Парадигма концепции управления глобальными информационными системами

В. Л. Литвинов
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)
vlad.litvinov61@gmail.com

Ф. В. Филиппов
Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича
9000096@mail.ru

***Аннотация*. Информационные системы должны обеспечивать высокую точность и достоверность предоставляемой информации. Для построения глобального информационного ресурса, обладающего свойствами общедоступности и достоверности, а также широким диапазоном данных требуется определить эффективные концепции управления. Предметом исследования являются глобальные информационные системы, построенные на базе RDF-хранилищ, позволяющих объединять любые домены онтологий по принципу связанных открытых данных (Linked Open Data). Предлагается концептуальная схема согласования сущностей при разработке RDF-ресурсов и построении SPARQL-запросов. Описывается методика повышения скорости обработки запросов.**

Ключевые слова: глобальные информационные системы; онтологии; RDF; Linked Open Data; Linked Open Vocabularies; SPARQL

I. Введение

В отличие от ограниченного информационного ресурса, будь то корпоративная или общегосударственная информационная система (ИС), где система управления достаточно строго регламентирована, система управления глобальной ИС является сама по себе «открытой». Парадигма концепции управления глобальной ИС основывается на взаимодействии с внешней средой, она не может быть самообеспечиваемой и полностью зависит от способов контентного наполнения, поступающего извне.

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

Среди важнейших проблем управления и формирования глобальной ИС следует выделить следующие:

- проблема выбора сущностей и согласования их назначения при разработке RDF-ресурсов;
- проблема подбора адекватных сущностей при построении SPARQL-запросов;
- проблема больших временных затрат на реализацию запросов.

В [1] отмечены некоторые проблемы, обнаруженные при систематическом изучении существующих связанных открытых данных. В своем исследовании авторы обнаружили много типов ошибок, относящихся к неправильному использованию семантики терминов и в частности, около 15% триплетов, включающих необъявленные свойства URI. Доказательством сложности создания аннотаций является и тот факт, что, согласно [2], более половины изученных ресурсов (56,58%) путают свойства объекта со свойствами типа данных. Эти факты, а также практический опыт авторов в создании RDF-ресурсов, говорят о том, что существует острая необходимость в формализации шагов в процессе создания аннотаций, которые могут положительно повлиять на их качество.

Поскольку создание контента LOD длится уже более 15 лет, то накопилось огромное количество словарей опубликованных в Интернете. Более того, образовалось пространство Linked Open Vocabularies (LOV) [3], предоставляющее огромное число доступных словарей и терминов. Очевидно, будет естественным использовать LOV для решения проблемы выбора сущностей и согласования их назначения при разработке RDF-ресурсов и подбора адекватных сущностей при построении SPARQL-запросов. Этому исследованию посвящена первая часть настоящей работы.

II. Согласование сущностей RDF-ресурсов

Формирование онтологий и RDF-ресурсов осуществляется в определенной предметной области, для которой и строится соответствующий словарь (пространство имен) используемых сущностей. Одним из первых подобных словарей был Dublin Core, который широко используется для описания семантики

электронных текстовых документов в веб-среде. По сути, подобные словари включают идентификаторы для сущностей, определяющих наборы классов и свойств. Поскольку идентификация пространств имен с помощью URI уникальна в веб-среде, определенные в них имена при квалификации их идентификатором пространства имен также являются глобально уникальными. Благодаря этому в одной RDF-спецификации возможно использовать имена свойств, которые принадлежат различным пространствам имен и тем самым имеют различный смысл, не опасаясь коллизий между ними.

Рассмотрим проблему подбора адекватных сущностей при работе с RDF-ресурсами. На сегодняшний день в мире разработано огромное число словарей для различных предметных областей. Поэтому, в большинстве случаев, нет необходимости разрабатывать новый словарь. При построении ресурса в первую очередь необходимо выбрать словарь, наиболее полно отражающий его предметную область [4, 5]. Для этой цели можно использовать различные инструменты LOV, например, данные из https://lov.linkeddata.es/.

