ПРИМЕНЕНИЕ RESOURCE DESCRIPTION FRAMEWORK ДЛЯ СЕМАНТИЧЕСКОГО ПОИСКА В УЗКОСПЕЦИАЛИЗИРОВАННЫХ БАЗАХ ЗНАНИЙ

Кайда А.Ю., Цой В.Г., Черний А.В.
Томский политехнический университет
ayk13@tpu.ru

Введение

В эпоху бурного развития информационных технологий всё больше уделяется внимания «интеллектуальным» технологиям. Не исключением является и RDF (Resource Description Framework) в частности семантический поиск с использованием RDF.

Описание алгоритма

RDF (Resource Description Framework) – это модель описания связанных данных, которая позволяет технологии Семантического веба интерпретировать информацию, представленную в вебе. Модель RDF основана на идее, которая заключается в следующем: всё, что существует в мире (будь то физический предмет или абстрактное понятие), имеет определенные свойства, а любое свойство имеет конкретные значения. Следовательно, любая сущность может быть описана с помощью элементарных выражений, которые называют эти свойства и их значения. Основу модели RDF представляет трехчастное утверждение, или триплет вида: Субъект – Предикат (свойство) – Объект (значение свойства). Например, утверждение «Томский Политехнический Университет основан в 1896 году» в RDF-терминологии можно представить следующим образом: субъект – «Томский Политехнический Университет», предикат – «основан», объект – «в 1896 году». Такое выражение принято представлять в виде графа, в котором субъект и объект – это узлы, а предикат изображается дугой или иной соединительной линей, направленной от субъекта к объекту [1].

Строение триплета представлено на рисунке 1.

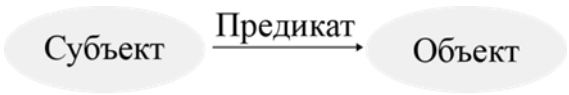

Рис. 1. Поиск фрагмента сцены в видеофайле

Иными словами, все, что хранится в RDF, вся информация, является набором триплетов. При этом информацию нужно интерпретировать только при ее формализации приложением в RDF и чтением агентом на другом конце. Однако стоит заметить, что создание предикатов предусматривает осторожность при работе с моделью – использование только общепризнанных словарей, а также установление связей между сущностями, которые представляют собой один и тот же объект.

В отношении одного и того же субъекта могут быть составлены и другие выражения, определяющие другие его свойства. Например, утверждение Технический вуз «Томский Политехнический Университет», основанный в 1896 году в городе Томске, включает в себя девять институтов можно представить в виде набора триплетов:

- вуз – имеет статус – Технический,
- вуз – имеет название – Томский Политехнический Университет,
- вуз – основан в году – в 1896 году,
- вуз – основан в городе – в городе Томске,
- вуз – включает в себя – девять институтов.

Кроме того, объекты (или значения свойств) могут являться субъектами других выражений, образуя значительно более сложные схемы. Множество RDF-утверждений образует ориентированный граф, в котором вершинами являются субъекты и объекты, а рёбра помечены предикатами [2].

Пример такого ориентированного графа представлен на рисунке 2.

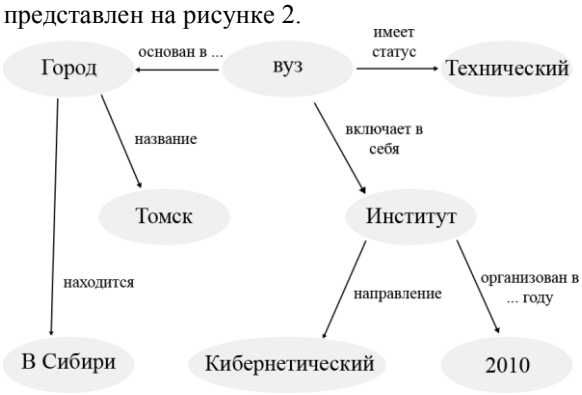

Рис. 2. Ориентированный граф, образованный множеством RDF-утверждений

Выражения в RDF должны быть составлены таким образом, чтобы они могли обрабатываться машинами. Для этого необходимы две вещи: система доступных машинной обработке уникальных идентификаторов для обозначения субъекта, предиката и объекта; доступный машинной обработке язык для представления выражений и обмена ими между машинами.

