Московский государственный университет им. М.В.Ломоносова
Факультет Вычислительной математики и кибернетики
Кафедра Системного программирования

Введение в диплом

Формирование дескриптора взаимосвязанных данных для семантической электронной библиотеки

Научный руководитель:
зав. отд., д.ф.-м.н Серебряков В.А.

Студент: Пушин К.П.
Группа 528

Москва, 2013

Содержание

1 Семантическая паутина

Семантическая паутина это движение, возглавляемое Консорциумом Всемирной паутины (W3C) [1]. Цель движения — создать единый способ совместного использования информации приложениями, людьми или организациями, то есть значительно упростить процесс обработки информации. Кроме того, паутина позволяет автоматически устанавливать новые связи между данными.

Стоит заметить, что семантическая паутина это не замена текущей системе, чаще использующей язык разметки гипертекста (HTML), а дополнение к этой системе.

Для реализации цели были разработаны и стандартизованы специальные инструменты, языки и форматы, позволяющие работать с данными из заданной предметной области. Рассмотрим подробнее некоторые технологии.

Среда описания ресурса [2] (*RDF, Resource Description Framework*) — модель для представления данных, в особенности — метаданных. Среда предоставляет утверждения о ресурсах в пригодном для машинной обработки виде. Ресурсом может быть абсолютно любая сущность. Утверждение о ресурсе имеет вид «субъект – предикат – объект» и называется *триплетом* или RDF-тройкой. Для обозначения каждого элемента используются *уникальные идентификаторы ресурса (URI)*. Множество всех утверждений образует ориентированный граф, в котором вершинами являются субъекты и объекты, а ребра помечены предикатами. RDF имеет несколько различных вариантов представлений. Самый популярный на сегодняшний день — *RDF/XML*. Для хранения данных используются специальные RDF-хранилища, предоставляющие различный функционал для работы с данными, в том числе, иногда SPARQL-точку доступа для публикации данных в Сети.

Язык описания онтологий [3] (*OWL, Web Onthology Language*) — язык для описания веб-онтологий. Термин *онтология*, взятый из философии, это раздел, описывающий формы бытия. Веб-онтология это набор знаний об определенной предметной области. Она может включать в себя описание классов, свойств, экземпляров классов и операций над ними. Кроме того, онтологии могут описывать, как получить из данной онтологии некоторые логические следствия.

Язык запросов SPARQL [4] (*SPARQL Protocol and RDF Query Language*) — язык запросов к данным, представленным в модели RDF и протокол для передачи этих запросов и ответов к ним. Предоставление SPARQL-точек доступа является рекомендацией Консорциума при публикации данных во всемирной паутине.

2 Связанные открытые данные

Связанные открытые данные [5] (LOD, Linked Open Data) — метод публикации структурированных данных в стандартных форматах семантической паутины, при котором все данные связываются с ранее опубликованными.

Основные принципы LOD, сформулированные Тимом Бернерсом-Ли:

- Использование URI для идентификации сущностей;
- Использование HTTP URI для того, чтобы эти сущности могли использоваться человеком;
- При обращении по URI предоставлять информацию о сущности в одном из стандартизованных форматов, например, RDF или SPARQL;
- При публикации данных в сети также публиковать ссылки на схожие данные (используя URI).

На рис. 1 показана диаграмма источников данных, опубликованных в LOD к 2012 году. Сейчас можно сказать, что в развитии LOD наступила вторая фаза. В начале развития идеи (2006–2007 года) основной задачей была публикация данных в сеть. Второй этап подразумевает улучшение «экосистемы» LOD. Основной проблемой всех данных является их избыточность и несвязность, поэтому появляются различные приемы по автоматизации поиска уже опубликованных данных, относящихся к данной предметной области и привязывания к ним своих данных.

