

Возможности использования семантического поискового образа в задаче выявления определений понятий

О.Л. Голицына, А.С. Гаврилкина

Национальный исследовательский ядерный университет «МИФИ»

olgolitsina@yandex.ru, asgavrilkina@yandex.ru

Аннотация

Представлены результаты эксперимента по выявлению определений понятий в полных слабоструктурированных текстах научно-технических статей в рамках задачи семантического индексирования текстов документов. Содержание текста рассматривается как совокупность отражаемых триплетными элементарными фактами, включающими имена сущностей и отношений (ситуативных и имманентных). Методика построения триплетов основана на применении лексико-синтаксических шаблонов, позволяющих формировать триплеты для отдельных предложений текста. Для типизации ситуативных отношений используется таксономия отношений, в которой классы включают множества лингвистических конструкций. Использование таксономии отношений как дополнительного лингвистического обеспечения позволяет для конкретной лингвистической конструкции определить тип отношения, построить унифицированную (с точностью до типов отношений, включенных в таксономию) теоретико-графовую модель текста и тем самым обеспечить сопоставимость смыслов, выраженных разными лингвистическими конструкциями. Эксперимент проведен для класса таксономии отношений «Определение понятия», в котором рассмотрены три лингвистические конструкции – «называть», «понимать под» и «подразумевать под». Для каждой лингвистической конструкции рассчитаны полнота и точность выявления определений понятий на материале полных текстов 115 научно-технических статей по тематике «Атомная энергетика». Приведены примеры фрагментов текста, соответствующих триплетам с отношением из класса «Определение понятия».

Ключевые слова: обработка текста, извлечение фактов, онтология, таксономия отношений, определение понятий

Библиографическая ссылка: Голицына О.Л., Гаврилкина А.С. Возможности использования семантического поискового образа в задаче выявления определений понятий // Компьютерная лингвистика и вычислительные онтологии. Выпуск 5 (Труды XXIV Международной объединенной научной конференции «Интернет и современное общество», IMS-2021, Санкт-Петербург, 24 – 26 июня 2021 г. Сборник научных статей). — СПб.: Университет ИТМО, 2021. С. 29-38. DOI: 10.17586/2541-9781-2021-5-29-38

Введение

Семантический поиск можно определить, как поиск не самих документов, а фрагментов, отвечающих в совокупности сформулированной информационной потребности. Тем самым, отличие семантического поиска от традиционного (информационно-библиографического), в первую очередь, состоит в представлении формы и содержания поискового запроса и результата поиска. Для решения задач семантического поиска уже недостаточно формулировать поисковый образ запроса (и, соответственно, документа)

в виде последовательности ключевых слов, в общем случае не связанных общим контекстом употребления, как это принято при традиционном поиске – семантический поисковый образ как запроса, так и документа должен включать не только ключевые слова, но и связи (отношения) между ними. При этом речь идет не только об использовании имманентных отношений, наличие которых определяется посредством тезаурусов, но и ситуативных отношений, зависящих от лингвистической формы изложения конкретных фактов.

В [1] предложена методика 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 [2] называют «поверхностными», а подходы к их извлечению – «открытым извлечением информации» – Open Information Extraction). Основу формирования элементарных фактов-триплетов представляет ориентированный на русский язык подход, использующий лексико-синтаксические шаблоны для извлечения имен сущностей и связей.

При формировании ситуативных связей (отношений) существенное значение имеет не только факт их обнаружения, но и возможность различения и унификации. Для этих целей обычно используются таксономии, построенные на универсальных категориях. Наличие такой таксономии позволяет решать поисковые задачи через указание класса отношений по принципу принадлежности связи конкретного триплета к указанному классу.

В работе приводится пример одной из подобных задач – выявление наличия в научно-технических текстах определений понятий. Задача выявления определений понятий в текстах не является новой и известны ее решения, в том числе, и с применением лексико-синтаксических шаблонов (например, [3]). В основе предлагаемого решения – использование семантического образа слабоструктурированного текста.

