматических отношений.
2. Представить текст документа в виде совокупности элементарных фактов. На этом этапе формируются, помимо имен сущностей, ситуативные отношения, которые могут быть типизированы в соответствии с таксономией, предложенной в [10]. Анализ имен сущностей позволяет сформировать дополнительно структурно-лингвистические отношения на основе распознавания аббревиатур, единиц измерения, членения длинных словосочетаний по правилам естественного языка и т.п., а также определить имена сущностей (понятий), являющихся точками входа в понятийную систему онтологии.

Выражение имен сущностей и отношений на знаковом уровне позволяет индексировать элементарный факт как триплет – последовательность знаков, в которой представлены не только имена, но и типы сущностей и отношений. Таким образом, могут быть построены как традиционные индексы (по ключевым словам), так и индексы, представляющие семантические связи. Наличие таких индексов позволяет в рамках традиционной теоретико-множественной модели информационного поиска (и средствами традиционного дескрипторного ИПЯ) реализовать отбор документов с учетом имманентных и ситуативных отношений между сущностями. ИПЯ при этом попадает в семантической классификации в класс языков, имеющих средства выражения (и различения) и имманентных, и ситуативных отношений.

1.5 Этапы построения семантического образа документа

Построение семантического образа документа как множества элементарных фактов, формирующих узлы и дуги графа онтологии, основывается на классической схеме семантического анализа текстов, которая включает этапы графематического, морфологического, семантико-синтаксического и концептуального анализа [12].

На этапе графематического анализа традиционно проводится выделение структурных элементов текста (разделов, глав, абзацев, заголовков), разбиение текста на токены, которые идентифицируются и (в случае необходимости) объединяются с помощью словарей и лингвистических правил. Выявляются именные группы, даты, числа с плавающей точкой, аббревиатуры, единицы измерения. Определяются границы предложений по знакам препинания с учетом идентифицированных специфических последовательностей символов.

В задачу этапа морфологического анализа входит определение основных морфологических характеристик (часть речи, род, число, падеж) токенов, идентифицированных как слова.

Этап семантико-синтаксического анализа начинается со снятия морфологической неоднозначности, порожденной на этапе морфологического анализа. Осуществляется выбор единственной парадигмы слова на основании анализа контекстного окружения и применения правил русского языка.

На этапе семантико-синтаксического анализа осуществляется формирование элементарных фактов на базе лексико-синтаксических шаблонов. В основе алгоритма формирования элементарных фактов лежит представление отдельного предложения в виде линейной последовательности отрезков, каждый из которых идентифицируется как «имя субъекта/объекта» или «связь (часть связи)». Например, предложение «Кронштейны нижней проставки опираются на роликовые опоры, установленные на перекрытие» будет поделено следующим образом:

(Кронштейны нижней проставки) < имя субъекта/объекта>|

(опираются) < связь (часть связи)>|

(на) < связь (часть связи)>|

(роликовые опоры) < имя субъекта/объекта>|

(установленные) < связь (часть связи)>|

(на) < связь (часть связи)>|

(перекрытие) < имя субъекта/объекта>

Язык описания шаблонов позволяет для триплета <субъект(S)><связь(L)><объект(O)> указать последовательности фрагментов предложения, которые должны определять (или входить в состав) каждую из частей триплета. Множество шаблонов может формироваться в зависимости от вида обрабатываемых текстов. Для приведенного примера будут построены триплеты:

<Кронштейн нижняя проставка (S)><опираться(L1)><на(L2)><роликовый опора (O)>

<роликовый опора (S)> <установить(L1)> <на(L2)> <перекрытие (O)>

Таким образом, линейный текст преобразуется в совокупность триплетов, формирующих узлы и дуги графа функциональной системы онтологии.

На этапе концептуального анализа решаются следующие задачи:

1. Классификация сформированных связей (отношений) в соответствии с таксономией отношений [13] и определение модальных свойств отношений с использованием морфологических характеристик и сигнальных слов. Модальности могут иметь следующие значения: *Достоверное (актуальное) / Предполагаемое (возможное) / Невозможное* и *Выполняющееся / Состоявшееся / Ожидаемое*.
2. Выявление имен субъектов/объектов (или частей имен), входящих в состав понятийной системы онтологии. Для таких понятий формируются самостоятельные узлы и дуги. Узлы в дальнейшем служат входами в граф

понятийной системы, а дуги представляют соответствующие структурно-лингвистические отношения.

3. Выявление частей имен субъектов/объектов, являющихся аббревиатурами, именными группами, и формирование дополнительных узлов и дуг, соответствующих структурно-лингвистическим отношениям.
4. Выявление имен субъектов/объектов (частей имен), являющихся единицами измерения. Создаются дополнительные узлы для единиц измерения и для имен свойств, которые определяются в соответствии с таксономией свойств и единиц измерения [14]. Имя свойства и соответствующая единица измерения соединяются дугой.
5. Формирование дуг структурно-лингвистических отношений по принципу лексикографического включения между именами субъектов/объектов.
6. Выявление имен субъектов/объектов, содержащих имя более чем одной сущности. Происходит деление таких имен на два и более в соответствии с правилами формирования словосочетаний, создаются дополнительные узлы (например, для имени «Кронштейн нижняя проставка» будет сформировано два дополнительных узла – «Кронштейн» и «Нижняя проставка»), которые соединяются с исходными узлами дугами, соответствующими структурно-лингвистическим отношениям.
7. Определение частотных характеристик имен субъектов/объектов. Расчет веса узлов на основе частоты встречаемости, роли, принадлежности значимым фрагментам текста.

В результате последовательного выполнения этапов формируется полный граф функциональной системы онтологии, служащий основой семантического образа документа.

2 Визуализация графа онтологии

Проблема визуализации графа онтологии обусловлена тем, что, с одной стороны, элементарный факт должен иметь визуальный образ, позволяющий различить отдельные сущности и отношения, а с другой – визуальный образ ситуативного или завершенного факта должен быть доступен для восприятия целиком и умещаться в пространство экрана. Онтология полного текста документа, как правило, характеризуется большой мощностью элементов (это могут быть тысячи сущностей и отношений даже для сравнительно небольшого текста), что предопределяет необходимость создания и использования инструментов отбора и визуализации фрагментов графа онтологии, адекватных типу решаемой задачи.

Процесс визуализации любых данных направлен на представление их в удобной для зрительного наблюдения и анализа форме, а эффективность методов визуализации во многом обеспечивается осмысленным использованием принципов восприятия информации, адекватным выбором метафоры и модели визуализации, соблюдением критериев визуализации.

В настоящей работе метафора визуализации понимается, как отображение множества объектов пространства данных исходной задачи во множество объектов пространства представления путем переноса признаков объектов первого множества на объекты второго интуитивно понятным образом [7, 15].

Некоторые примеры метафор визуализации подробно рассматриваются в работах [7, 15, 16, 17], в частности, такие как: метафора рабочего стола, комнаты, здания, молекулы. Для объектов, представляемых графом, часто используют цвет и размеры элементов графа для создания визуальных маркеров, позволяющих фокусировать внимание пользователя. Например, для когнитивных карт [16] цвет дуги соответствует типу связи, толщина – ее интенсивности, а цвет и размер вершины зависит от типа вершины или ее принадлежности к смысловой группе.

В контексте задач документального информационного поиска с точки зрения отображения графа онтологии на схему ситуации когнитивного состояния пользователя выделены две метафоры: метафора поиска пути и метафора анализа окрестности.

К критериям визуализации графа относят [18, 19]:

- различимость отдельных элементов графа (вершин, дуг, меток);
- удобная с точки зрения визуального восприятия укладка вершин на плоскости;
- приемлемое время ожидания и, как следствие, приемлемая (близкая к линейной) вычислительная сложность алгоритмов укладки вершин;
- сохранность ментальной карты графа, задающей требование схожести укладок в случае незначительного изменения графа.

На критерии влияют правила визуализации, которые часто называют эстетическими критериями [16, 18, 20]:

- минимизация размеров занимаемой области;
- равномерность распределения вершин на плоскости;
- минимизация числа пересечений дуг (приближение к планарному отображению);
- минимизация суммарной длины всех дуг;
- минимизация числа сгибов дуг;
- направленность дуг («сверху-вниз» и «слева-направо»);
- соблюдение симметрии.

Однако удовлетворить всем правилам обычно бывает невозможно. Во-первых, из-за противоречивости правил друг другу (например, минимизация числа сгибов дуг может нарушать равномерность распределения вершин на плоскости и приводить к увеличению размеров занимаемой области) [20]. Во-вторых, из-за возрастания вычислительной сложности алгоритмов.

