ИНФОРМАЦИОННЫЙ ПОИСК

УДК 81’322.2’37

О.Л. Голицына, А.С. Гаврилкина

Об одном подходе к выделению имён сущностей и связей в задаче построения семантического поискового образа*

Представлены методы и средства выделения имён сущностей и связей на основе лексико-синтаксических шаблонов в рамках задачи семантического индексирования текстов документов. Содержание текста рассматривается как совокупность отражаемых триплетами элементарных фактов, включающих имена сущностей и отношений (имманентных, ситуативных и структурно-лингвистических). Для типизации ситуативных отношений используется таксономия отношений, в которой классы включают лингвистические конструкции; имманентные отношения формируются на основе сети понятий (тезауруса). Для идентификации свойств сущностей используется таксономия свойств и единиц измерения. Предложенный подход позволяет использовать в качестве поискового запроса имена сущностей, имена отношений, а также элементарные факты и составленные из них завершённые смысловые конструкции.

Ключевые слова: *семантический поиск, семантический поисковый образ, обработка текста, извлечение фактов, онтология*

DOI: 10.36535/0548-0027-2021-03-3

ВВЕДЕНИЕ

Семантический поиск в настоящее время ассоциируется с двумя классами задач.

Первый класс – отбор документов, отвечающих информационной потребности, выраженной запросом. Это хорошо известные механизмы: фактографический и тематический поиски, поиск аналогов, которые в совокупности нацелены на снятие неопределенности объекта/предмета поиска (лингвистической, семантической, прагматической).

Второй класс – комплексный аналитический (семантический) поиск, ориентированный на задачи синтеза нового знания (мониторинг проектов, выявление оснований и ограничений, анализ новизны, выявление потенциально возможных связей; оценка соответствия документации установленным критериям и нормативным требованиям и т.п.).

Отличие семантического поиска от традиционного (информационно-библиографического) состоит, в первую очередь, в представлении формы и содержания поискового запроса и результата поиска. Если поисковый образ запроса (и, соответственно, документа) при традиционном поиске обычно формулируется в виде последовательности ключевых слов, в общем случае не связанных общим контекстом употребления, то семантический поисковый образ, как запроса, так и документа, может включать не только ключевые слова, но и связи (отношения) между ними. При этом речь идет как об использовании имманентных отношений, наличие которых определяется посредством тезаурусов, так и ситуативных отношений, зависящих от лингвистической формы изложения конкретных фактов. Кроме того, решения задач, обеспечиваемые семантическим поиском (особенно задач второго класса), основываются на соотнесении и систематизации информации и характеризуются комплексностью результата и вариантностью его использования. Это позволяет определить семантический поиск как поиск не самих документов, а фрагментов, отвечающих в совокупности информационной потребности.

Качество и возможности поиска в значительной степени определяются индексированием массива документов, которое призвано обозначить смысл каждого документа. И если традиционно смысл представляется поисковым образом, включающим множество независимых ключевых слов, то при семантическом поиске поисковый образ – это совокупность связанных фактов, включающих имена сущностей и отношений.

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

В настоящей работе рассматривается методика построения семантического поискового образа путем преобразования текста документа в совокупность элементарных фактов-триплетов – пар имен сущностей, между которыми в отдельном предложении выделяется фрагмент текста, возможно представляющий семантическую связь. Такие связи в [1] называют «поверхностными», а подходы к их извлечению – «открытым извлечением информации» (Open Information Extraction).

Задача «открытого извлечения информации» была впервые сформулирована в [2] применительно к системе TextRunner, использующей для получения триплетов машинное обучение с частичным привлечением учителя. Подход TextRunner взят за основу при создании системы WOE [3], в которой удалось значительно повысить точность и полноту извлечения триплетов.

Для английского языка коллективом авторов был разработан ряд систем [4-9] (ReVerb, R2A2, ArgLearner, OLLIE, SRLIE, RelNoun, BONIE, OpenIE4, CALM), в которых для извлечения информации используются шаблоны из частей речи. В [4] представлен метод, при котором в качестве отношений рассматриваются глаголы и глагольные конструкции. При формировании отношений в [5] помимо глагольных связей привлекаются также связи между существительным и прилагательным (с учетом контекста). Кроме того, триплет уже сам может выступать в роли объекта и/или субъекта (аналогично реификации в RDF). В одну из методик, рассматриваемых в [8], включена возможность построения *n*-арных отношений. В [6] представлены ролевые связи, построение которых основано на разборе составных существительных. В [7] при формировании сущностей привлекаются правила извлечения числовых аргументов как единиц измерения. В [10] описаны многоязычные системы, такие как DepOE, ArgOE, LSOE и др.