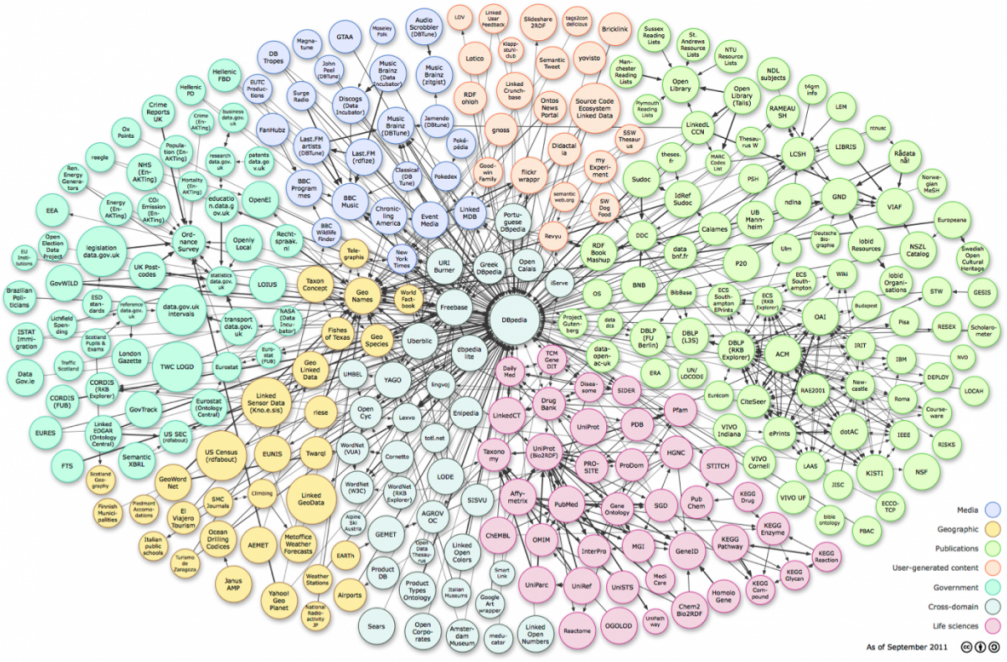

Рис. 1: Диаграмма LOD

3 Словарь взаимосвязанных данных

Словарь взаимосвязанных данных [6, 7] (VoID, Vocabulary of Interlinked Datasets) — словарь, описывающий метаданные об RDF-хранилище. Для описания словаря используется RDF Schema [8] (набор классов для описания онтологий). VoID был придуман как связующее звено между теми кто публикует данные в Сети, и теми, кто их будет использовать. Эти словари могут использоваться в различных ситуациях, от каталогизации и архивирования хранилищ до исследования данных, но чаще всего они используются для поиска необходимой информации.

VoID покрывает четыре аспекта метаданных: основные метаданные, метаданные доступа, структурные метаданные и связи с другими хранилищами.

Основные метаданные, соответсвуют стандарту Дублинского ядра (Dublin Core) [9]. Они содержат такую информацию, как название и краткое описание хранилища, лицензия, под которой распространяются данные, а также информацию о предметной области. Эти метаданные помогают пользователю понять, насколько информация, содержащаяся в хранилище, подходит для целей этого пользователя.

Метаданные доступа, описывающие как можно получить доступ к данным. Наборы данных в VoID определяются как наборы RDF-троек, но на самом деле, эти триплеты не описываются в словаре. Вместо этого с помощью метаданных доступа описываются методы доступа непосредственно к триплетам, используемым в описании VoID. Например, здесь можно указать SPARQL-точку доступа, которая предоставляет доступ к описываемому хранилищу.