В настоящей работе элементам онтологии (элементам множеств каждой из трех систем) в соответствие ставятся элементы графа – помеченные вершины и дуги разных типов, характеризующих происхождение элемента. Вершина графа изображается прямоугольником и помечается именем сущности – понятием, именем объекта, единицей измерения и т.п. Размер вершины (размер шрифта метки) вычисляется на основе веса соответствующего имени в тексте. Цвет вершины задается в зависимости от типа вершины или роли имени. Дуга графа изображается изогнутой кривой[6], что позволяет располагать метку дуги и метки инцидентных вершин на разных горизонталях, что в случае длинных имен меток позволяет уменьшить число их пересечений и избежать высокой концентрации пересечений. Рисование изогнутой кривой требует больше вычислительных ресурсов (соответственно, занимает больше времени), чем рисование прямой линии, однако получаемое изображение графа позволяет легче различать отдельные его элементы. Дуги графа помечаются разными способами, в том числе именем отношения или именем класса отношения, в зависимости от типа дуги и параметров визуализации.

2.1 Технология визуализации графа онтологии

Технология визуализации графа онтологии, построенной по тексту на естественном языке, включает в себя следующие этапы:

- отбор элементов графа в соответствии с задачей пользователя;
- формирование представления в соответствии с метафорой визуализации;

[6] В настоящей работе для рисования графа используется библиотека визуализации с открытым исходным кодом «vis-network.js» [21]. Библиотека реализует различные методы изображения дуг кривыми. Авторами выбран метод изображения дуг квадратичными кривыми Безье из эстетических соображений.

— формирование изображения в соответствии с моделью визуализации.

На этапе отбора элементов графа используются фильтры, задаваемые автоматически (в случае, например, формирования аспектной проекции) или вручную. Все фильтры разделены на три группы: фильтры множества вершин, фильтры множества дуг и фильтры, относящиеся к графу в целом. Фильтры множества вершин предполагают отбор вершин по имени сущности, типу вершины[7], местоположению в исходном тексте, роли термина[8], весу[9]. После применения фильтров вершин в графе останутся вершины, удовлетворяющие условиям фильтров, и дуги, инцидентные таким вершинам. Фильтры множества дуг предполагают отбор дуг по имени отношения, классу[10], модальности, местоположению в исходном тексте. После применения фильтров дуг в графе останутся дуги, удовлетворяющие условиям фильтров, и инцидентные им вершины. Фильтры, относящиеся к графу в целом, позволяют задать дополнительные условия отбора вершин и дуг. Например, оставить в графе вершины, которые не удовлетворяют фильтрам вершин, но входят с таковыми в одну компоненту связности.

Метафора визуализации обеспечивает отображение графа онтологии на схему ситуации когнитивного состояния и цели пользователя (нахождение решения практической задачи или анализ проблемной ситуации).

Модель визуализации (представления данных) задается множеством правил формирования визуальных объектов и их графических атрибутов, таких как цвет, форма, размер, и должна обеспечить вариантную геометрию сформированного множества дуг и вершин. В соответствии с моделью визуализации осуществляется укладка вершин графа на плоскости, т.е. вычисление их координат в пространстве.

Таким образом, технология визуализации обеспечивает построение динамической графической формы, которая предоставляет пользователю следующие возможности интерактивного взаимодействия:

— просмотр свойств вершин и дуг;
— перемещение вершин на плоскости;
— изменение, удаление и создание новых вершин и дуг;
— приближение и прокрутка графа, позволяющие взаимодействовать с отдельными фрагментами графа с разной степенью детальности;
— поиск вершин и перемещение фокуса на найденные вершины;
— переход от вершин к соответствующим фрагментам текста;
— построение пути между двумя вершинами, если такой путь существует;
— построение окрестности вершины;
— построение аспектных проекций графа;
— отбор элементов графа в соответствии с фильтрами;
— объединение и пересечение графов.

Этапы технологии соответствуют традиционному подходу к визуализации данных, предусматривающему выполнение следующих шагов: задание исходных данных, фильтрация, мэппинг и рендеринг – которые в совокупности называют конвейером визуализации [22]. Соответствие приведено на рис. 1.

---

[7] Тип вершины характеризует происхождение соответствующего имени: из текста, из тезауруса, из таксономии свойств и единиц измерения, часть более длинного термина и объединяющие вершины.

[8] Роль имени сущности определяется в соответствии с функциональной моделью, аналогичной IDEF0.

[9] Вес вершины рассчитывается на основе частоты встречаемости, роли, принадлежности значимым фрагментам текста.

[10] Класс отношения определяется в соответствии с таксономией классов функциональных отношений [13].

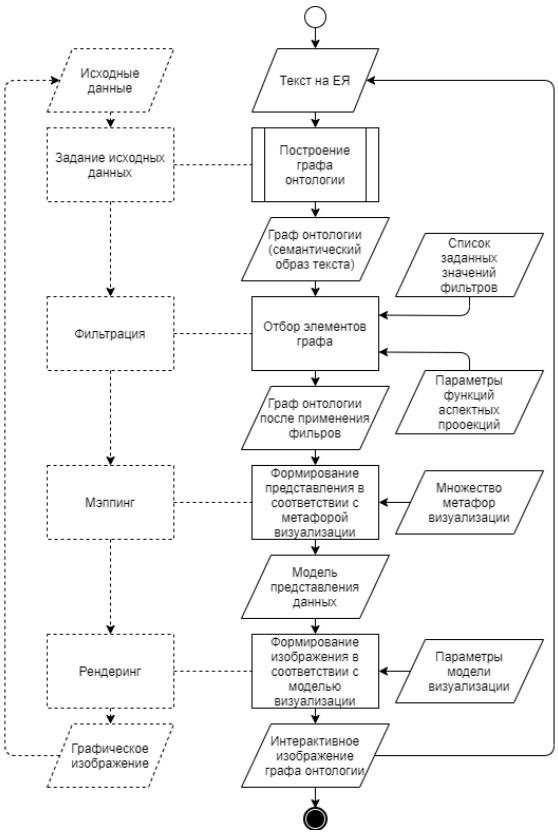

Рисунок 1 – Технология визуализации графа онтологии в соответствии с этапами конвейера визуализации, представленного в [22]

2.2 Модель визуализации

В настоящей работе модель визуализации графа онтологии соответствует способу отображения отдельного элемента онтологии и алгоритму укладки множества элементов на плоскости.

Типология алгоритмов укладки вершин графа на плоскости приведена в работе [23] и включает следующее:

— алгоритмы, использующие теорию графов и алгоритмы на графах;
— алгоритмы, использующие силовые (например, пружинные), термодинамические или биоинформатические модели (генетические алгоритмы) и другие симуляции;
— эвристические алгоритмы;
— комбинации вышеперечисленных алгоритмов.

В целом использование различных вариантов отображения во время решения пользовательской задачи позволяет взглянуть на ее возможные решения с разных сторон, в разных аспектах, что обеспечивает взаимодополнительность подходов и позволяет повысить восприятие и глубину понимания проблемы.

Для графа онтологии предлагается использовать следующие алгоритмы укладки вершин на плоскости:

— укладка вершин силовым методом Barnes-Hut;
— укладка вершин в порядке употребления (появления) имен сущностей в тексте;
— укладка вершин в соответствии со значимостью путей (длинной пути или суммарным весом вершин);
— укладка вершин в соответствии с некоторой схемой (например, функциональной моделью IDEF0).

Силовые алгоритмы основаны на физических аналогиях, однако на практике силовые алгоритмы неприемлемы с точки зрения времени построения укладки для больших графов [24], поэтому часто используют разные их модификации. Например, известный в астрофизике метод Barnes-Hut, сформулированный в терминах решения гравитационной задачи n-тел, позволяет ускорить укладку графа за счет аппроксимации сил отталкивания.

В работе рассматриваются следующие модели визуализации графа онтологии:

— модель визуализации кратчайшего пути между двумя сущностями, обеспечивающая представление цепочки элементарных фактов;

— модель визуализации путей на основе укладки с упорядоченностью по значимости, где в качестве значимости используется длина пути или суммарный вес вершин;

— модель визуализации окрестности сущности, основанная на укладке вершин силовым методом Barnes-Hut.

Рассмотрим подробнее модель визуализации путей с упорядоченностью по значимости. Для укладки вершин разработан следующий алгоритм:

1. Поиск множества путей между всеми вершинами графа (например, с использованием алгоритма Флойда-Уоршелла), длина или суммарный вес вершин которых превышает величину заданного порогового значения (по умолчанию эти параметры равны 1). Поиск проводится с учетом (или без) ориентации дуг.
2. Сортировка множества путей по убыванию длины пути или суммарного веса вершин.
3. Последовательное расположение на плоскости (присвоение вершинам координат) путей таким образом, что наиболее значимые пути располагаются выше, причем учитываются следующие особенности:
   — каждый путь укладывается слева направо, исходя из порядка следования составляющих его вершин;
   — вершины, входящие в несколько путей, не дублируются, а инцидентные им дуги крепятся к первому встретившемуся экземпляру вершины (такие дуги отражают связи между путями).
4. Анализ плотности подграфов и связности путей графа. Связность между двумя путями рассчитывается исходя из количества общих вершин и количества дуг между вершинами первого и второго пути. Пути, имеющие большую связность, располагаются рядом. Наиболее плотные подграфы и компоненты связности отделяются друг от друга на плоскости. Висячие вершины наиболее значимых путей приближаются на плоскости к ним.