Приведенные выше системы ориентированы преимущественно на обработку текстов на английском языке. Для разработки приложений, работающих с текстами на русском, создан ряд отечественных инструментальных систем: RCO Pattern Extractor [11], Alex [12], LSPL [13], DSTL [14], Томита-парсер [15]. Все эти инструментальные средства обработки текстов объединяет общий подход, основанный на распознавании и извлечении конкретных языковых конструкций. Для описания лингвистических свойств этих конструкций используется формальный язык, и само описание свойств осуществляется в форме специальных шаблонов и правил, с помощью которых происходит распознавание в тексте различных объектов и фактов. Совокупность шаблонов и правил формирует образец, настроенный на решение конкретной задачи, например, на выявление наименований юридических лиц, товаров, адресов; поиск информации о месте или дате рождения; выявление определений, понятий и т.п. Отметим, что принципиальным отличием этого подхода является то, что вначале создаются шаблоны, исходя из особенностей предметной области и решаемой задачи, после чего извлекаются факты. Однако, поскольку для информационного поиска характерно большое разнообразие форм представления смыслов и видов потребностей, описанный подход к построению семантического поискового образа не представляется полностью адекватным.

В настоящей работе предложен ориентированный на русский язык подход, использующий шаблоны для извлечения связей и формирования триплетов. Для унификации отношений применяется таксономия отношений, построенная на универсальных категориях.

НЕДОСТАТКИ КООРДИНАТНОГО ИНДЕКСИРОВАНИЯ

Для всех известных механизмов документального информационного поиска общим является сопоставление поискового образа документа и поискового образа запроса. При этом качество поиска определяется структурой этих образов и используемым критерием смыслового соответствия.

Основа информационного документального поиска – координатное индексирование – процесс, который заключается в формировании описания содержания документа в виде совокупности дескрипторов, выбираемых из заранее созданных словарей понятий либо из текста документа, и обозначающих основные понятия этого документа.

Метод координатного индексирования базируется на положении, что основное смысловое содержание документа и информационной потребности может быть с достаточной степенью точности и полноты выражено соответствующим списком ключевых слов[1], которые явно или в скрытом виде содержатся в тексте.

При так называемом «чистом» координатном индексировании ключевые слова в поисковых образах никак не связаны. В простейшем случае документ считается соответствующим информационному запросу (и подлежит выдаче), если в поисковом образе этого документа содержатся все ключевые слова поискового предписания. При этом необходимо понимать, что сами поисковые механизмы не имеют средств «угадывания» смысла (например, принадлежность определенной предметной или проблемной области) или интерпретации термина, используемого в качестве ключевого слова. Отсутствие возможности исключить влияние синонимии, полисемии и омонимии естественного языка, а также выразить ситуативные и имманентные связи между реальными объектами, процессами и т.п., представленными на вербальном уровне в тексте, существенно снижает семантическую силу информационно-поисковых языков, основанных на координатном индексировании.

К недостаткам «чистого» координатного индексирования [16], в основном, относят ложную или неполную координацию, когда в запросе недостаточно

[1] Под ключевыми словами в этом случае понимаются наиболее существенные для этой цели слова и словосочетания, обладающие назывной (номинативной) функцией. Назывные слова не обозначают предмет, а выделяют его путем указаний. К категории назывных слов относятся также имена собственные. Кроме назывных в качестве ключевых слов могут выступать и соответствующие численные характеристики, хронологические данные, диапазоны температур, давления и т. д.

использовать только координатную связь между ключевыми словами (например, результат запроса «Поставщики баз данных» содержит документы, представляющие базы данных поставщиков), а также отсутствие возможностей выражения в запросе парадигматических и синтагматических связей между ключевыми словами (например, запрос «Продажа лука», не доопределенный контекстом, приведет к формированию результата, содержащего и документы, представляющие лук как растение, и документы, в которых лук – это оружие).

Частично устранить отдельные недостатки могут, например, технологии расширения запроса (с использованием тезаурусов, лингво-процессоров, статистических связей и т.п.), однако они существенно зависят от особенностей понятийно-знаковой системы предметной области. Но, главное, их использование в автоматическом режиме на практике скорее ухудшает интегральные показатели эффективности поиска.

Для существенного повышения качества информационного поиска, основанного на применении координатного индексирования, необходимо разработать синтаксис информационно-поискового языка, который бы позволял использовать при построении поисковых образов документов и запросов не только простую координацию дескрипторов, но и существенные парадигматические и синтагматические связи.

СЕМАНТИЧЕСКИЙ ПОИСКОВЫЙ ОБРАЗ ДОКУМЕНТА

Поисковый образ линейной структуры, формируемый в процессе координатного индексирования, уже не отвечает задаче фиксирования не только ключевых слов, но и связей (отношений) между ними. В [17] предложен онтологический подход к построению поискового образа, который позволяет представить семантику документа системой понятий и отношений. Тогда при поиске в качестве запроса можно будет использовать завершенные смысловые конструкции.

Онтология определяется с позиций Общей теории систем как совокупность трёх взаимосвязанных систем: $O =< S_f, S_c, S_t \equiv>$ – функциональной (S_f), понятийной (S_c), терминологической (S_t) – и операции сопоставления элементов различных систем на уровне знаков (≡), обеспечивающей их тождество.