Этот ресурс позволяет найти подходящий словарь для идентификации сущности, будь то класс или свойство. При выборе словаря немалое значение имеет удобство его использования. Неоспоримое преимущество имеет проект wikidata, прежде всего обеспечением многоязычности и возможностью копировать, изменять, распространять и обрабатывать контент в любых целях, включая коммерческое использование.

Составим SPARQL-запрос, позволяющий получить названия всех стран, их столицы и население столиц (табл. 1, слева). Если не расшифровать коды сущностей, (*wdt:P31* – экземпляр класса, *wd: Q3624078* – суверенное государство, *wdt:P36* – столица, *wdt:P1082* – численность населения), то понять назначение запроса невозможно. С другой, стороны запрос с использованием словаря dbpedia (табл. I, справа) вполне понятен.

ТАБЛИЦА I SPARQL-ЗАПРОСЫ В WIKIDATA И DBPEDIA

https://query.wikidata.org/	**https://dbpedia.org/sparql**
SELECT ?a ?b ?c *WHERE {* *?a wdt:P31 wd:Q3624078 .* *?a wdt:P36 ?b .* *?b wdt:P1082 ?c. }*	*SELECT ?a ?b ?c* *WHERE {* *?a a dbo:Country.* *?a dbo:capital ?b.* *?b dbo:populationTotal ?c . }*

Пример одной строки результата запроса (табл. II, первая строка) показывает, что данные из wikidata также закодированы, в то время, как ответ из dbpedia очевиден (табл. II, вторая строка).

ТАБЛИЦА II ПРИМЕР РЕЗУЛЬТАТА SPARQL-ЗАПРОСА

a	***b***	***c***
wd:Q142	*wd:Q90*	2206488
http://dbpedia.org/ resource/Francia	*http://dbpedia.org/ resource/Paris*	2229621

Для получения раскодированных данных следует в запросе к wikidata заменить ?*a* на ?*aLabel*, ?*b* на ?*bLabel* и добавить строку:

SERVICE wikibase:label { bd:serviceParam wikibase:language "ru"},

которая заставит вывести метку, соответствующую коду. В данном случае результатом будет триплет на русском языке: *Франция Париж 2206488*. Несовпадение значений «*c*» реальному населению Парижа является результатом качества сопровождения баз знаний.

Несмотря на некоторые неудобства, связанные с кодированием сущностей, использование инструмента wikidata вполне допустимо.

Во-первых, при использовании поискового сервиса https://www.wikidata.org/w/index.php?sort=relevance&search значительно упрощается процесс нахождения адекватных кодов сущностей.

Во-вторых, при необходимости создания и публикации новых сущностей просто и удобно использовать сервис https://www.wikidata.org/wiki/Special:NewItem.

Очевидно, что вполне возможны и другие подходы. В частности, используя различные сервисы LOV [1, 3] можно выбрать минимальную совокупность словарей, покрывающую предметную область, создаваемого RDF-ресурса. При этом, целесообразно в первую очередь выбирать словари с наибольшим рейтингом.

III. СНИЖЕНИЕ ВРЕМЕННЫХ ЗАТРАТ НА РЕАЛИЗАЦИЮ SPARQL-ЗАПРОСОВ

Парадигма концепции управления LOD допускает прямой доступ к набору данных RDF-хранилища. Это является необходимой предпосылкой для решения проблемы больших временных затрат на реализацию запросов. В частности, использования теоретико-множественного подхода для реализации процедуры поиска триплетов, определяемых SPARQL-запросом [6].

На концептуальном уровне набор RDF-данных представляет набор троек

$$t = (s, p, o) \in (U \cup B) \times U \times (U \cup B \cup L),$$

где s – субъект, p – предикат, а o – объект. Счетно-бесконечные множества U, B и L являются обозначениями унифицированных идентификаторов ресурсов URI, пустых узлов и литералов соответственно. Ниже предлагается подход формализованного описания больших объемов триплетов, который позволяет существенно сократить время поиска информации за счет замены ряда процедур SPARQL на операции над множествами, доступными в универсальном языке программирования.

Хранилище данных естественным образом может быть отображено в трехмерном пространстве с координатами, представленными множествами всех значений субъектов, объектов и предикатов, используемых в триплетах. Подобное представление использовалось в [7, 8].