Четвертый шаг алгоритма направлен на сокращение размеров занимаемой области, уменьшение числа пересечений дуг и сокращение их длины. Однако из-за высокой вычислительной сложности, которая приводит к увеличению времени ожидания визуализации, этот шаг не является обязательным.

Модель визуализации окрестности сущности обеспечивает представление всех выделенных из текста элементарных фактов, в которых участвует соответствующая вершине сущность. Последовательное применение модели позволяет выбирать следующую вершину для анализа и таким образом выбирать направление и формировать семантическую окрестность. Тем самым у исследователя формируется представление об искомом объекте, глубина и полнота которого определяется содержанием текста.

Рассмотрим алгоритм работы с использованием модели визуализации окрестности:

1. Отбор вершин графа онтологии по заданному имени сущности (может соответствовать выражению поискового запроса), в результате чего будет сформирован некоторый подграф, содержащий вершины $\{v_1, \ldots, v_n\}$.

2. Построение окрестности вершины $v_i \in \{v_1, \ldots, v_n\}$, в результате чего будет сформирован подграф, состоящий из вершин $\{v_1, \ldots, v_n\}$ и связанных с вершиной v_i в некоторой окрестности («радиус» окрестности определяет пользователь) вершин $\{v_{n+1}, \ldots, v_m\}$.
3. Выполнение п. 2 алгоритма для любой другой вершины из множества $\{v_1, \ldots, v_n, v_{n+1}, \ldots, v_m\}$.

3 Информационный поиск на графах онтологий

Задача семантического поиска может быть сведена к итеративному последовательному решению двух задач: классической задаче информационного поиска и глубинному анализу найденных документов с использованием графа онтологии документа в качестве интерактивного инструмента навигации по тексту и понятиям. Обобщенная схема поиска приведена на рис. 2. Традиционная схема дополнена этапами построения и анализа графа онтологии полного текста документа, комбинирования фрагментов текста (информационных блоков) и оценки результатов комбинирования.

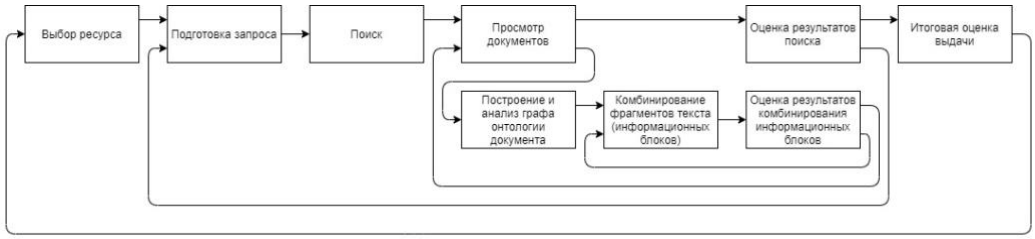

Рисунок 2 – Обобщенная схема поиска

Визуальный анализ графа онтологии документа позволяет обнаружить пути или компоненты связности, релевантные поисковой потребности пользователя. Переход от вершин графа к фрагментам исходного текста и комбинирование таких фрагментов позволяет обнаруживать новые знания, а также проверять на непротиворечивость уже имеющиеся.

3.1 Типы задач информационного поиска на графах онтологии

Форма представления результатов поиска должна быть адекватна характеру решаемой пользователем задачи, поэтому обеспечение человеку лучшего восприятия в зависимости от типа решаемой задачи является важным аспектом визуализации информации.

Задачи информационного поиска можно разделить на два типа – задачи поиска решения проблемы основной деятельности и задачи информационно-аналитического характера, такие как предварительные исследования, поиск гипотез или возможных путей решения.

Задачи первого типа предполагают поиск решения, которое всегда может быть представлено процессом, то есть направленной последовательностью событий и действий над объектами. Это предопределяет необходимость представления результатов поиска в форме, отражающей направленность (от исходных положений к «ответу») решения и представлять собой в идеале алгоритм этого решения. Для такого рода задач подходит метафора поиска пути, которая предполагает выстраивание последовательности пунктов, соответствующих объектам, событиям, действиям, выраженным понятиями (цепочки элементарных фактов), от опорных понятий к понятиям в контексте потенциального решения задачи основной деятельности. Формой представления результатов визуализации в графе онтологии, соответствующей этой метафоре, является путь от вершин исходных данных (опорных понятий, терминов запроса) к вершинам, содержащим понятия решения.

Для задач информационно-аналитического характера может быть использована метафора поиска окрестности, которая предполагает визуализацию контекста опорных понятий. Группирование вокруг опорных понятий позволяет пользователю рассматривать окрестность, углубляясь в изучение темы исследования.

Таким образом, информационный поиск на графах онтологий сводится к следующим схемам:

— поиск цепочки фактов, соответствующих фрагментам исходного текста, в совокупности содержащих решение задачи основной деятельности;
— поиск окрестности элементарного (ситуативного) факта, где в качестве отправной точки используется опорное понятие (вершина графа);
— комбинация первых двух схем.

Схема «поиск цепочки фактов» предполагает отбор вершин, выстроенных в ориентированную или неориентированную цепь (аналогия – поиск пути между вершинами) от исходного элементарного факта к целевому (аналогия – поиск причинно-следственных связей).

Схема «поиск окрестности» предполагает отбор вершин в окрестности исходного элементарного факта (аналогия – поиск в ширину в графе), и построение из отобранных вершин законченной конструкции – завершенного факта (аналогия – схема снежинки).

Комбинация схем «поиск цепочки фактов» и «поиск окрестности» предполагает последовательное или комбинированное их применение.

Для решения указанных задач разработана программа «Сервис визуального онтологического анализа научно-технических текстов» [25], реализующая рассматриваемые в статье модели и технологию визуализации. Программа включает компонент лингвистического разбора текста, использующий методы документальной информационно-аналитической системы «xIRBIS» [26] и лексикографическую БД [27]. Для рисования графа онтологии используется библиотека визуализации с открытым исходным кодом «vis-network.js» [21]. Фрагмент интерфейса разработанной программы приведен на рис. 3. Программа может функционировать как самостоятельно (в режиме веб-сервера) и применяться для целей семантического анализа текста, так и в составе системы «xIRBIS» [26], что позволяет применять разработанные модели и технологию непосредственно в процессе информационного поиска.

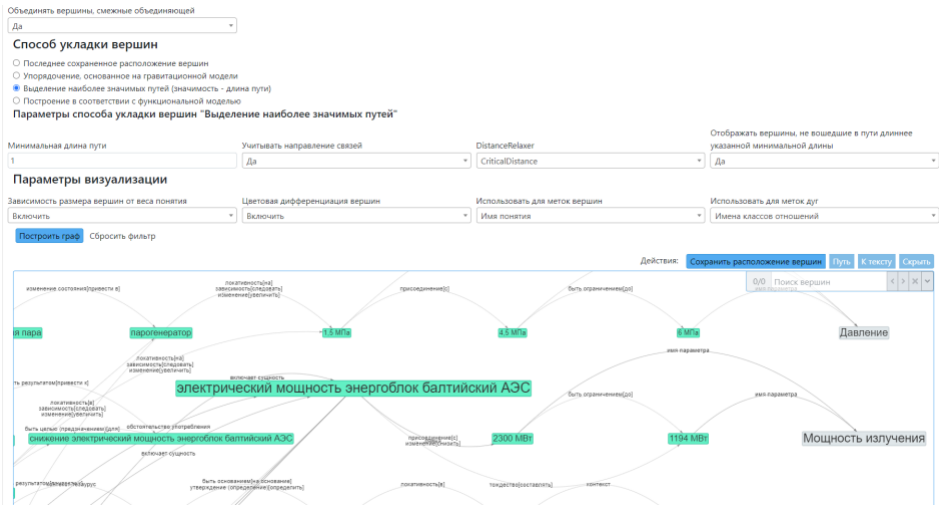

Рисунок 3 – Фрагмент интерфейса сервиса визуального онтологического анализа научно-технических текстов[11]

[11] Далее используется следующее изобразительное соглашение. Зеленые вершины имеют тип «из текста», синие – «из тезауруса», серые – «имена свойств», желтые – части более длинных терминов. В качестве меток дуг функциональных отношений используются имена классов отношений, рядом с каждым именем класса в квадратных скобках указывается нормализованная лингвистическая конструкция, по которой был определен класс.

3.2 Поиск пути на графе онтологии

Проиллюстрируем применение разработанных инструментов на примере проблемы неполной востребованности мощностей строящейся Балтийской АЭС. В результате информационного поиска в информационном ресурсе по проблеме найдены документы [28, 29. 30], релевантные фрагменты которых объединены в новый текст. По этому тексту построен граф онтологии[12], который ввиду большого объема здесь приводится не полностью. Фрагмент графа онтологии, содержащий вершину «Мощность», изображен на рис. 4.