Функциональная система представляет объекты и отношения действительности средствами знакового уровня. Эти отношения имеют функциональную окраску, так как определяют способы и характер совместного существования и использования объектов. Логико-семантическим базисом онтологии является понятийная система, объектами которой служат устойчивые понятия предметной области, а набор отношений ограничен родовидовыми и ассоциативными (фиксируется в форме тезаурусов, рубрикаторов, классификационных схем и т.п.). Терминологическая система в онтологии отражает свойства естественного языка на уровне знаков – терминов, которые могут быть связаны отношениями эквивалентности (синонимии) и включения (образования словосочетаний). В качестве термина выступает отдельное слово или словосочетание естественного языка (или искусственного, например, шифр классификации), которое может применяться для описания понятия или объекта. Наличие понятийной и терминологической систем позволяет использовать для уточнения-расширения поискового запроса парадигматические отношения, формировать словосочетания, с разной степенью точности отражающие смысл.

В качестве моделей предлагаемых систем онтологии применяются помеченные ориентированные графы. При этом типологии вершин и дуг для графов понятийной и терминологической систем зафиксированы и однозначно заданы.

Семантической основой (смысловым «атомом») формирования графа функциональной системы является понятие элементарного факта – образа, фиксирующего некоторое состояние отдельного взаимодействия пары сущностей, где в роли сущности выступает понятие, объект, субъект и т.п., а взаимодействие представлено ситуативной связью (отношением). Элементарному факту в графе онтологии соответствует триплет «сущность – отношение – сущность», а множество вершин и множество дуг графа в совокупности соответствуют множеству элементарных фактов.

Таким образом, формирование онтологии, как поискового образа с сетевой организацией, требует:

1) задать понятийную систему онтологии с множеством парадигматических отношений;

2) представить текст документа в виде совокупности элементарных фактов.

Выражение имен сущностей и отношений на знаковом уровне позволяет индексировать элементарный факт как 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 позволит средствами традиционной теоретико-множественной модели информационного поиска, а также традиционного дескрипторного информационно-поискового языка реализовать отбор документов уже с учетом имманентных и ситуативных отношений между сущностями.

МЕТОДИКА ФОРМИРОВАНИЯ ЭЛЕМЕНТАРНОГО ФАКТА

Отображение смысла документа на множество элементарных фактов, формирующих узлы и дуги графа функциональной системы онтологии, основывается на классической схеме семантического анализа текстов, которая традиционно включает этапы графематического, морфологического, семантико-синтаксического и концептуального анализа [18].

На этапе графематического анализа выделяются структурные элементы текста (разделы, главы, абзацы, заголовки), текст разбивается на токены, которые идентифицируются и, в случае необходимости, объединяются с помощью словарей и лингвистических правил. Далее выявляются именные группы, даты, числа с плавающей точкой, аббревиатуры, единицы измерения и определяются границы предложений по знакам препинания с учётом идентифицированных специфических последовательностей символов. При

этом используются разделители элементов данных, разделители токенов, правила идентификации дат и аббревиатур, словари наименований и единиц измерения, разделители предложений.

Более точная идентификация токенов достигается за счёт учёта контекста их употребления, для чего используется таксономия свойств и единиц измерения [19]. Семантически значимые элементы текста, извлекаемые в границах одного или нескольких предложений, образующих семантическую окрестность токена, соотносятся с соответствующими компонентами онтологии, что позволяет восстанавливать недостающие смысловые фрагменты или выявлять противоречия, в частности, находить расхождения в обозначениях.

На этапе морфологического анализа происходит определение основных морфологических характеристик (часть речи, род, число, падеж) токенов, идентифицированных как слова, с использованием лингвистического процессора.

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. После согласования морфологических характеристик каждое предложение преобразуется во множество триплетов – элементарных фактов.

В основе методики формирования элементарных фактов лежит представление отдельного предложения в виде линейной последовательности токенов, разделенной на синтаксические отрезки, идентифицируемые в соответствии с типологией, где тип задается множеством частей речи, к которым могут относиться токены отрезка. На начальном уровне каждый отрезок типизируется как «имя субъекта/объекта», «связь (часть связи)» или «разделитель». На следующем уровне отрезок типа «имя субъекта/объекта» доопределяется подтипами: «группа имени сущности», «группа подлежащего», «имя собственное», «аббревиатура», «значение и единица измерения». Отрезок типа «связь (часть связи)» типизируется как «действие» или «обстоятельство» и далее тип «действие» – подтипами «действие-глагол», «действие-причастие», «действие-краткое причастие», а тип «обстоятельство» – подтипами «предлог» и «контекст».

Для кодирования отрезков применяется система кодирования с единичной длиной кода, алфавит которой содержит заглавные буквы латинского алфавита и символ «#», кодирующий разделитель. В результате кодирования предложения формируется символьная строка, в которой выполняется поиск на соответствие лексико-синтаксическим шаблонам. Пример разбиения и кодирования фрагмента текста представлен на рисунке.

«Для площадки Курской АЭС-2 выполнен анализ возможных сценариев аварийных ситуаций, приводящих к возникновению воздушной ударной волны (ВУВ) от источников взрывной опасности, находящихся внутри площадки. Определены безопасные расстояния от внутриплощадочных источников возникновения ВУВ с давлением во фронте ВУВ, не превышающим 30 кПа по ПиН АЭ-5.6».

