Автоматизация формирования SPARQL-запросов для глобального семантического поиска

А. М. Велюго[1], А. Н. Губин[1], В. Л. Литвинов[2], Ф. В. Филиппов[1]
[1]*Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича*
[2]*Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)*
velugo.a@gmail.com, gan50_60@mail.ru, vlad.litvinov61@gmail.com, 9000096@mail.ru

***Аннотация*. Информационные системы должны обеспечивать высокую точность и достоверность предоставляемой информации. Для построения глобального информационного ресурса, обладающего свойствами общедоступности и достоверности, требуется определить эффективные концепции управления. Использование имеющихся онтологических словарей помогает повысить эффективность поисковых систем. Для оптимизации процесса обращения к словарю онтологий необходимо реализовать соответствующий сервис, способный к масштабированию и параллельному использованию. В работе предлагается способ обработки поисковой фразы с естественного языка на язык SPARQL-запроса. Описывается методика повышения эффективности семантического поиска.**

Ключевые слова: глобальные информационные системы; семантический поиск; онтологии; когнитивный поиск; SPARQL-запрос

I. Введение

Использование информационных систем не является эффективным в полной мере, если конечному пользователю требуется одновременно взаимодействовать сразу с несколькими ресурсами. Эффективно взаимодействовать в одном пространстве сразу с несколькими информационными системами позволяют технологии интеграции данных. Наиболее подходящим решением при проектировании онтологических информационно-справочных WEB-сервисов является семантическая сервисно-ориентированная архитектура SOA (Service Oriented Architecture) [1].

Фактически глобальные системы строятся в WEB-пространстве, и концепция управления глобальной информационной системой (ИС) как открытой системой означает переход к технологиям Linked Open Data (LOD) и предоставлению потребителю релевантной информации, полученной из достоверных источников [2]. Каждая система управления глобальной информационной системой, функционирующая в открытой среде, должна самостоятельно решать не только внутренние проблемы, но и всю совокупность проблем глобального информационного ресурса, связанных с внешней средой [3].

Таким образом, современные поисковые системы должны использовать онтологические технологии для максимального удовлетворения пользователя. Онтологии помогают искать информацию не по конкретному слову, а по предметной области и связанным с ней объектам. В качестве сервиса с онтологическим словарём можно использовать проект **wikidata**, который позволяет решить большинство базовых задач при поиске связей между объектами [4–6].

Информация в онтологии хранится в виде триплетов: субъект → предикат → объект. Наиболее частым запросом (в понятии триплетов) для поисковой системы является поиск субъекта по заданным предикатам и объектам. Открытое API **wikidata** позволяет находить субъекты с помощью предикатов и объектов.

В качестве входных данных может использоваться поисковая строка или оцифрованный контент документа. Фраза или текст передаются в сервис, который производит лемматизацию – приведение слов в начальную, словарную форму. Данный процесс необходим для дальнейшей передачи в **wikidata** с целью более точного определения элемента.

SPARQL-запрос использует коды элементов для взаимодействия с базой данных. Однако данные коды невозможно сопоставить с человекочитаемой информацией, поэтому необходим сервис, который поможет определить код по введённому слову.

Примером такого сервиса является **wbsearchentities**. Он возвращает код элемента по входящему названию и наоборот. При этом достаточно передать синоним названия, сервис вернёт подходящие элементы. Пользователю понадобится только подтвердить нужное значение, если метод вернул больше одного элемента. Сервис должен получать данные, обработанные после лемматизатора.

II. Процедура формирования запросов

Для запроса предиката обратимся к сервису:

https://www.wikidata.org/w/api.php?action=wbsearchentities&language=ru&format=json&type=property&search=Parent%20org.

Параметры приведены в табл. 1.

ТАБЛИЦА I. Параметры запроса предиката

Параметр	Значение	Комментарий
action	*wbsearchentities*	Указание, к какому методу общедоступного API wikidata необходимо обратиться
language	*ru*	Указание на каком языке передаётся запрос в search
format	*json*	Определяет формат ответа. Json выбран как самый универсальный для REST-сервиса
type	*property*	Указание типа поиска элемента. *Property* = предикат
search	*Parent org*	Значение слова/фразы для поиска. Передаётся нормальная форма слова (слов).