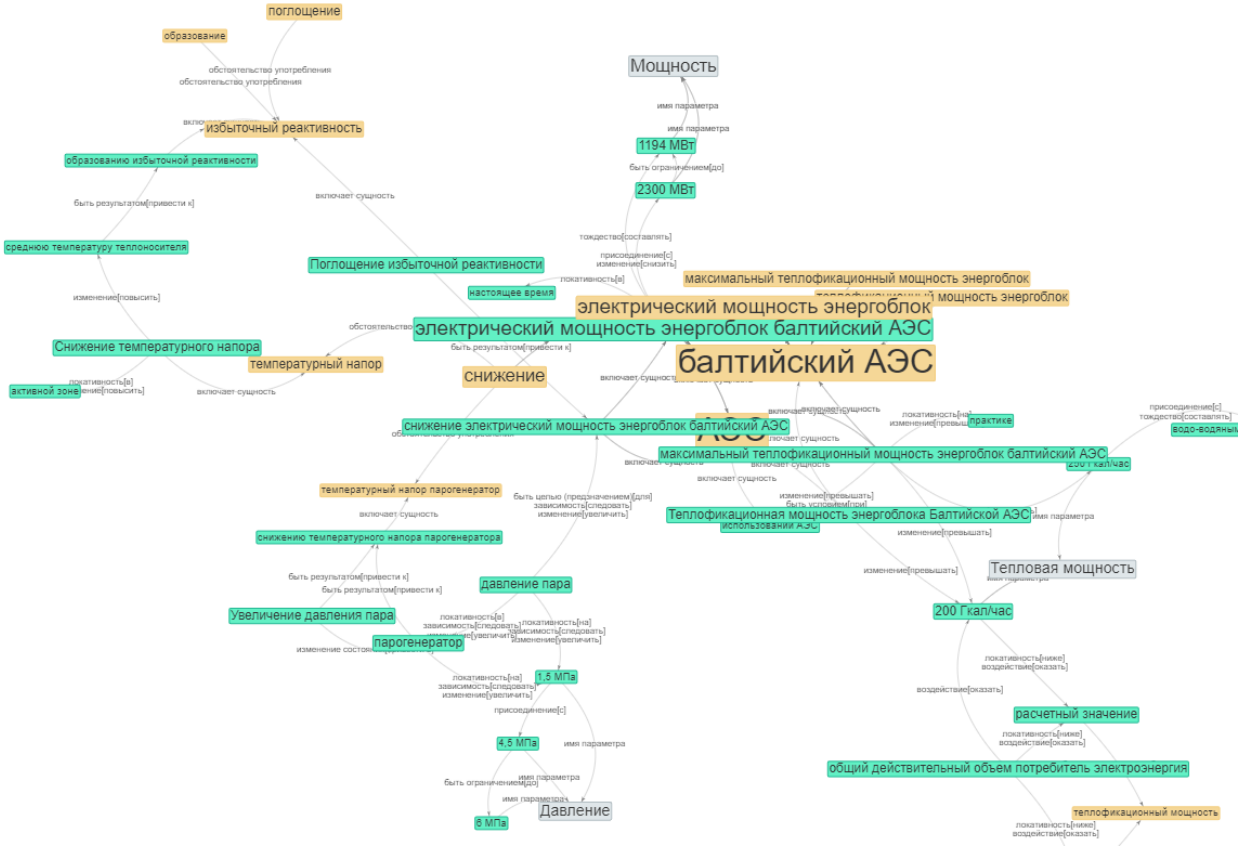

Рисунок 4 – Фрагмент графа онтологии текста «проблема неполной востребованности мощностей строящейся Балтийской АЭС» с укладкой вершин методом Barnes-Hut

В результате поиска вершин в графе по запросу «Мощность» найдено 12 вершин, одна из которых (серая вершина в верхней части рис. 4) выделена в соответствии с таксономией свойств и единиц измерения [14] из-за наличия вершин «1194 МВт» и «2300 МВт», содержащих единицу измерения мегаватт. Другие свойства, выделенные аналогичным образом – вершина «Давление» (серая вершина в нижней части рис. 4) и «Тепловая мощность» (серая вершина в правой части рис. 4). Построим кратчайший путь между вершинами «Давление» и «Мощность» (см. рис. 5, путь выделен красным цветом) с целью проследить связь между соответствующими параметрами.

Меткой «имя параметра» помечается дуга, связывающая величину из текста с соответствующим ей именем свойства из таксономии свойств и единиц измерений.

[12] Граф также может быть построен и путем объединения релевантных фрагментов графов онтологий текстов документов.

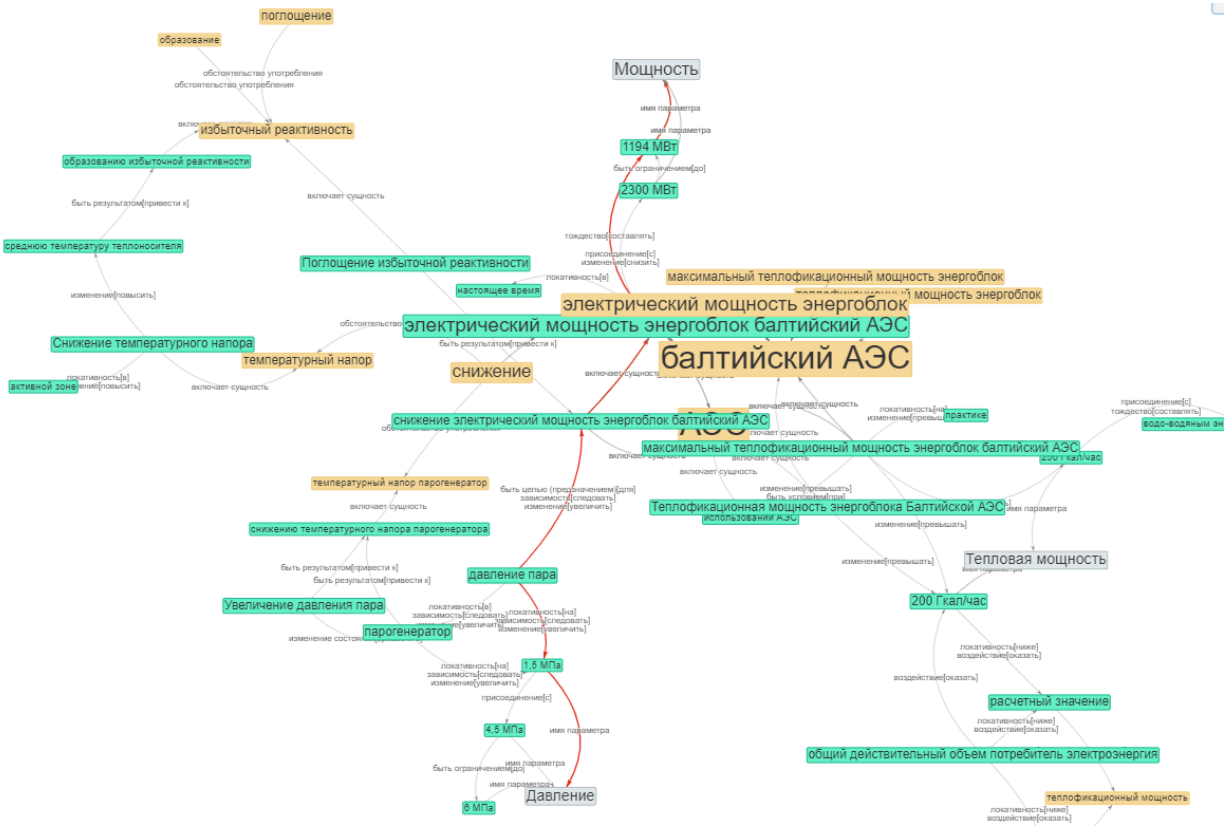

Рисунок 5 – Фрагмент графа онтологии текста «проблема неполной востребованности мощностей строящейся Балтийской АЭС» с укладкой вершин методом Barnes-Hut и кратчайшим путем (выделен красным) между вершинами «Давление» и «Мощность»

Путь (см. рис. 5) содержит элементарный факт «давление пара» – «быть целью (предназначением) [для] зависимость [следовать] изменение [увеличить]» – «снижения электрической мощности энергоблока Балтийской АЭС». Рассмотрим далее элементарный факт «давление пара» – «локативность [в]» – «парогенератор». Из этих двух элементарных фактов следует, что параметры давление пара парогенератора и электрическая мощность энергоблока связаны. Теперь необходимо установить характер связи.

Рядом с рассматриваемым путем выделим вершину «снижение» (желтая вершина в центре рис. 5, левее красного пути), которая визуально отличается цветом и размером шрифта, что является визуальным маркером и говорит о возможной значимости вершины.

Построим кратчайший путь между вершинами «Давление» и «Мощность» через промежуточную вершину «снижение» (см. рис. 6, путь выделен красным цветом).