Для	площадки Курской АЭС-2	выполнен
F	N	K

анализ возможных сценариев аварийных ситуаций	,
S	#

приводящих	к	возникновению воздушной ударной волны	(	ВУВ	)	от
W	F	N	#	X	#	F

источников взрывной опасности	,	находящихся	внутри	площадки	.
N	#	W	F	N	#

Определены	безопасные расстояния	от
K	S	F

внутриплощадочных источников возникновения ВУВ	с	давлением	во	фронте ВУВ	,
N	F	N	F	N	#

не превышающим	30 кПа	по	ПиН АЭ-5.6	.
W	M	F	X	#

S - группа подлежащего; N - группа имени сущности; X – аббревиатура;
M - величина и единица измерения; K - действие-краткое причастие; W - действие-причастие; F – предлог; # - разделитель.

Пример разбиения и кодирования фрагмента текста

Токены «30» (распознан как числовое значение) и «кПа» объединены в результате проверки токена «кПа» на принадлежность к единицам измерения в таксономии свойств и единиц измерения. Для токена «кПа» однозначно определено свойство с наименованием «Давление»[2].

Для описания шаблонов разработан язык, позволяющий для триплета вида

<субъект(S)><связь(L)><объект(O)>

указать последовательности отрезков предложения, которые должны определять каждый компонент триплета. Шаблон состоит из двух частей, разделенных символом «=»: в левой части приводится подстрока для поиска в строке, кодирующей отдельное предложение, а в правой – задается порядок формирования триплета (элементарного факта) в следующем формате:

S<десятичная цифра>**L**{<десятичная цифра>}[…n(7)]**O**<десятичная цифра>

Десятичная цифра указывает на позицию символа в левой части шаблона. В соответствии с форматом после символа **L** может быть указано несколько десятичных цифр (не более 7-ми), каждая из которых определяет позицию отдельной части связи (т.е. связь может быть составной и не обязательно представляется одним непрерывным отрезком). При формировании подстроки можно задавать наборы символов в квадратных скобках, которые позволяют указать, что на данном месте в исходной строке может стоять последовательность из перечисленных символов произвольной длины. Например, простейший шаблон SVN=S1L2N3 дает возможность выявить в предложении элементарный факт, представленный триплетом <подлежащее><сказуемое><дополнение>. В зависимости от вида и жанра обрабатываемых документов могут быть сформированы разные наборы шаблонов.

Пример формирования элементарных фактов на материале приведенного нами фрагмента текста представлен в табл. 1 (буква «U» обозначает имя сущности, уже использованное в каком-либо триплете).

Таким образом, результатом этапа семантико-синтаксического анализа является формализация линейного текста до уровня совокупности триплетов, формирующих узлы и дуги графа функциональной системы онтологии. При этом связи в триплетах отражают выявленные в тексте ситуативные отношения между сущностями.

Формирование имманентных и структурно-лингвистических отношений, а также типизация ситуативных отношений, – задача, которая решается уже на этапе концептуального анализа.

Имманентные отношения на уровне функциональной системы выделяются для обеспечения входа в понятийную систему. Проводится проверка имен сущностей элементарного факта на равенство или «подобие» терминам понятийной системы (например, тезауруса). «Подобие» в этом контексте означает наличие среди отдельных слов имени всех слов, составляющих термин понятийной системы, в произвольном порядке (например, имя «система информационного поиска» будет эквивалентно термину тезауруса «информационно-поисковая система») или наличие всех слов термина среди составляющих имен обеих сущностей триплета (например, триплет S«язык»-L«поддерживать»-O«информационный поиск» будет соответствовать термину тезауруса «информационно-поисковый язык»). В случае обнаружения равенства или подобия термин понятийной системы включается во множество имен сущностей и формирует элементарный факт (триплет) с отношением «термин понятийной системы». Термин в этом типе отношения выступает в роли объекта, а субъектом становится имя сущности, равное или подобное термину (в случае подобия на уровне элементарного факта – имя сущности-субъекта).

Таблица 1

Шаблоны и соответствующие элементарные факты

Триплет	Шаблон
<анализ возможный сценарий аварийный ситуация> <выполнить для><площадка курский АЭС-2>	FNKS=S4L31O2
<источник взрывной опасность><находиться внутри><площадка>	N[#]WFN=S1L23O4
<анализ возможный сценарий аварийный ситуация> <приводить к><возникновение воздушный ударный волна>	S[#]WFN=S1L23O4
<возникновение воздушный ударный волна><контекст><ВУВ>	N[#]X= S1L0O2
<безопасный расстояние><определить от> <внутриплощадочный источник возникновение ВУВ>	KSFN=S2L13O4
<фронт ВУВ><не превышать><30 кПа>	N[#]WN=S1L2O3
<внутриплощадочный источник возникновение ВУВ><с><давление>	U[#]FN= S1L2O3
<30 кПа><по><ПиН АЭ-5.6>	U[#]FX= S1L2O3

[2] В случае, когда единице измерения соответствуют несколько наименований свойств из разных областей применения, для разрешения многозначности на этапе концептуального анализа может быть проведён поиск в тексте по связанным с идентифицированными свойствами компонентам таксономии (в пределах обрабатываемого предложения), или при помощи тезауруса проанализирована терминология в предложении и тексте в целом и уточнена область применения. Например, в рассматриваемом нами предложении присутствует термин «давление», который совпадает с названием свойства «Давление» в таксономии свойств и единиц измерения.