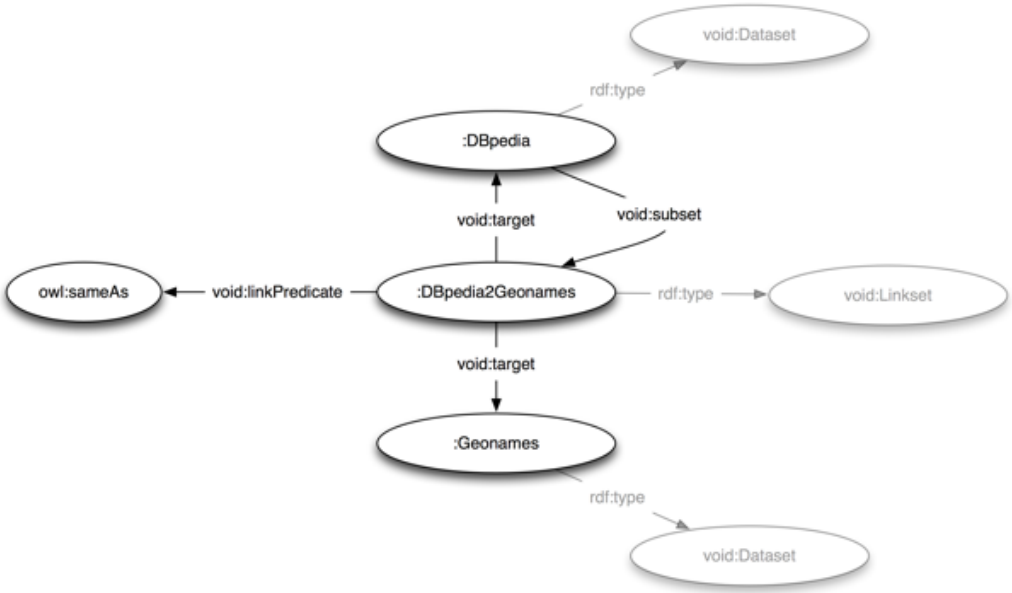

Рис. 2: Диаграмма типичного набора связей DBPedia2Geonames

Структурные метаданные, описывающие структуру и схему хранилища и используемые, например, для запросов к данным. Эти метаданные дают информацию о схеме и внутренней структуре хранилища и используются для исследования хранилища и выполнения запросов к нему. Например, в этом разделе указывается информация о словарях, используемых в хранилище, размере хранилища и примеры ресурсов типичных, для данного хранилища.

Наборы связей (Linksets), описывающие как различные хранилища соотносятся друг с другом и используются вместе.

На рис. 2 показан пример набора связей. Можно видеть, что DBPedia содержит подмножество связей *owl:sameAs*, соединяющих ресурсы DBPedia и Geonames.

4 Постановка задачи

В рамках дипломной работы необходимо разработать программу, которая будет генерировать VoID-дескрипторы для заданного RDF-хранилища.

За основу будет взята статья [10], в которой описывается алгоритм генерации VoID-дескрипторов для большого объема данных, полученных автоматически с помощью специальной программы (crawler), на примере Billion Triples Challenge (BTC) [11].

Алгоритм, предложенный в статье, должен быть модифицирован. В исследуемых в ходе работы хранилищах присутствует механизм автоматического связывания с другими хранилищами, основанными на Silk-framework [12]. Предполагается, что на основе правил связывания будут построены соответствующие связи в VoID-дескрипторах.

Список литературы

[1] T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web,” *Scientific American Magazine*, March 26, 2008.

[2] *RDF*. http://www.w3.org/TR/rdf-concepts/.

[3] *OWL*. http://www.w3.org/TR/owl-semantics/.

[4] *SPARQL*. http://www.w3.org/TR/rdf-sparql-query/.

[5] T. Heath and C. Bizer, “Linked data: Evolving the web into a global data space,” *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2011.

[6] *VoID*. http://www.w3.org/TR/void/.

[7] K. Alexander, R. Cygainiak, M. Hausenblas, and J. Zhao, “On the design and usage of void, the "vocabulary of interlinked datasets ” 2009.

[8] *RDF Schema*. http://www.w3.org/TR/rdf-schema/.

[9] *Dublin Core Metadata initiative*. http://dublincore.org/.

[10] C. Böhm, J. Lorey, D. Fenz, E. Kny, M. Pohl, and F. Naumann, “Creating voiD Descriptions for Web-scale Data,” 2010.

[11] *Billion Triples Challenge Dataset*. http://km.aifb.kit.edu/projects/btc-2012/.

[12] *Silk Framework*. http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/.