В ответ получим структуру, содержащую *id* предиката, который можно использовать в дальнейшем при построении запроса:

{"searchinfo":{"search":"Parent org"},"search":[{"id":"P749","title":"Property:P749","pageid":16095355,"display":{"label":{"value":"parent organization","language":"en"},"description":{"value":"parent organization of an organization, opposite of subsidiaries (P355)","language":"en"}},"repository":"wikidata","url":"//www.wikidata.org/wiki/Property:P749","datatype":"wikibase-item","concepturi":"http://www.wikidata.org/entity/P749","label":"parent organization","description":"parent organization of an organization, opposite of subsidiaries (P355)","match":{"type":"label","language":"en","text":"parent organization"}}],"success":1}.

Для запроса объекта обратимся к тому же сервису, но с другими параметрами (табл. 2):

https://www.wikidata.org/w/api.php?action=wbsearchentities&search=%D0%A1%D0%9F%D0%B1%D0%93%D0%A3%D0%A2&language=ru&format=json&type=item.

ТАБЛИЦА II. ПАРАМЕТРЫ ЗАПРОСА ОБЪЕКТА

Параметр	Значение	Комментарий
action	*wbsearchentities*	Указание, к какому методу общедоступного API wikidata необходимо обратиться
language	*ru*	Указание на каком языке передаётся запрос в search
format	*json*	Определяет формат ответа. Json выбран как самый универсальный для REST-сервиса
type	*item*	Указание типа поиска элемента. *item* = объект
search	*СПбГУТ*	Значение слова/фразы для поиска. Передаётся нормальная форма слова (слов).

В ответ придёт структура, содержащая *id* объекта, который можно использовать в дальнейшем при построении запроса:

{"searchinfo":{"search":"\u0421\u041f\u0431\u0413\u0423\u0422"},"search":[{"id":"Q4407705","title":"Q4407705","pageid":4209876,"display":{"label":{"value":"Saint Petersburg State University of Telecommunications","language":"en"},"description":{"value":"university in Saint Petersburg, Russia","language":"en"}},"repository":"wikidata","url":"//www.wikidata.org/wiki/Q4407705","concepturi":"http://www.wikidata.org/entity/Q4407705","label":"Saint Petersburg State University of Telecommunications","description":"university in Saint Petersburg,
Russia","match":{"type":"alias","language":"ru","text":"\u0421\u041f\u0431\u0413\u0423\u0422"},"aliases":["\u0421\u041f\u0431\u0413\u0423\u0422"]}}.

Имея идентификаторы *id* предикатов и объектов, поисковая система может сформировать SPARQL-запрос для поиска конкретного субъекта.

Например, пользовательский поисковый запрос «Какой университет находится в Санкт-Петербурге и принадлежит Министерству цифрового развития?» будет трансформирован в SPARQL-запрос:

```
SELECT ?item ?itemLabel
WHERE
{
  ?item wdt:P31 wd:Q3918;
        wdt:P131 wd:Q656;
        wdt:P749 wd:Q4294667.
  SERVICE wikibase:label {bd:serviceParam
wikibase:language "ru". }
},
```

где:
?item – переменная хранения субъекта;
?itemLabel – переменная хранения названия субъекта;
wdt:P31 – предикат «частный случай»;
wd:Q3918 – объект «университет»;
wdt:P131 – предикат «административно-территориальная единица»;
wd:Q656 – объект «Санкт-Петербург»;
wdt:P749 – предикат «материнская компания»;
wd:Q4294667 – объект «Министерство цифрового развития, связи и массовых коммуникаций Российской Федерации».

В результате выполнения запроса вернётся одна запись, показанная в табл. 3.

ТАБЛИЦА III. РЕЗУЛЬТАТ ВЫПОЛНЕНИЯ ЗАПРОСА

item	*itemLabel*
wd:Q4407705	Санкт-Петербургский государственный университет телекоммуникаций имени проф. М. А. Бонч-Бруевича

Полученные данные можно использовать в дальнейшем по назначению, например, используя *item*, запросить дополнительные связи объекта с другими объектами с целью улучшения релевантности поисковой выдачи.

Кроме операций сортировки и быстрого условного поиска имеются большие возможности по группировке и агрегированию данных RDF-хранилищ.