Таблица 2

Триплеты имманентных и структурно-лингвистических отношений

Имя сущности	Отношение	Имя сущности
анализ возможный сценарий аварийный ситуация	«Содержит сущность»	аварийный ситуация
аварийный ситуация	«Обстоятельство употребления»	анализ
аварийный ситуация	«Обстоятельство употребления»	возможный сценарий
площадка курский АЭС-2	«Содержит сущность»	АЭС-2
площадка курский АЭС-2	«Содержит сущность»	курский АЭС-2
курский АЭС-2	«Обстоятельство употребления»	площадка
источник взрывной опасность	«Содержит сущность»	взрывной опасность
взрывной опасность	«Обстоятельство употребления»	источник
возникновение воздушный ударный волна	«Содержит сущность»	воздушный ударный волна
воздушный ударный волна	«Обстоятельство употребления»	возникновение
внутриплощадочный источник возникновение ВУВ	«Содержит сущность»	ВУВ
фронт ВУВ	«Содержит сущность»	ВУВ
30 кПа	«Параметр»	ДАВЛЕНИЕ
Площадка	«Нижестоящий»	площадка курский АЭС-2
ВУВ	«Нижестоящий»	внутриплощадочный источник возникновение ВУВ
ВУВ	«Нижестоящий»	фронт ВУВ
анализ возможный сценарий аварийный ситуация	«Тезаурус»	АВАРИЙНЫЕ СИТУАЦИИ
возникновение воздушный ударный волна	«Тезаурус»	УДАРНЫЕ ВОЛНЫ

Структурно-лингвистические отношения формируются на основе распознавания аббревиатур внутри имени сущности, членения длинных словосочетаний, выявления имен сущностей – величин и единиц измерения, определения лексикографического включения имен сущностей.

При обнаружении аббревиатуры внутри имени сущности создается новый триплет со структурно-лингвистическим отношением: S<имя сущности> L<«Включает сущность»>O<аббревиатура>. Деление длинного словосочетания на два и более происходит в соответствии с правилами формирования словосочетаний тогда, когда оно содержит имя более чем одной сущности, например, при следовании прилагательного за существительным. В этом случае формируются новые триплеты со структурно-лингвистическими отношениями «Включает сущность» и «Обстоятельство употребления» (имени). Такие операции приводят к увеличению весового показателя коротких имен сущностей в тексте и позволяют выделить объекты-действия.

Имена сущностей (или части имен), идентифицированные на этапе графематического анализа как величины и единицы измерения, связываются отношением с наименованием измеряемого свойства в соответствии с таксономией свойств и единиц измерения [19].

В формировании структурно-лингвистических отношений лексикографического включения участвуют только имена сущностей элементарных фактов, построенных на этапе семантико-синтаксического анализа. Отношения строятся по принципу «от самого короткого имени к самому длинному» («Нижестоящий»), например, *нагрузка → пожарная нагрузка → постоянная пожарная нагрузка*.

В табл. 2 приведены триплеты имманентных и структурно-лингвистических отношений, построенные для содержимого табл. 1 на этапе концептуального анализа. В частности, имена АВАРИЙНЫЕ СИТУАЦИИ и УДАРНЫЕ ВОЛНЫ принадлежат тезаурусу, выполняющему роль понятийной системы онтологии.

Задача приведения различных естественно-языковых конструкций к единой модели на структурном уровне решается путем построения единой последовательности-тройки по шаблонам, представляющим различные последовательности текстовых единиц в предложении. Например, фрагментам текста «определены безопасные расстояния от внутриплощадочных источников возникновения ВУВ», «безопасные расстояния определены от внутриплощадочных источников возникновения ВУВ», «от внутриплощадочных источников возникновения ВУВ определены безопасные расстояния» будет соответствовать один триплет <безопасный расстояние><определить от> <внутриплощадочный источник возникновение ВУВ>, включающий ситуативное отношение «определить от». Однако такая структурная формализация не выявляет ситуаций, когда одна и та же семантика может быть реализована разными с точки зрения языкового выражения отношениями.

Таблица 3

Пример применения таксономии отношений

Отношение	Класс	Модальность
выполнить для	быть целью (предназначением)	достоверное| состоявшееся
находиться внутри	локативность в пространстве	достоверное| выполняющееся
приводить к	быть результатом	достоверное| выполняющееся
определить от	быть ограничением	достоверное| состоявшееся
не превышать	изменение	невозможное| выполняющееся
с	присоединение	
по	быть основанием	

Лексических конструкций, представляющих отношения в тексте, довольно много, и их значение зачастую доопределяется или изменяется контекстом их употребления.

Минимальной основой для построения исчисления семантики является иерархия классов сущностей, отношений и свойств, базирующаяся на общей системе характеристических признаков. Хотя обзор существующих решений показывает, что нет строгой и всеобщей классификации, тем не менее, есть ряд вполне самодостаточных решений [20].