Стоит отметить, что данные могут быть независимыми от модели любого конкретного приложения, в котором они используются, т.е. набор фактов существует сам по себе. Можно

осуществлять операции добавления, удаления, делать запросы, интерпретировать, но они остаются логически независимыми.

Синтаксис RDF

Для публикации RDF-графов в вебе, их необходимо представить в последовательной, понятной для машин и пригодной для обмена данными форме. Для этого могут использоваться форматы, различающихся конкретным способом записи описания ресурса:

- RDF/XML – выражение графа RDF в виде документа XML;
- RDFa – запись внутри атрибутов произвольного HTML- или XHTML-документа;
- N3 (Notation 3) – краткий способ записи моделей RDF, компактнее и удобнее для чтения, чем XML-запись RDF.

Форматы RDF/XML и RDFa являются стандартом Консорциума W3C.

Семантический поиск

Семантический поиск – это метод информационного поиска, в котором релевантность документа запросу определяется семантически, а не синтаксически. Используя RDF-модель, прежде, чем рассмотреть задачи, с которыми легко справляется семантический поиск, рассмотрим самые сложные задачи. Существуют задачи, требующие продолжительного вычисления. Эти задачи не имеют ничего общего с пониманием семантики слова. На ранней стадии существования «Семантического Веба» считалось, что с его помощью возможно решить даже сверхсложные задачи, но это не так. Есть пределы того, что можно вычислить, и есть класс задач с огромным числом возможных решений, и на данный момент нет такой возможности решить эти задачи только представив информацию в RDF [3].

Но существую такие задачи, с которыми Semantic Web справляется великолепно. Они решаются при помощи тематической базы данных. Нужно учесть, что семантические технологии помогают нам отыскать тематическую информацию, рассредоточенную по всей сети – следовательно нет ничего удивительного в том, что семантические поисковые системы превзойдут тематические запросы.

Рассмотрим пример-сравнение семантического поиска и поиска через знакомые всем поисковые системы. Мы можем ввести запрос «Список студентов Томского Политехнического Университета» и получить что угодно, но только не то, что требовалось в запросе в порядке релевантности. Во-первых, коренным отличием являются ссылки. Стандарт Semantic Web использует URI – универсальный идентификатор ресурса, в то время как поисковики указывают на URL – местоположение ресурса. При этом URI может указывать и на URL. Поисковики ищут вхождения слов из запроса в тексте документов и возвращают документы, а не факты.

Но если система понимает, что нам нужны объекты «студент» субъекта «Томский Политехнический Университет», а формальные описания этих объектов будут доступны для индексирования в модели RDF (например, в форме записи RDFa на странице, чтобы поисковая машина могла их проиндексировать), будет получен набор искомых объектов.

С другой стороны, вполне обоснованно можно обратиться к привычному методу поиска, сделать несколько уточняющих запросов и найти всё, что нужно, что ставит под вопрос концепцию семантического поиска и RDF среди скептиков. Однако в ряде случаев предлагаемый вариант поиска, несомненно, гораздо эффективнее, удобнее и проще – избежать поиска среди множества документов с помощью построения сложных запросов [4].

Также у RDF-модели данных есть одно несомненное преимущество: при осуществлении семантического поиска, благодаря представлению данных в виде триплетов с набором некоторых свойств, можно получить выявленные «неявно» результаты поиска за счет образования новых связей между триплетами, незаданных вручную [5].

Если говорить о том, как RDF обеспечивает семантический поиск, RDF-модель обеспечивает формальные описания. Там, где присутствует формальное описание, поисковый агент может искать факты и знания.

Список использованных источников

1. Tim Berners-Lee et al. Tabulator: Exploring and analyzing linked data on the semantic web. In Proceedings of the 3rd International Semantic Web User Interaction Workshop, 2006. [Электронный ресурс]. – URL: http://swui.semanticweb.org/swui06/papers/Berners-Lee/Berners-Lee.pdf.

2. Powers S. Practical RDF – O'Reilly Media, 2003. – 352 p.

3. Semantic Web Standarts. RDF – Resource Description Framework. [Электронный ресурс]. – URL: https://www.w3.org/RDF/ (дата обращения 20.09.2016).

4. Hitzler P., Krötzch M., Rudolph S., Foundations of Semantic Web Technologies – FL.: Chapman & Hall/CRC, 2009. – 455 p.

5. SPARQL Query Language for RDF. [Электронный ресурс] – URL: https://www.w3.org/TR/rdf-sparql-query