1. Семантический поисковый образ слабоструктурированного текста

В [4] предложен подход к построению онтологического поискового образа, который позволяет представить семантику документа системой понятий и отношений (что дает возможность использовать при поиске в качестве запроса завершенные смысловые конструкции). Особенностью этого подхода является формирование онтологического поискового образа как трех взаимосвязанных систем – функциональной, понятийной и терминологической.

Функциональная система представляет объекты и отношения действительности, которые определяют способы и характер совместного существования и использования объектов средствами знакового уровня. Понятийная система является логико-семантическим базисом онтологии – ее объектами служат устойчивые понятия предметной области, а набор отношений ограничен родовидовыми и ассоциативными (фиксируется в форме тезаурусов, рубрикаторов, классификационных схем и т.п.). Свойства естественного языка на уровне знаков – терминов, которые могут быть связаны отношениями эквивалентности (синонимии) и включения (образования словосочетаний), отражает терминологическая система. 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 включать в поисковый образ парадигматические отношения, формировать словосочетания, с разной степенью точности, отражающие смысл.

Даталогическими моделями таких систем онтологии являются помеченные ориентированные графы, причем 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.

2. Методика формирования элементарного факта

Отображение смысла документа на множество элементарных фактов, формирующих узлы и дуги графа функциональной системы онтологии, происходит в соответствии с классической схемой семантического анализа текстов, которая традиционно включает этапы графематического, морфологического, семантико-синтаксического и концептуального анализа [5].

Методика формирования элементарных фактов довольно полно представлена в [1] и далее, для понимания существа, приводится ее краткое изложение.

В основе методики – формирование (на этапе семантико-синтаксического анализа) отдельного предложения текста в виде линейной последовательности токенов, разделенной на синтаксические отрезки. Отрезки идентифицируются в соответствии с типологией, где тип задается множеством частей речи, к которым могут относиться токены отрезка. На первом 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» (для глаголов и деепричастий), «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. На рисунке 1 представлен пример разбиения и кодирования следующего фрагмента текста:

«Под быстрой обратной связью понимается мощностная связь, действующая в течение 30-60 с после возмущения реактивности.»

Под F	быстрой обратной связью N	понимается V	мощностная связь S	, #
действующая W	в течение F	30-60 с N	после F	возмущения реактивности N
				. #

S – группа подлежащего; N – группа имени сущности; V – действие-глагол; M – величина и единица измерения; W – действие-причастие; F – предлог; # – разделитель.

Рис. 1. Пример разбиения и кодирования фрагмента текста

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 части 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 $SVN=S1L2O3$ 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 выше фрагмента текста представлен в таблице 1 (буква «U» обозначает имя сущности, уже использованное ранее в каком-либо триплете).

Таблица 1. Шаблоны и соответствующие элементарные факты

Триплет	Шаблон
<мощный связь><понимать под> <быстрый обратный связь>	FNVS=S4L31O2
<мощный связь><действовать в течение><30-60 с> <30-60 с><после><возмущение реактивность>	N[#]WFN=S1L23O4 U[#]FN= S1L2O3

Таким образом, происходит формализация линейного текста до уровня совокупности триплетов, формирующих узлы и дуги графа. При этом связи в триплетах отражают выявленные в тексте ситуативные отношения между сущностями.

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 «под быстрой обратной связью понимается мощная связь», «под быстрой обратной связью понимаем мощную связь», «мощная связь понимается под быстрой обратной связью» будет соответствовать один триплет <мощный связь><понимать под><быстрый обратный связь>, включающий ситуативное отношение «понимать под». Однако такая структурная формализация не выявляет ситуаций, когда одна и та же семантика может быть реализована разными с точки зрения языкового выражения отношениями. Лексических конструкций, представляющих отношения в тексте, довольно много, и их значение зачастую доопределяется или изменяется контекстом их употребления.

Для типизации ситуативных отношений (происходит на этапе концептуального анализа) в предложенной методике используется онтология отношений [6], основанная на трехуровневой иерархической классификации. 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.

Таблица 2. Пример применения таксономии отношений

Отношение	Класс
понимать под действовать в течение после	определение (понятия) воздействие, локативность во времени локативность

Каждый класс содержит множество конкретных лингвистических конструкций, которые отражают его семантику. Классы нижнего уровня открыты для пополнения, и содержимое их 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 таблице 1, представлен в таблице 2.