Для типизации ситуативных отношений в рассматриваемой нами методике используется онтология отношений, основанная на трехуровневой иерархической классификации[3]: первый уровень отражает соотношение реальность/модель; второй – комбинации соотношений отдельного (часть) и агрегатного (целое); третий уровень построен по признаку формы проявления отношения – действие-ориентированные, объект-ориентированные и результат-ориентированные. Листьями иерархического дерева являются классы отношений, обладающие комбинацией свойств верхних уровней. Каждый класс содержит множество конкретных лингвистических конструкций, которые отражают его семантику. Классы нижнего уровня открыты для пополнения, и их содержимое может зависеть от вида текстового массива, подлежащего обработке. Для удобства восприятия и эксплуатации на основе онтологии путем линейно-иерархического упорядочения построена таксономия отношений, пример применения которой для типизации отношений, отраженных в табл. 1, представлен в табл. 3.

Отношения типа «Действие» дополнительно характеризуются модальными свойствами. Текущее состояние таксономии отношений позволяет выявить свойство, определяющее возможность осуществления действия со значениями «Достоверное (Актуальное)»/ «Предполагаемое» (возможное)/«Невозможное» и свойство, характеризующее состояние завершённости действия со значениями «Выполняющееся» / «Состоявшееся» / «Ожидаемое»[4].

Для установления значения первого модального свойства лингвистические конструкции отношений проверяются на наличие сигнальных слов и/или их сочетаний. Например, отношение «не превышать» в табл. 3 включает сигнальное слово «не», по которому определено значение модального свойства «Невозможное». Если сигнальные слова не обнаружены, устанавливается значение «Достоверное». Для определения значения второго модального свойства анализируются морфологические характеристики (время и вид) глаголов и отглагольных частей речи, входящих в отношение.

ИСПОЛЬЗОВАНИЕ ЭЛЕМЕНТАРНЫХ ФАКТОВ В ЗАДАЧАХ СЕМАНТИЧЕСКОГО ПОИСКА

В результате семантического поиска из фрагментов найденных по запросу документов должна быть построена новая единица знания. В этом контексте элементарный факт может рассматриваться как некий маркер конкретного смысла, содержащегося в отдельном предложении текста, – сохраненная при индексировании связь триплета с предложением дает возможность прямого перехода к изложению факта.

Такое использование элементарного факта существенно снижает требования к его семантической согласованности и завершенности – в качестве «проводника» к смыслу может рассматриваться только отдельный компонент (или пара компонентов). Например, решение задачи поиска определений понятий в отдельном тексте или в информационном массиве возможно свести к поиску предложений, в которых при индексировании был выявлен триплет, содержащий отношение из класса «Определение (понятия)». Отбор элементарных фактов по признаку принадлежности определенному классу отношений не ориентирован на формирование самого определения на базе триплета и, тем самым, не требует обязательного наличия корректного полного представления имен определяемого понятия и определяющих его сущностей. В табл. 4 приведены примеры фактов-триплетов и соответствующие им фрагменты текстов, отобранных по запросу на поиск определений.

[3] Более подробно см. в [20]

[4] Таким образом, всем отношениям типа «Действие» дополнительно присваиваются два значения модальности, однако возможно определение двух значений первого модального свойства для одного отношения.

Таблица 4

Фрагменты текста, соответствующие триплетам с отношением из класса «Определение (понятия)»

Триплет с отношением из класса «Определение (понятия)»	Фрагмент текста
<энергосистема планета> <*быть понимать под*> <энергетический система>	Под энергетической системой будем понимать энергосистему планеты либо континента, либо группы стран, либо страны, либо области страны, либо района области, либо в виде одной энергоплощадки, в которой размещены энергоблоки, или даже один энергоблок.
<текущий концепция конечный захоронение> <*являться так называть*> <Pollux-концепция>	Текущей концепцией конечного захоронения облученных топливных сборок является так называемая Pollux-концепция.
<пространство><<*называться*> <гермообъем>	Пространство, ограничиваемое защитной оболочкой, называется гермообъемом.
<параметр i> <*назвать*> <предел Высикайло>	Параметр *i* - безразмерное число, назовем его пределом Высикайло и равно оно с большой точностью 0,9•10-18 для любых квазинейтральных КДС Космоса, состоящих из любых химических элементов или веществ.
<способ> <*пониматься под*> <альтернативный источник энергия>	Соответственно, под альтернативным источником энергии понимается способ, устройство или сооружение, позволяющее получать электрическую энергию (или другой требуемый вид энергии) и заменяющее собой традиционные источники энергии, функционирующие на углеводородах.