Каждый отдельный триплет будет представлять точку в этом пространстве. Положим, ось x соответствует субъектам из множества $s \in U \cup B$, ось y – объектам из множества $o \in U \cup B \cup L$ и ось z – предикатам из $p \in U$. Тогда для представления набора данных хранилища будем использовать три упорядоченных множества X, Y и Z, включающих соответственно все субъекты, объекты и предикаты, записанные в той последовательности, в которой они встречаются в триплетах.

Более наглядное представление этих данных предполагает рассмотрение плоскостей параллельных координатным плоскостям. Для примера из табл. 1 плоскость, параллельная координатной плоскости $X0Y$ с ординатой z = *"a"* (что эквивалентно *rdf:type*), будет включать точки, определяющие координаты пар субъект – объект, связанные отношением принадлежность к классу и, в частности, к классу *dbo:Country*. Реализация возможности такого компактного способа представления предполагает обеспечить независимую группировку и упорядочивание триплетов. Основной вопрос, с точки зрения сокращения времени выполнения запросов, состоит в скорости упорядочивания триплетов в соответствии с логикой запроса.

Известно много эффективных алгоритмов обработки больших массивов и, в частности, ассоциативных массивов. Наибольший выигрыш по скорости получается с использованием процедур бинарного поиска и сортировки с использованием ключей, аналогичных ключам ассоциативных массивов. Именно такие возможности предоставляет пакет data.table [9].

Множества триплетов в этом пакете представляются в виде таблицы, которая может быть сформирована автоматически из дампа хранилища. Назначение ключей осуществляется в соответствии со структурой и логикой запросов. Пусть, набор данных dbpedia в сериализации Turtle представлен с помощью множеств X, Y и Z, как описано выше. Тогда запрос на языке R, эквивалентный SPARQL-запросу (табл.1, справа) будет иметь вид:

```
dt <- data.table(x =X, y = Y, z = Z)
dt[z==(“a”),.(a = x & y = “dbo:Country”)]
dt[z==(“dbo:capital”),.(b = y & x = a)]
dt[z==(“dbo:populationTotal”),.(c = z & x = b)].
```

Для оценки эффективности использования пакета data.table в [6] описан эксперимент над хранилищами данных, включающих от одного до пятидесяти миллионов триплетов. Использование традиционных процедур, реализованных в стандарте SPARQL, затрачивает время на два порядка превышающее затраты на базе процедур пакета data.table языка R.

Кроме операций сортировки и быстрого условного поиска пакет предоставляет большие возможности по группировке и агрегированию данных RDF-хранилища. Это в полной мере позволяет использовать теоретико-множественный подход для значительного повышения эффективности использования хранилищ большого объема в системах поиска информации в глобальных ИС.

Список литературы

[1] I. Stavrakantonakis, A. Fensel, D. Fensel. Linked Open Vocabulary Recommendation based on Ranking and Linked Open Data. URL: https://www.researchgate.net/publication/299492189_Linked_Open_Vocabulary_Recommendation_Based_on_Ranking_and_Linked_Open_Data. – DOI: 10.1007/978-3-319-31676-5_3 (дата обращения 05.09.2019).

[2] R. Meusel, H. Paulheim. Heuristics for fixing common errors in deployed schema.org microdata. In The Semantic Web. Latest Advances and New Domains // ESWC 2015. P. 152–168. DOI:10.1007/978-3-319-18818-8_10.

[3] 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 05.09.2019).

[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.

[7] A. Matono, S.M. Pahlevi, and I. Kojima. RDFCube: A P2P-based Three-dimensional Index for Structural Joins on Distributed Triple Stores [Электронный ресурс]. URL: https://link.springer.com/chapter/10.1007/ 978-3-540-71661-7_31 (дата обращения 05.09.2019).

[8] M. Atre and J. A. Hendler. BitMat: A Main-memory Bit-Matrix of RDF Triples. In SSWS workshop at ISWC, 2009 [Электронный ресурс]. URL: http://www.cs.rpi.edu/~zaki/PaperDir/WWW10.pdf (дата обращения 05.09.2019).

[9] Package data.table [Электронный ресурс]. URL: https://cran.rproject.org/ web/packages/data.table/data.table.pdf (дата обращения 05.09.2019).