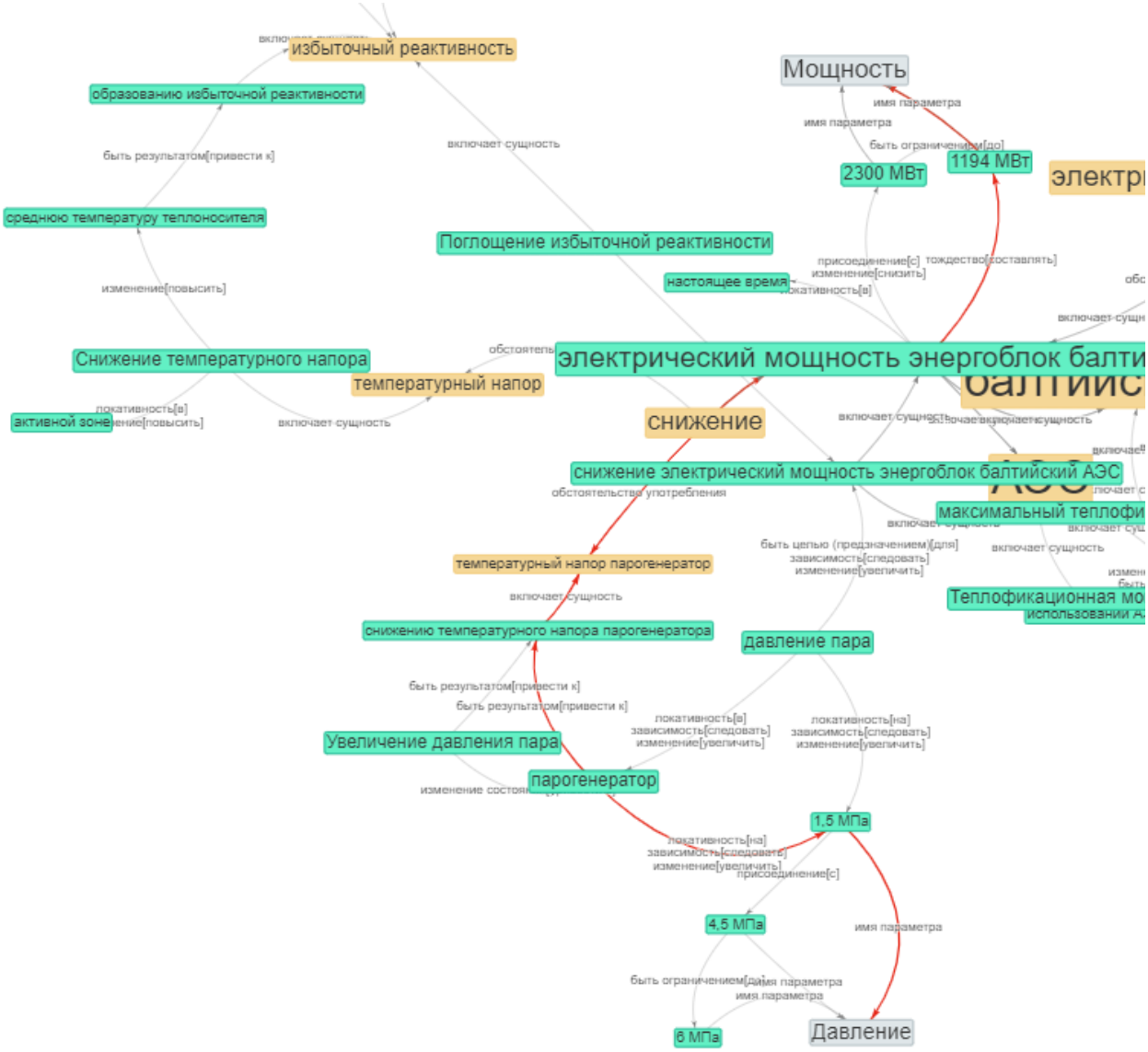

Рисунок 6 – Фрагмент графа онтологии текста «проблема неполной востребованности мощностей строящейся Балтийской АЭС» с укладкой вершин методом Barnes-Hut и кратчайшим путем (выделен красным) между вершинами «Давление» и «Мощность» через промежуточную вершину «Снижение»

Рассмотрим подробнее вершины на построенном пути (рис. 6) в контексте смежных вершин и увидим следующую цепочку элементарных фактов: «Увеличение давления пара» – «изменение состояния [привести в]» – «парогенератор» – «быть результатом [привести к]» – «снижению температурного напора парогенератора». То есть увеличение давления парогенератора приводит к снижению температурного напора. Далее на пути «температурный напор парогенератора» – «обстоятельство употребления» – «снижение» – «обстоятельство употребления» – «электрическая мощность энергоблока Балтийской АЭС». Рассмотрим подробнее обстоятельства употребления электрической мощности энергоблока Балтийской АЭС в контексте снижения. Вершина «снижение» имеет еще одну дугу «обстоятельство употребления», инцидентную вершине «температурный напор», которая через цепочку элементарных фактов связана с вершиной «избыточная реактивность».

Построим кратчайший путь между вершинами «Мощность» и «Давление» через промежуточные вершины «температурный напор» и «избыточная реактивность» (см. рис. 7, путь выделен красным цветом).

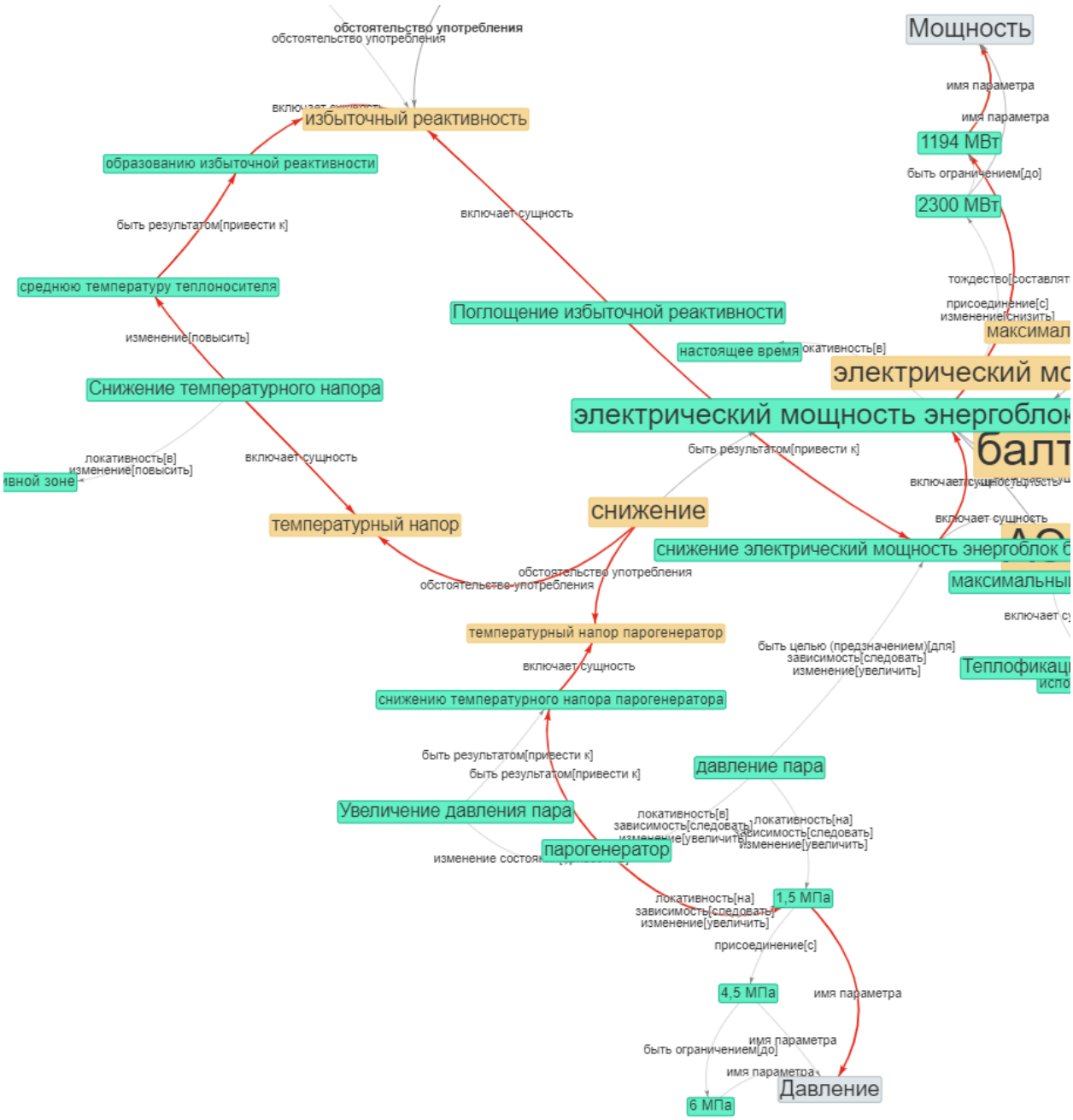

Рисунок 7 – Фрагмент графа онтологии текста «проблема неполной востребованности мощностей строящейся Балтийской АЭС» с укладкой вершин методом Barnes-Hut и кратчайшим путем (выделен красным) между вершинами «Давление» и «Мощность» через промежуточные вершины «Температурный напор» и «Избыточная реактивность»

Путь содержит следующую цепочку элементарных фактов: «снижение температурного напора» – «изменение [повысить]» – «среднюю температуру теплоносителя» – «быть результатом [привести к]» – «образование избыточной реактивности» – «включает сущность» – «избыточная реактивность» – «включает сущность» – «поглощение избыточной реактивности» – «быть результатом [привести к]» – «снижение электрической мощности энергоблока Балтийской АЭС».