Таблица 5

Фрагменты текста, соответствующие триплетам с отношением из класса «Результат (исследования)»

Триплет с отношением из класса «Результат (исследования)»	Фрагмент текста
<массив экспериментальный данные> <*быть получить в результат*> <испытание материал>	Массив экспериментальных данных, на основании которого установлены наши зависимости, был получен в результате испытаний материалов, которые подвергались облучению, в основном при температуре ~270°С.
<зависимость потеря транспортный свойство линия> <*не быть обнаружить в*> <эксперимент>	В экспериментах не было обнаружено зависимости потери транспортных свойств линии от наличия масляной пленки на поверхности электрода.
<расчетный анализ циклический прочность> <*быть провести применить к*> <контейнер>	Поскольку конструкции контейнеров, размещаемых на выгородке и на корпусе реактора ВВЭР-1000, принципиально не различаются, расчетный анализ циклической прочности был проведен применительно к контейнерам, устанавливаемым на выгородке, поскольку они подвергаются более жестким условиям нагружения.
<расчет упругий термический напряжение> <*провести в*> < работа>	В работе проведен расчет упругих термических напряжений, возникающих при работе ядерного реактора в таблетках ядерного топлива цилиндрической формы.
<использование код KESS> <*показать для*> <анализ процесс>	Показано использование кодов KESS, FREQN, FRADEMO, IDEMO и других для анализа процессов при авариях.

Однако следует отметить, что полнота и точность поиска во многом определяются составом и наполнением классов отношений. Развитие онтологии отношений, начиная с нижнего уровня классификационного дерева, происходит путем деления класса, сформированного на основе комбинаций значений свойств верхних уровней, с дальнейшим наполнением полученных подклассов конкретными лингвистическими конструкциями. При этом, если количество классов нижнего уровня ограничено произведением количества значений классификационных признаков первых трех уровней, то дальнейшее деление может происходить без привязки к системе признаков верхнего уровня. Это означает использование в рамках отдельного класса собственных признаков деления, возможное несбалансированное развитие отдельных типологий и даже нарушение принципа иерархии (возможность построения класса как объединения подклассов двух и более родительских классов). Значения признаков деления формируют семантику класса и определяют его конкретное наполнение. Такой способ построения онтологии позволяет иметь несколько видов ее существования: как универсальной, с ориентацией на предметную область, так и с ориентацией на множество решаемых задач. Иными словами, каждая онтология может предполагать свое лингвистическое наполнение.

В табл. 5 приведены примеры поиска триплетов и соответствующих им фрагментов текста по принадлежности отношений классу «Результат (исследования)» в научных статьях по атомной энергетике.

Формирование структурно-лингвистических отношений, типы которых предопределены, направлено на упорядочение лексики семантического поискового образа. С одной стороны, при выявлении связей типа «Включает сущность» и «Обстоятельство употребления (имени)» происходит деление имен сущностей в триплетах с целью выявления внутренних взаимосвязей: из длинных словосочетаний вычленяются более общие понятия, что приводит к увеличению частоты таких понятий и, соответственно, к возрастанию значения меры их семантической значимости в отдельном тексте или в документальном массиве в целом. С другой стороны, построение над именами сущностей лексикографических деревьев позволяет проследить тематическое развитие отдельного понятия от общего к более частному употреблению и использовать эти точные имена сущностей при поиске. Например, ветви лексикографического дерева понятия «электрод» содержат словосочетания:

- «никелевый электрод», «висмутовый электрод», «жидкометаллический электрод» и т.п., формируя родовидовые связи;
- «взрыв электрода», «нагрев электрода», «поляризация электрода» и т.п., отражая связи типа «объект-процесс»;
- «неприемлемое увеличение эффективного сопротивления электродов», «изучение поведения материала электродов МИТЛ», «нагрев электрода магнитоизолированной транспортирующей линии протекающим током» и т.д., указывающие на конкретные ситуативные факты.

Таким образом, можно утверждать о формировании при индексировании мини-тезаурусов или ситуативных рубрикаторов, которые способны служить для пользователя своеобразными когнитивными проводниками в задачах извлечения знания.

ЗАКЛЮЧЕНИЕ

Предложенная в настоящей работе методика формирования семантического поискового образа рассматривает его как часть трехуровневой онтологии, включающей множество элементарных фактов, множество точек входа в понятийную систему (тезаурус) и деревья лексикографического включения. Представляющие элементарные факты триплеты, извлекаемые из текстов способами, подобными изложенному в настоящей статье, хотя и отражают так называемые поверхностные связи, тем не менее довольно полно идентифицируют содержащийся в тексте смысл.

Использование таксономии отношений как дополнительного лингвистического обеспечения позволяет для конкретной лингвистической конструкции определить тип отношения и построить унифицированную (с точностью до типов отношений, включенных в таксономию) теоретико-графовую модель текста, и тем самым обеспечить сопоставимость смыслов, выраженных разными лингвистическими конструкциями. Индексирование текста как совокупности триплетов позволит в рамках традиционной теоретико-множественной модели информационного поиска (и средствами традиционного дескрипторного информационно-поискового языка) реализовать отбор документов уже с учетом имманентных и ситуативных отношений между сущностями. Приведенные здесь примеры показывают конструктивность применения предложенной методики в задачах семантического индексирования и поиска.