3. Использование элементарных фактов в задаче поиска определений понятий

В общем случае можно принимать, что в результате обработки графа из фрагментов 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).

Рассмотрим, например, задачу поиска определений понятий в отдельном тексте или в коллекции текстов. С использованием семантического поискового образа, построенного как совокупность элементарных фактов, решение задачи можно 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. Однако, результат такого отбора, представленный в виде последовательности предложений (или более длинных фрагментов текста), будет уже содержать (и связывать контекстно) все или почти все необходимые составляющие определения. Вместе с тем, следует отметить, что полнота и точность такого поиска во многом определяются наполнением классов отношений, то есть составом лингвистических конструкций, включенных в класс.

4. Эксперимент

В целях оценки качества представленной выше методики формирования элементарных фактов был проведен эксперимент по выявлению определений понятий в текстах научно-технических статей по тематике «Атомная энергетика» (всего было обработано 115 статей). В классе «Определение понятия» были рассмотрены три лингвистические конструкции – «называть», «понимать под» и «подразумевать под». Для каждой конструкции было подсчитано общее количество вхождений в тексты (с учетом различных значений морфологических признаков – род, число, падеж, время и т.п.) и экспертным путем определено, сколько из этих вхождений действительно указывают на контексты, содержащие определения понятий. При экспертном оценивании определениями не считались обозначения в формулах (например, «Где K будем называть радиационным коэффициентом токсичности i -го нуклида.») и фрагменты текста, не содержащие разъяснения понятия (например, «Указанные диапазоны включают в себя значения так называемой псевдокритической температуры.»)

В таблице 3 приведены примеры фактов-триплетов и соответствующие им фрагменты текстов, отобранных по запросу на поиск определений.

Для оценки эффективности работы методики использовались частные критерии – полнота (г) и точность (р).

Таблица 3. Фрагменты текста, соответствующие триплетам с отношением из класса «Определение понятия»

Триплет с отношением из класса «Определение понятия»	Фрагмент текста
<текущий концепция конечный захоронение> <являться так называемый> <Pollux-концепция>	Текущей концепцией конечного захоронения облученных топливных сборок является так называемая Pollux-концепция.
<пространство> <называть> <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 $0,9 \cdot 10^{-18}$ для любых квазинейтральных КДС Космоса, состоящих из любых химических элементов или веществ.
<гомогенный подход> <представлять так называемый> <логичным>	Более логичным представляется так называемый гомогенный подход к трансмутации МА в быстром реакторе. В нем МА, трансмутируя и превращаясь в плутоний, смешиваются с основным топливом и при 21 233 Ра.
<корпускулярный или электромагнитный излучение> <принять называемый> <жесткий>	Жестким принято называть такое корпускулярное или электромагнитное излучение, которое ионизует и возбуждает газ, но слабо взаимодействует непосредственно с электронами образовавшейся плазмы.
<электрон> <быть называемый> <ядерный электрон>	Электроны, входящие в состав ядра, будем называть ядерными электронами.
<агрегат> <называть> <ядерный энергетический реактор>	Ядерным энергетическим реактором называют агрегат, предназначенный для получения тепла из ядерного горючего путём самоподдерживающийся управляемой цепной реакции, деления атомов этого горючего.
<способ> <понимать 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.
<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.
<мощностный связь> <понимать под> <быстрый обратный связь>	Под быстрой обратной связью понимается мощностная связь, действующая в течение 30-60 с после возмущения реактивности.
<каталог> <понимать под> <база данные>	Под базой данных понимается каталог с набором файлов, содержащих всю информацию по измеряемым и расчётным параметрам, а также уставки, константы и другие данные.
<повреждение сосудистой стенка> <понимать под> <васкулит>	Под васкулитом понимали повреждение сосудистой стенки с ее инфильтрацией клеточными элементами.
<совокупность пакет прикладной программа> <подразумевать под> <компьютерный свод>	В нашем понимании под компьютерным сводом подразумевается совокупность пакетов прикладных программ, специализированных рабочих программ, служебных программ, а также библиотек констант и специальных программ для получения (генерации) этих библиотек из файлов оцененных ядерных данных.