Например, замена объекта *wd:Q656* в запросе на переменную *?b*, позволяет получить информацию о всех университетах принадлежащих Министерству цифрового развития, с указанием кодов городов, в которых они находятся:

```
SELECT ?item ?itemLabel ?b
WHERE
{
  ?item wdt:P31 wd:Q3918;
        wdt:P131 ?b;
        wdt:P749 wd:Q4294667.
  SERVICE wikibase:label { bd:serviceParam
wikibase:language "ru". }
}.
```

В результате выполнения запроса получим следующую информацию (табл. 4).

ТАБЛИЦА IV. РЕЗУЛЬТАТ ПЕРВОГО ЗАПРОСА С ПАРАМЕТРОМ

item	*itemLabel*	*b*
wd:Q4304299	Московский технический университет связи и информатики	*wd:Q1382*
wd:Q4366452	Поволжский государственный университет телекоммуникаций и информатики	*wd:Q894*
wd:Q4407705	Санкт-Петербургский государственный университет телекоммуникаций имени проф. М. А. Бонч-Бруевича	*wd:Q656*
wd:Q4418283	Сибирский государственный университет телекоммуникаций и информатики	*wd:Q883*

Модификация запроса с заменой *wd:Q4294667* на переменную *?b* позволяет получить информацию о всех университетах расположенных в г. Санкт-Петербурге с указанием в переменной *b* кодов министерств, к которым относятся эти университеты (в табл. 5 показаны с сокращением).

```
SELECT ?item ?itemLabel
WHERE
{
  ?item wdt:P31 wd:Q3918;
        wdt:P131 wd:Q656;
        wdt:P749 ?b.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "ru". }
}
```

ТАБЛИЦА V. РЕЗУЛЬТАТ ВТОРОГО ЗАПРОСА С ПАРАМЕТРОМ

item	***itemLabel***	***b***
wd:Q27621	Санкт-Петербургский государственный университет	*wd:Q1140115*
wd:Q323681	Российский государственный педагогический университет им. А. И. Герцена	*wd:Q53738184*
wd:Q2652597	Художественное училище А.Л. Штиглица	*wd:Q53579434*
wd:Q2654435	Санкт-Петербургский государственный технологический институт	*wd:Q53579434*
wd:Q1628690	Санкт-Петербургский государственный аграрный университет	*wd:Q4294685*
wd:Q1628699	Санкт-Петербургский академический университет — научно-образовательный центр нанотехнологий РАН	*wd:Q53579434*
	……	
wd:Q4398064	Российский государственный гидрометеорологический университет	*wd:Q53579434*
wd:Q4407660	Санкт-Петербургская государственная академия ветеринарной медицины	*wd:Q4481679*
wd:Q4407665	Санкт-Петербургская государственная химико-фармацевтическая академия	*wd:Q2624248*
wd:Q4407667	Северо-Западный государственный медицинский университет имени И.И. Мечникова	*wd:Q2624248*
wd:Q4407693	Санкт-Петербургский государственный морской технический университет	*wd:Q53579434*
wd:Q4407698	Санкт-Петербургский государственный институт культуры	*wd:Q2465416*
wd:Q4407700	Санкт-Петербургский государственный университет гражданской авиации	*wd:Q4353455*
wd:Q4407701	Санкт-Петербургский государственный университет технологии и дизайна	*wd:Q53579434*
wd:Q4407705	Санкт-Петербургский государственный университет телекоммуникаций имени проф. М. А. Бонч-Бруевича	*wd:Q4294667*
wd:Q4407708	Санкт-Петербургский электротехнический университет	*wd:Q53579434*
wd:Q19908493	Государственный университет морского и речного флота имени адмирала С.О. Макарова	*wd:Q4481711*

Чтобы убрать из результата объекты по определенному параметру, необходимо использовать оператор MINUS. Добавим данный оператор, очистив из результатов университеты, принадлежащие Министерству науки и высшего образования Российской Федерации. Для удобства понимания информации выведем названия объектов, отсортируем по названию материнской организации (табл. 6).

```
SELECT ?itemLabel ?bLabel
WHERE
{
  ?item wdt:P31 wd:Q3918;
        wdt:P131 wd:Q656;
        wdt:P749 ?b.
  MINUS { ?item wdt:P749 wd:Q53579434. }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "ru". }
}
ORDER BY (?bLabel)
```