«Прочтение» графа позволяет заключить, что увеличение давления пара в парогенераторе приведет к снижению температурного напора парогенератора, вследствие чего повысится средняя температура теплоносителя, что приведет к образованию избыточной реактивности. Поглощение избыточной реактивности приведет к снижению электрической мощности энергоблока Балтийской АЭС.

Таким образом, установлена связь между параметрами давления пара в парогенераторе и электрической мощности энергоблока и установлен характер связи – увеличение давления приведет к снижению мощности.

Далее можно интерактивно изменить укладку вершин (см. рис. 8), удалить лишние вершины и дуги и сохранить результат, сформировав таким образом завершенный факт по рассматриваемой проблеме. В будущем к нему можно будет обратиться и использовать для навигации по тексту. В частности, использование ссылки на завершенный факт в когнитивном рубрикаторе [31] позволяет реализовать принцип сохранения и накопления знаний.

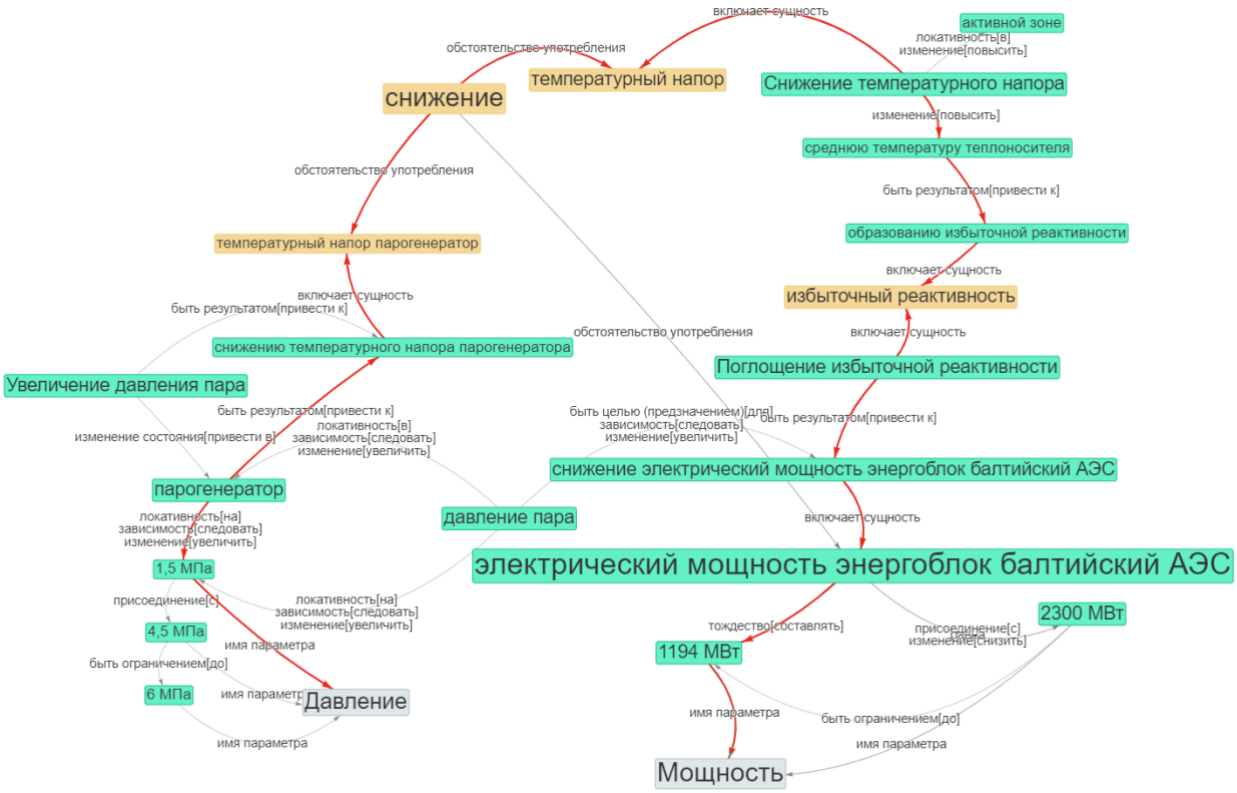

Рисунок 8 – Фрагмент графа онтологии текста «проблема неполной востребованности мощностей строящейся Балтийской АЭС» с интерактивной укладкой вершин

3.3 Анализ окрестности

Проиллюстрируем поиск по схеме «анализ окрестности». Построим граф по фрагменту текста документа конструкторской документации «Главный циркуляционный насосный агрегат» (ГЦНА), содержащего описание конструкции насоса. Проведем отбор вершин по имени сущности «ГЦНА» и построим окрестность вершины «ГЦНА» радиуса 1. Получим подграф, изображенный на рис. 9.

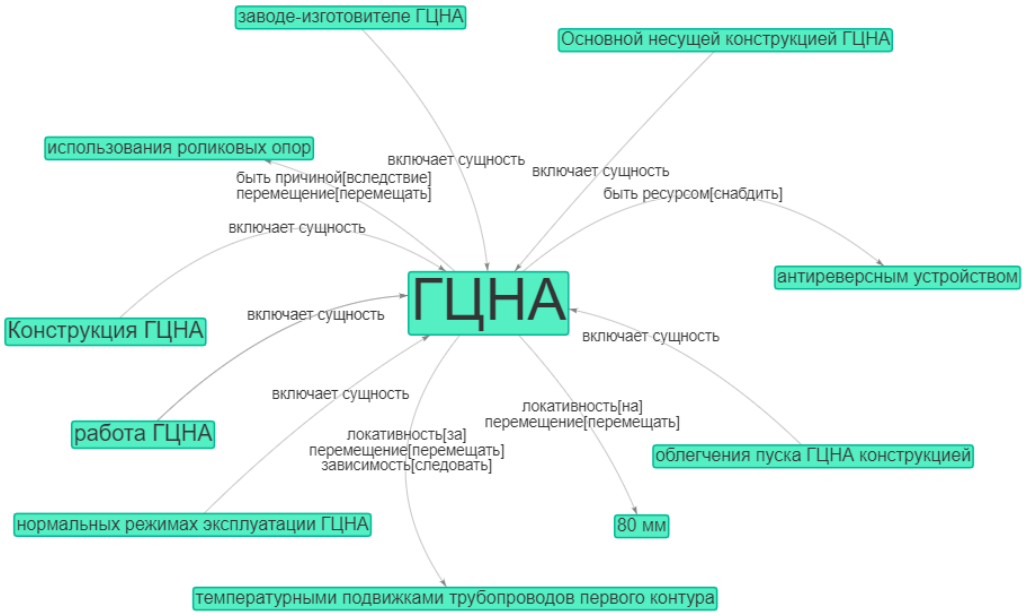

Рисунок 9 — Окрестность вершины «ГЦНА»

Все вершины сформированного подграфа (рис. 9) имеют происхождение «из текста», о чем говорит цвет вершин. Смежные вершины для вершины «ГЦНА» имеют одинаковый размер шрифта (размер зависит от веса), что говорит о равновесности имен сущностей в тексте.

Для элементарного факта «антиреверсное устройство» – «быть ресурсом [снабдить]» – «ГЦНА» построим окрестность вершины «антиреверсное устройство» (рис. 10).

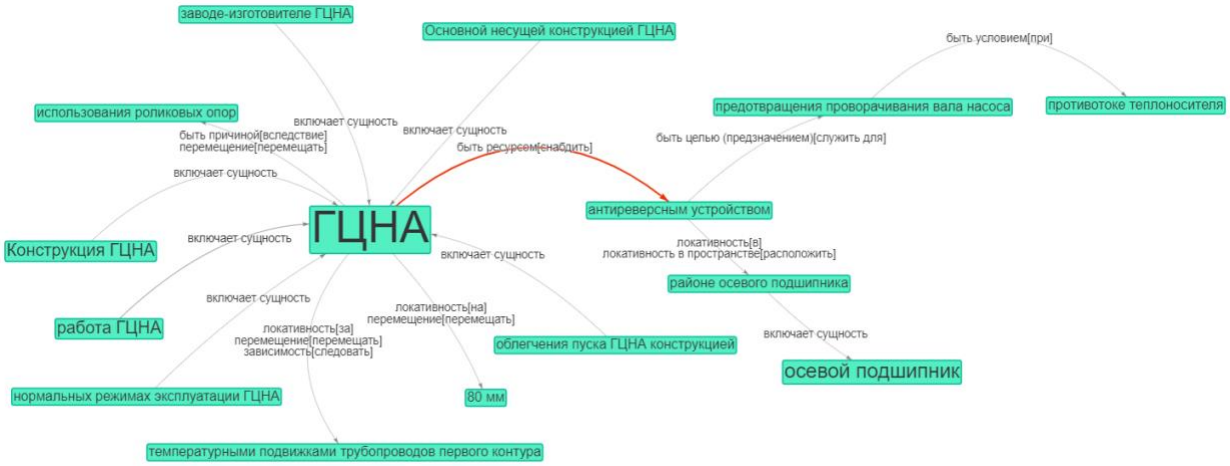

Рисунок 10 — Расширение окрестности термина «ГЦНА» в направлении «антиреверсного устройства»

Построенная окрестность включает две цепочки элементарных фактов, представляющих местонахождение и предназначение антиреверсного устройства:

— «антиреверсное устройством» – «быть целью (предназначением) [служить для]» – «предотвращение проворачивания вала насоса» – «быть условием [при]» – «противоток теплоносителя»;

— «антиреверсное устройство» – «локативность [в]» – «район осевого подшипника» – «включает сущность» – «осевой подшипник».