Триплет с отношением из класса «Определение понятия»	Фрагмент текста
<часть орбита><подразумевать под><виток орбита>	Под витком орбиты подразумевается часть орбиты, соответствующая полному обороту спутника вокруг Земли.
<газовый цикл брайтон><подразумевать под><сверхкритический цикл брайтон>	Под сверхкритическим циклом Брайтона подразумевается такой газовый цикл Брайтона, в котором на стадии сжатия рабочее тело находится вблизи своей критической точки.

Полнота рассчитывалась как частное от деления количества выделенных триплетов с отношением из класса «Определение понятия», действительно являющихся маркерами определений ($N_{tripledef}$), на общее число определений, выявленных в текстах статей (N_{def}). Точность рассчитывалась как частное от деления количества выделенных триплетов, являющихся маркерами определений ($N_{tripledef}$), на общее число выделенных триплетов из класса «Определение понятия» (N_{triple}):

$$r = \frac{N_{tripledef}}{N_{def}}; p = \frac{N_{tripledef}}{N_{triple}}$$

Оценки были построены отдельно для каждой лингвистической конструкции. В табл. 4 приведены результаты расчета полноты и точности.

Таблица 4. Результаты эксперимента

Лингвистическая конструкция	Общее количество вхождений (N)	N_{def}	N_{triple}	$N_{tripledef}$	r	p
«называть»	84	44	75	39	0,89	0,52
«понимать под»	76	69	62	59	0,86	0,95
«подразумевать под»	26	17	24	13	0,76	0,54
Всего	186	130	161	111	0,85	0,69

Результаты эксперимента показали, что для класса «Определение понятия», представленного тремя лингвистическими конструкциями, полнота выявления определений составляет 85%. Точность определения оказалась ниже (около 70%) за счет невысокой точности для конструкций «называть» и «подразумевать под», что объясняется большой разницей между общим количеством вхождений конструкций в тексты и контекстов, содержащих определения понятий (например, для конструкции «называть» фиксировалось частое определение словосочетания «так называемый» без разъяснения понятия).

В ходе эксперимента были выявлены и типизированы случаи, когда по лингвистической конструкции не был построен триплет. Наиболее частыми оказались следующие:

- прилагательное в тексте определено как причастие, что не позволило образовать группу имени сущности (например, «Будем называть значимыми такие риски, ...»);
- обработка предложений с двусторонним как списков со структурно-лингвистическими отношениями (например, «Причем, под эффективностью следует понимать: оптимальность тактических характеристик противодействия основным угрозам, ...»);
- отсутствуют шаблоны, позволяющие, в том числе, учитывать сложносочиненные предложения (например, «Геополитику нередко связывают с геоэкономикой, под которой чаще всего понимается...»).

Тем не менее, представляющие элементарные факты триплеты, извлекаемые из текстов способами, подобными предложенному, хотя и отражают так называемые поверхностные связи, довольно полно идентифицируют содержащийся в тексте смысл.

Представленная в работе методика реализована в информационно-аналитической системе «xIRBIS» [7].

Заключение

Целью настоящей работы было показать возможность использования универсального поискового образа для решения частной задачи. Действительно, формирование поискового образа как множества элементарных фактов позволяет строить и традиционные индексы (по ключевым словам), и индексы, представляющие семантические связи. Наличие таких индексов (в данном случае индексируются типизированные отношения) позволяет в рамках традиционной теоретико-множественной модели информационного поиска (и средствами традиционного дескрипторного ИПЯ) реализовать отбор фрагментов документов с учетом ситуативных отношений между сущностями.

Авторы не ставили задачу сравнить полученные результаты с результатами работ, целенаправленно ориентированных на выявление определений понятий: задачей эксперимента было проверить на практике результаты типизации ситуативных отношений, с одной стороны, а с другой – продемонстрировать возможности семантического поиска в части формирования поискового результата как совокупности фрагментов документов, отвечающих поисковой потребности.