ТАБЛИЦА VI. РЕЗУЛЬТАТ ЗАПРОСА С СОРТИРОВКОЙ

itemLabel	***bLabel***
Северо-Западный государственный медицинский университет имени И.И. Мечникова	Министерство здравоохранения Российской Федерации
Санкт-Петербургская государственная химико-фармацевтическая академия	Министерство здравоохранения Российской Федерации
Санкт-Петербургский государственный университет кино и телевидения	Министерство культуры Российской Федерации
Санкт-Петербургский государственный институт культуры	Министерство культуры Российской Федерации
Российский государственный педагогический университет им. А. И. Герцена	Министерство просвещения Российской Федерации
Санкт-Петербургский государственный аграрный университет	Министерство сельского хозяйства Российской Федерации
Национальный государственный университет физической культуры, спорта и здоровья имени П. Ф. Лесгафта	Министерство спорта Российской Федерации
Санкт-Петербургский государственный университет телекоммуникаций имени проф. М. А. Бонч-Бруевича	Министерство цифрового развития, связи и массовых коммуникаций Российской Федерации
Санкт-Петербургский государственный университет	Правительство Российской Федерации
Санкт-Петербургская государственная академия ветеринарной медицины	Федеральная служба по ветеринарному и фитосанитарному надзору
Санкт-Петербургский государственный университет гражданской авиации	Федеральное агентство воздушного транспорта
Государственный университет морского и речного флота имени адмирала С. О. Макарова	Федеральное агентство морского и речного транспорта

III. ЗАКЛЮЧЕНИЕ

Предложенная методика автоматизации формирования SPARQL-запросов, основанная на использовании сервиса **wbsearchentities**, позволяет существенно сократить временные затраты и обеспечить высокую релевантность поисковой выдачи. Предоставляемые возможности по группировке и агрегированию данных с возможностью дополнительной очистки обеспечивают удобство и простоту доступа к информационно-справочным WEB-сервисам глобальных информационных систем и могут быть использованы при построении глобальных информационных ресурсов.

Список литературы

[1] Кашалкин Д.Ю., Курчидис В.А. Принципы построения семантической сервис-ориентированной архитектуры // Моделирование и анализ информационных систем, 2007. Том 14. №1. С. 48–53.

[2] P.-Y. Vandenbussche, B. Vatant. Linked Open Vocabularies // ERCIM News 96. 2014. P. 21–22. URL: https://ercim-news.ercim.eu/en96/special/linked-open-vocabularies (дата обращения 02.07.2023).

[3] Litvinov V.L., Filippov F.V. Paradigm of controls concept for global information systems // Proceedings of 2019 3rd International Conference on Control in Technical Systems, CTS 2019. 2019. P. 228-230.

[4] Губин А.Н., Литвинов В.Л., Литвинов Д.В., Филиппов Ф.В. Анализ методов проектирования пользовательских интерфейсов на базе онтологии предметной области // Актуальные проблемы инфотелекоммуникаций в науке и образовании. VII Международная научно-техническая и научно-методическая конференция; сб. науч. ст. в 4 т. / Под. ред. С.В. Бачевского; сост. А.Г. Владыко, Е.А. Аникевич. СПб.: СПбГУТ, 2018. Т. 2. С. 253-257.

[5] Губин А.Н., Литвинов В.Л., Турушева В.А., Филиппов Ф.В. Обеспечение заданного уровня доступа к данным в RDF-хранилищах. // Актуальные проблемы инфотелекоммуникаций в науке и образовании. VII Международная научно-техническая и научно-методическая конференция; сб. науч. ст. в 4 т. / Под. ред. С.В. Бачевского; сост. А.Г. Владыко, Е.А. Аникевич. СПб. : СПбГУТ, 2018. Т. 2. С. 183-187.

[6] Губин А.Н., Литвинов В.Л., Филиппов Ф.В. Теоретико-множественный подход к поиску информации в RDF-хранилищах. // Актуальные проблемы инфотелекоммуникаций в науке и образовании. VII Международная научно-техническая и научно-методическая конференция; сб. науч. ст. в 4 т. / Под. ред. С.В. Бачевского; сост. А.Г. Владыко, Е.А. Аникевич. СПб. : СПбГУТ, 2018. Т. 2. С. 262-266.