Вершина «район осевого подшипника» связана структурно-лингвистическим отношением «включает сущность» с вершиной «осевой подшипник». Причем вершина

«осевой подшипник» выделяется большим размером по отношению к смежным, что сигнализирует о ее большем весе и, соответственно, значимости в тексте.

Построим ситуативный факт для элементарного факта «район осевого подшипника» – «включает сущность» – «осевой подшипник» путем применения функции построения окрестности к вершине «осевой подшипник». Получим подграф, изображенный на рис. 11.

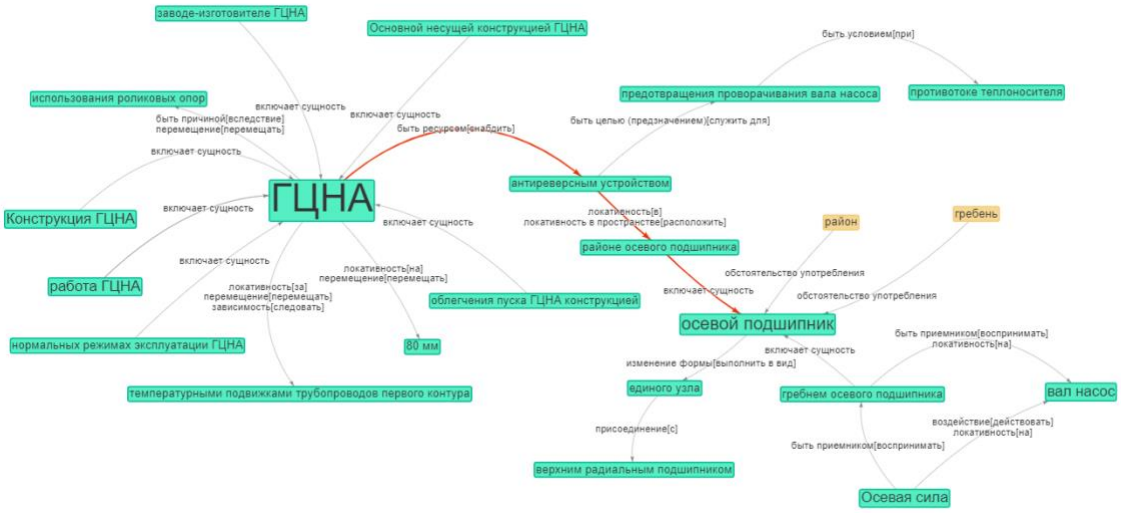

Рисунок 11 — Пример расширения окрестности термина «ГЦНА» через окрестность термина «антиреверсное устройство» в направлении термина «осевой подшипник»

Аналогичным способом могут быть изучены окрестности других терминов.

Таким образом, по построенному графу можно «восстановить» текст: «ГЦНА обладает таким ресурсом, как антиреверсное устройство, которое предназначено для предотвращения проворачивания вала насоса при условии противотока носителя и находится в районе осевого подшипника, который имеет вид единого узла с верхним радиальным подшипником. При этом гребень осевого подшипника воспринимает осевую силу, которая оказывает воздействие на вал насоса».

При помощи функции перехода от вершин к исходному тексту были получены фрагменты текста, использованные для построения графа: «…Осевая сила, действующая на вал насоса, воспринимается гребнем осевого подшипника. Осевой подшипник конструктивно выполнен в виде единого узла с верхним радиальным подшипником. … ГЦНА снабжен антиреверсным устройством, служащим для предотвращения проворачивания вала насоса при противотоке теплоносителя, расположенным в районе осевого подшипника».

Сравнение с исходным текстом показало, что смысл, в основном, не был искажен.

Заключение

В работе предложена технология построения и визуализации семантического образа полного текста документа, представляемого онтологией.

Типология информационных компонентов графа онтологии по степени контекстной определенности и завершенности смысловой конструкции обеспечивает соответствие даталогического и семантического уровней: элементарный факт соответствует отдельному действию (событию), ситуативный факт – отдельному утверждению, а завершенный факт – решению.

Разработана обобщенная схема решения задач информационного поиска, которая помимо этапов, соответствующих классическому информационному поиску, предполагает построение, визуализацию и анализ графа онтологии документа. Информационный поиск на графах онтологий сводится к схемам, включающим поиск

цепочки фактов и поиск окрестности элементарного (ситуативного) факта. Технология вариантной визуализация графа онтологии включает следующие этапы:

— отбор элементов графа в соответствии с задачей пользователя;
— формирование представления в соответствии с метафорой визуализации;
— формирование изображения в соответствии с моделью визуализации.

Исходя из типологии задач документального информационного поиска, определены две метафоры визуализации. Метафора «поиска пути» соответствует построению направленной цепочки фактов от исходных положений к целевым. Метафора «анализа окрестности» соответствует исследованию окружения (контекста) исходного факта. В целом это позволяет повысить эффективность восприятия за счет целенаправленного и управляемого сокращения размерности операционного пространства и профилирования.

Модель визуализации задает логику укладки элементов графа на плоскости. Реализована укладка вершин силовым методом Barnes-Hut; укладка вершин в порядке употребления (появления) имен сущностей в тексте; укладка вершин в соответствии со значимостью путей (длинной пути или суммарным весом вершин); укладка вершин в соответствии с некоторой схемой (например, функциональной моделью IDEF0).

Разработанное программное обеспечение позволяет строить граф онтологии по тексту на естественном языке, а также предоставляет возможности отбора, конфигурирования и манипулирования фрагментами онтологий в соответствии с метафорами поиска пути и анализа окрестности. Граф онтологии здесь выступает в роли технологического пространства «точек входа» в информационный массив, обеспечивая возможность непосредственного перехода от вершин графа к соответствующим фрагментам текста документа.

Разработанные инструменты интерактивного взаимодействия с графом онтологии реализуют принцип динамизма отображения (что обеспечивает возможность последовательного восприятия рассматриваемого объекта или процесса), а инструменты снижения размерности графа онтологии до уровня, приемлемого с точки зрения восприятия человеком – принцип минимизации временных затрат на анализ данных.

В то же время опытная эксплуатация и примеры применения показали, что качество визуализации в значительной степени зависит от качества построения семантического образа, в частности, от точности выделения и идентификации сущностей и отношений в тексте.

Благодарности

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

Список источников

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. Peirce Ch.S., Sowa J. Existential Graphs: MS 514 by Ch. S. Peirce with comment by J.F. Sowa., URL: HTTP://www.jfsowa.com/peirce/ms514.htm (дата обращения: 20.05.2020)
4. Михайлов А.М. Основы информатики. / Михайлов А.М., Черный А.И., Гиляревский Р.С. – М.: Наука, 1968. 756с.

5. Захарова А., Шкляр А. Основные принципы построения визуальных моделей данных на примере интерактивных систем трехмерной визуализации // Научная визуализация. – 2014. – Т. 6(2). – С. 62–73.
6. Гордеев Д. С. Обзор техник визуализации алгоритмов на графах // Научная визуализация. – 2018. – Т. 10(1). – С. 18–48.
7. Авербух В. Семиотический подход к формированию теории компьютерной визуализации // Научная визуализация. – 2013. – Т. 5(1). – С. 1–25.
8. 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.
9. 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.
10. 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 // НТИ. Сер. 2. – 2019 – №7. – С. 8–19.
11. Скороходько Э.Ф. Лингвистические проблемы обработки текстов в автоматизированных информационно-поисковых системах. // Вопросы информационной теории и практики. Сб.№25, – М.: ВИНИТИ. 1974.
12. Белоногов Г. Г. и др. Автоматический концептуальный анализ текстов // Сб. «Научно-техническая информация», сер. – 2002. – Т. 2. – С. 26–32.
13. Максимов Н.В., Гаврилкина А.С., Андронова 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.
14. Maksimov N. et al. Ontology of Properties and its Methods of Use: Properties and Unit extraction from texts //Procedia Computer Science. – 2020. – Т. 169. – P. 70–75.
15. Захарова А., Шкляр А. Метафоры визуализации // Научная визуализация. – 2013. – Т. 5(2). – С. 16–24.
16. Подвесовский А.Г., Исаев Р.А. Метафоры визуализации нечетких когнитивных карт // Научная визуализация. – 2018. – Т. 10(4). – С. 13–29.
17. Авербух В.Л., Бахтерев М.О., Манаков Д.В. Оценка метафор визуализации и видов отображения в контексте представления трасс выполнения и графов вызовов // Научная визуализация. – 2017. – Т. 9(5). – С. 1–18.
18. Касьянов В., Касьянова Е. Визуализация информации на основе графовых моделей // Научная визуализация. – 2014. – Т. 6(1). – С. 31–50.
19. Пупырев С.Н., Тихонов А.В. Визуализация динамических графов для анализа сложных сетей // Моделирование и анализ информационных систем. – 2010. – Т. 17(1). – С. 117–135.
20. Касьянов В.Н., Золотухин Т.А., Гордеев Д.С. Методы и алгоритмы визуализации графовых представлений функциональных программ // Программирование. – 2019. – № 4. – С. 19–27.
21. Библиотека визуализации сетей с открытым исходным кодом «vis-network.js» [Электронный ресурс] https://github.com/visjs/vis-network (дата обращения: 12.08.2020)
22. Пилюгин В.В., Мильман И. Визуальная аналитика и ее использование в деятельности лаборатории «Научная визуализация» НИЯУ МИФИ // Научная визуализация. – 2019. – Т. 11(5). – С. 46–55.
23. Sugiyama K. Graph Drawing and Applications for Software and Knowledge Engineers. Series on Software Engineering and Knowledge Engineering (2002).
24. Апанович З.В. Современные силовые алгоритмы для визуализации информации большого объема. В кн.: XIV Международная конференция «Проблемы управления