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

Литература

- [1] Голицына О.Л., Гаврилкина А.С. Об одном подходе к выделению имен сущностей и связей в задаче построения семантического поискового образа // Научно-техническая информация. Сер. 2. 2021. № 3. С. 17-26.
- [2] 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. 47-61.
- [3] Ефремова Н.Э., Большакова Е.И., Носков А.А., Антонов В.Ю. Терминологический анализ текста на основе лексико-синтаксических шаблонов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог. 2010. С. 124-130.
- [4] 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.
- [5] 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.
- [6] Максимов Н.В., Гаврилкина А.С., Андропова В.В., Тазиева И.А. Систематизация и идентификация семантических отношений в онтологиях научно-технических предметных областей // Научно-техническая информация. Сер. 2. 2018. № 11. С. 32–42.
- [7] 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.

Using of the Semantic Search Image in the Task of Identifying Definitions of Concepts

O.L. Golitsina, A.S. Gavrilkina

National Research Nuclear University MEPhI

The article presents the results of an experiment to identify definitions of concepts in full semi-structured texts of scientific and technical articles under the task of semantic indexing of document texts. The content of the text is considered as a set of elementary facts represented by triplets, including the names of entities and relationships (situational and immanent). The technique for constructing triplets is based on the use of lexical-syntactic templates, which allow to form triplets for individual sentences of the text. To typify situational relationships, a relationship taxonomy is used, the classes in which include a set of linguistic constructions. The use of the relationship taxonomy as an additional linguistic support allows for a specific linguistic construction to determine the type of relationship and build a unified (up to the types of relations of the taxonomy) graph-theoretic model of the text, and thereby ensure the comparability of meanings expressed by different linguistic constructions.

The experiment was carried out for the class of the relationship taxonomy «Definition of the concept», in which three linguistic constructions are considered - «называть» (to call), «понимать под» (to understand by) and «подразумевать под» (to mean by). For each linguistic construction, the recall and precision of identifying definitions of concepts are calculated based on the full texts of 115 scientific-technical articles on the topic "Nuclear Power". Examples of text fragments corresponding to triplets with a relationship from the «Definition of the concept» class are given.

Keywords: text processing, fact extraction, ontology, taxonomy of relationships, definition of concepts

Reference for citation: Golitsina O.L., Gavrilkina A.S. Using of the Semantic Search Image in the Task of Identifying Definitions of Concepts // Computer Linguistics and Computing Ontologies. Vol. 5 (Proceedings of the XXIV International Joint Scientific Conference «Internet and Modern Society», IMS-2021, St. Petersburg, June 24-26, 2021). - St. Petersburg: ITMO University, 2021. P. 29 – 38. DOI: 10.17586/2541-9781-2021-5-29-38

Reference

- [1] Golitsina O.L., Gavrilkina A.S. Ob odnom podhode k vydeleniyu imen sushchnostej i svyazej v zadache postroeniya semanticheskogo poiskovogo obraza [On one approach to the extraction of entity and relationships names in the task of building a semantic search image], Nauchno-tehnicheskaya informatsiya. Ser. 2 [Scientific and technical information. Ser. 2]. 2021. Vol. 3. P. 17-26. (In Russian).
- [2] Shelmanov A.O., Isakov V.A., Stankevich M.A., Smirnov I.V. Otkrytoe izvlechenie informatsii iz tekstov. CHast' I. Postanovka zadachi i obzor metodov [Open Information Extraction. Part I. The Task and the Review of the State of the Art], Iskusstvennyj intellekt i prinyatie reshenij [Artificial intelligence and decision making]. 2018. Vol. 2. P. 47-61. (In Russian).
- [3] Efremova N.E., Bol'shakova E.I., Noskov A.A., Antonov V. Iu. Terminologicheskij analiz teksta na osnove leksiko-sintaksicheskikh shablonov [Analysis of text terminology based on lexicosyntactic patterns], Komp'yuternaya lingvistika i intellektual'nye tekhnologii [Computational Linguistics and Intelligent Technologies]. 2010. P. 124-130.
- [4] 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.

- [5] 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.
- [6] 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.
- [7] Maksimov, N. V., Golitsyna, O. L., Monankov, K. V., & Gavrilkina, A. S. (2020). Document Information-Analytical System xIRBIS. State Registration Certificate, (2020661683). (In Russian).