СПИСОК ЛИТЕРАТУРЫ

1. Шелманов А.О. и др. Открытое извлечение информации из текстов. Часть I. Постановка задачи и обзор методов // Искусственный интеллект и принятие решений. – 2018. – № 2. – С. 47.
2. Banko M. et al. Open information extraction from the web // Proceedings of the 20th International Joint Conference on Artifical Intelligence. – San Francisco: Morgan Kaufmann Publishers Inc., 2007. – P. 2670–2676/
3. Wu F., Weld D.S. Open information extraction using Wikipedia // Proceedings of the 48th annual meeting of the association for computational linguistics. – Uppsala: Association for Computational Linguistics, 2010. – P. 118-127.
4. Fader A., Soderland S., Etzioni O. Identifying relations for open information extraction // Proceedings of the 2011 conference on empirical methods in natural language processing. – Edinburgh: Association for Computational Linguistics, 2011. – P. 1535-1545.
5. Schmitz M. et al. Open language learning for information extraction // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Lan-

guage Learning. – Jeju Island: Association for Computational Linguistics, 2012. – P. 523-534.
6. Pal H. et al. Demonyms and compound relational nouns in nominal open IE // Proceedings of the 5th workshop on automated knowledge base construction. – San Diego: Association for Computational Linguistics, 2016. – P. 35-39.
7. Saha S. et al. Bootstrapping for Numerical Open IE // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers). – Vancouver: Association for Computational Linguistics, 2017. – P. 317-323.
8. Mausam M. Open information extraction systems and downstream applications //Proceedings of the twenty-fifth international joint conference on artificial intelligence. – Palo Alto: AAAI Press, 2016. – P. 4074-4077.
9. Saha S. et al. Open information extraction from conjunctive sentences // Proceedings of the 27th International Conference on Computational Linguistics. – Santa Fe: Association for Computational Linguistics, 2018. – P. 2288-2299.
10. Glauber R., Claro D.B. A systematic mapping study on open information extraction // Expert Systems with Applications. – 2018. – Vol. 112. – P. 372-387.
11. Ермаков А.Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor: компонент выделения особых объектов в тексте // Сб. трудов XII Международной научной конференции «Информатизация и информационная безопасность правоохранительных органов».– М.: Акад. упр. МВД России, 2003 – С. 312-317.
12. Жигалов В.А. и др. Система Alex как средство для многоцелевой автоматизированной обработки текстов // Компьютерная лингвистика и интеллектуальные технологии. – М.: ФГУП "Изд-во "Наука", 2002. – С. 192-208.
13. Большакова Е.И., Ефремова Н.Э., Шариков Г.Ф. Инструментальные средства для разработки систем извлечения информации из русскоязычных текстов // Новые информационные технологии в автоматизированных системах. – 2015. – № 18. – С. 533-543.
14. Скатов Д.С., Ливерко С.В., Окатьев В.В. Язык описания правил в системе лексического анализа ЕЯ-текстов Dictascope Tokenizer // Компьютерная лингвистика и интеллектуальные технологии: по материалам Международной конференции. «Диалог» (Бекасово, 26-30 мая 2010 г.). – 2010. – Т. 9, № 16. – С. 442-449.
15. Томита-парсер. Руководство разработчика. – URL: https://yandex.ru/dev/tomita/doc/dg/concept/about.html (дата обращения: 28.12.2020).
16. Михайлов А.М. Черный А.И., Гиляревский Р.С. Основы информатики. – М.: Наука, 1968. – 756 с.
17. 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; Golitsyna O.L., Maksimov N.V., Okropishina O.V., Strogonov V.I. The ontological approach to the identification of information in tasks of document retrieval // Automatic Documentation and Mathematical Linguistics. – 2012. – Vol. 46, № 3. – P. 125-132.
18. Белоногов Г.Г., Быстров И.И., Новоселов А.П., Козачук М.В., Хорошилов Ал-др А., Хорошилов Ал-сей А. Автоматический концептуальный анализ текстов // Научно-техническая информация. Сер. 2. – 2002. – № 10. – С. 26-32. Belonogov G.G., Bystrov I.I., Novoselov A.P., Kozachuk M.V., Khoroshilov A.A., Khoroshilov A.A. Automatic conceptual text analysis // Automatic Documentation And Mathematical Linguistics. – 2002. – Vol. 36, № 5. – P. 57-65.
19. Maksimov N. et al. Ontology of Properties and its Methods of Use: Properties and Unit extraction from texts // Procedia Computer Science. – 2020. – Vol. 169. – P. 70-75.
20. 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; Maksimov N.V., Gavrilkina A.S., Andronova V.V., Tazieva I.A. Systematization and identification of semantic relations in ontologies for scientific and technical subject areas // Automatic Documentation And Mathematical Linguistics. – 2018. – Vol. 52, № 6. – P. 306-317.

Материал поступил в редакцию 30.12.20.

Сведения об авторах

ГОЛИЦЫНА Ольга Леонидовна – доцент, кандидат технических наук, доцент института Интеллектуальных кибернетических систем Национального исследовательского ядерного университета «МИФИ», Москва.
e-mail: olgolitsina@yandex.ru

ГАВРИЛКИНА Анастасия Сергеевна – аспирант Национального исследовательского ядерного университета «МИФИ», Москва.
e–mail: asgavrilkina@yandex.ru