и моделирования в сложных системах», Самара, 2012: материалы. Самара: Самарский научный центр РАН, 2012. С. 164–171.
25. 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.
26. 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.
27. Максимов Н. В. Лексикографическая база данных для лингвистической поддержки задач документального информационного поиска: база данных / Максимов Н.В., Голицына О.Л., Тамеев А.А., Монанков К.В., Гаврилкина А.С., Абдулова Л.Л., Тазиева И.А., Андронова В.В., Кузьмина В.А. // Свидетельство о гос. регистрации №2019622150 от 22.11.2019.
28. Проект АЭС-2006, ОАО «СПбАЭП» [Электронный ресурс]: http://atomenergoprom.ru/u/file/npp_2006_rus.pdf (дата обращения: 20.05.2020).
29. Балтийская атомная электростанция [Электронный ресурс]: https://energybase.ru/power-plant/Baltic_NPP (дата обращения: 23.05.2020).
30. Проблемы повышения маневренности АЭС [Электронный ресурс]: https://tesiaes.ru/?p=9250 (дата обращения: 20.05.2020)
31. Максимов Н.В., Голицына О.Л., Усенко А.Л. Структура и компоненты операционного визуального пространства интерактивного поиска научной информации // Научная визуализация. – 2014. – Т. 6, № 4. – С 96–106.

Scientific Visualization, 2021, volume 13, number 1, pages 138 - 161, DOI: 10.26583/sv.13.1.10

Methods of visual graph-analytical presentation and retrieval of scientific and technical texts

N.V. Maksimov[1], O.L. Golitsina[2], K.V. Monankov[3], A.S. Gavrilkina[4]

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)

[1] ORCID: 0000-0002-8191-1521, nv-maks@yandex.ru
[2] ORCID: 0000-0002-3848-4755, olgolitsina@yandex.ru
[3] ORCID: 0000-0002-9267-3987, kmonankov@yandex.ru
[4] ORCID: 0000-0003-2167-1287, asgavrilkina@yandex.ru

Abstract

The technology of constructing and visualizing a semantic image of the full text of the document represented by the ontology as a system of three systems is offered: functional, conceptual and terminological. Objects and connections of the functional system correspond to the names of entities and relations extracted from the text; to objects of the conceptual system - descriptors of the thesauri of subject areas. The problem of the variable representation of entities at the sign level is solved using the rules for the formation of phrases of different lengths. 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 includes nodes and arcs of different types, which makes it possible to formalize operations on ontologies. Constructing a display of set elements of ontology into graph elements in such way that elements of different sets of different systems are distinguishable, recognizable and depicted in different ways, allows to implement the principle of correspondence of the graphic image with the semantics of the visualized data.

Based on the search tasks typology, metaphors for visualizing the ontology graph are proposed: the “pathfinding” metaphor, characterized by the construction of a directed chain of facts, and the “neighborhood analysis” metaphor, which is characterized by the study of the environment (context) of a fact.

The technology and software for the construction and variant visualization of the ontology graph have been developed.

Examples of using the proposed models for information retrieval through document texts are given.

Keywords: semantic search, text processing, graph representations of ontologies, visualization of ontology graphs, visualization metaphor.

References

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. Peirce Ch.S., Sowa J. Existential Graphs: MS 514 by Ch. S. Peirce with comment by J.F. Sowa. Retrieved May 20, 2020, from http://www.jfsowa.com/peirce/ms514.htm
4. Mikhailov, A. M., Chernyj, A. I., & Giljarevskij, R. S. (1968) Fundamentals of Informatics. [in Russian]

5. Zakharova, A., & Shklyar, A. (2014). Basic principles of data visual models construction, by the example of interactive systems for 3D visualization. Scientific Visualization, 6(2), 62-73. [in Russian]
6. Gordeev, D. S. (2018) A survey of visualization techniques of algorithms on graphs. Scientific Visualization, 10(1), 18-48. [in Russian]
7. Averbukh, V. (2013). Semiotic approach to forming the theory of computer visualization. Scientific Visualization, 5(1), 1-25.
8. 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]
9. 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.
10. 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.
11. Skorokhod'ko, E. F. (1974). Linguistic problems of text processing in automated information retrieval systems. Vopr. Inf. Teor. Prakt, (25), 5-120. [in Russian]
12. Belonogov, G. G., Bystrov, I. I., Novoselov, A. P., Kozachuk, M. V., Khoroshilov, A. A., & Khoroshilov, A. A. (2002). Automatic conceptual text analysis. Automatic Documentation and Mathematical Linguistics, 36(5), 57-65.
13. 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.
14. Maksimov, N., Gavrilkina, A., Kuzmina, V., & Borodina, E. (2020). Ontology of Properties and its Methods of Use: Properties and Unit extraction from texts. Procedia Computer Science, 169, 70-75.
15. Zakharova, A., & Shklyar, A. (2013). Visualization metaphors. Scientific Visualization, 5(2), 16-24. [in Russian]
16. Podvesovskii, A. G., & Isaev, R. A. (2018). Visualization metaphors for fuzzy cognitive maps. Scientific Visualization, 10(4), 13-29.
17. Averbukh, V. L., Bakhterev, M. O., & Manakov, D. V. (2017). Evaluations of visualization metaphors and views in the context of execution traces and call graphs. Scientific Visualization, 9(5), 1-18.
18. Kasyanov, V., Kasyanova, E. (2014). Information visualization on the base of graph models. Scientific Visualization, 6(1), 31-50. [in Russian]
19. Pupyrev, S. N., & Tikhonov, A. V. (2010). The analysis of complex networks with dynamic graph visualization. Modelirovanie i Analiz Informatsionnykh Sistem, 17(1), 117-135. [in Russian]
20. Kasyanov, V. N., Zolotukhin, T. A., & Gordeev, D. S. (2019). Visualization Methods and Algorithms for Graph Representation of Functional Programs. Programming and Computer Software, 45(4), 156-162.
21. Open Source Visualization Library to Display Networks “vis-network.js”. Retrieved August 12, 2020, from https://github.com/visjs/vis-network
22. Pilyugin, V. V., & Milman, I. (2019). Visual analytics and its use in the NRNU MEPhI “Scientific Visualization” laboratory activities. Scientific Visualization, 11(5), 46 – 55.
23. Sugiyama, K. (2002). Graph Drawing and Applications for Software and Knowledge Engineers. Series on Software Engineering and Knowledge Engineering.

24. Apanovich, Z. V. (2012). Modern Force-directed Algorithms for Visualization of Large Volumes of Information. In Problems of Management and Design in Complex Systems (pp. 164-171). [in Russian]
25. 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]
26. Maksimov, N. V., Golitsyna, O. L., Monankov, K. V., & Gavrilkina, A. S. (2020). Document Information-Analytical System xIRBIS. State Registration Certificate, (2020661683). [in Russian]
27. Maksimov, N.V., Golitsyna, O. L., Tameev, A.A., Monankov, K. V., Gavrilkina, A. S. et al. (2019). Lexicographic Database for Linguistic Support of Documentary Information Retrieval Tasks. State Registration Certificate, (2019622150). [in Russian]
28. Project AES-2006, JSC "SPbAEP". Retrieved May 20, 2020, from http://atomenergoprom.ru/u/file/npp_2006_rus.pdf [in Russian]
29. Baltic nuclear power plant. Retrieved May 23, 2020, from https://energybase.ru/power-plant/Baltic_NPP [in Russian]
30. Problems of increasing the maneuverability of nuclear power plants. Retrieved May 20, 2020, from https://tesiaes.ru/?p=9250 [in Russian]
31. Maksimov, N. V., Golitsyna, O. L., & Usenko, A. L. (2014). The structure and components of the operational visual space for scientific interactive information retrieval. Journal on Scietific Visualization, 6(4), 96-106.