Big Data Management

Coursework 2 Assignment:

Linked Data and Semantic Technologies

Ronan Smith, H00189534, rs6@hw.ac.uk

BSc Computer Science

Introduction

The world wide web (WWW), created by Tim Berners-Lee gives us access to astronomic amounts of data and may be considered as one of the most impressive inventions in human history. However, it has limitations, and with the growth of the concept of big data it is becoming more and more important that we find ways to solve these limitations. According to Hongyan Wu and Atsuko Yamaguchi in 2014 [1], more scientific data had been generated in the previous 5 years, than in the entire history of mankind. This is a startling fact. But if we are producing more and more data as time goes on, surely it makes sense that we want to be able to analyse and process it in more improved and more efficient ways, too? Linked Data and Semantic Technologies may offer a possible solution.

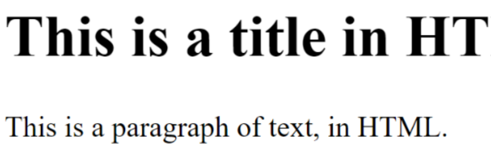
Objective of the Semantic Web

The current version of the World Wide Web (WWW) is useful in the way that it allows documents to be made available to anyone and connects documents from all around the world via hyperlinks. The current WWW provides a ‘web of documents’, documents that can be displayed to humans and machines, but only really *understood* by humans. A document is designed to be useful and easily interpretable for humans, and documents can be created to display all sorts of information about any topic you can imagine. The limitations begin when a computer tries to understand the information on a webpage/document. This is impossible using only HTML – the current standard for webpages. Below, a simple piece of HTML code is shown, and you can see how it may be interpreted completely differently by a human and a machine.

HTML:

```
<html>
      <h1>This is a title in HTML</h1>
      <p>This is a paragraph of text, in HTML.</p>
</html>
```

How a human user sees it (after rendering by a web browser):



This is a title in HTML

This is a paragraph of text, in HTML.

How a computer sees it:

```
<html>
      <h1>blah blah blah…</h1>
      <p>blah blah blah…</p>
</html>
```

Both the user and the machine see very different results from the same HTML document, with the user seeing a webpage that they can understand, and the machine seeing a document which contains no useful information which it cannot *understand* or *process.* The computer knows the tags (e.g. <p> is a paragraph) as they are built into HTML, but anything that is not predefined in HTML is complete gibberish to a computer (denoted above using '*blah blah blah*').

The Semantic Web (SW) is an extension or evolution of the WWW [2]. It's objective can be defined as follows:

> *"The objective of the semantic web is to provide a new form of content that is meaningful and processible by both humans and computers"* [3].

With the rise of the Internet of Things (IoT) which, in theory, looks to connect virtually every physical real-world device to the internet, it is becoming more and more important that the WWW is machine-friendly. As users of internet connected 'things' we expect everything to be automatic. We expect our smart devices to be able to look up the web to tell us the weather, plan our day, or tell us interesting facts quickly and easily, and this is why the web must be evolved from a 'web of (structured) documents' – unreadable documents for automated computer systems, to a 'web of (well-defined) data' – which can be understood, processed and linked by machines.

Key technologies

RDF (Resource Description Framework) is a key technology and standard model for data interchange [4] used in the semantic web. It is designed to allow for merging between different data sources, even if they have different underlying schemas, and it allows for meaning to be expressed using triples where each RDF item has at least a subject, predicate and object. It can be modelled using XML to provide an extension of HTML, allowing for meaning to be defined inside tags as well as user-defined tags. Usually, the subject is the current object that we want to talk about, the predicate (or list of predicates) is attributes that apply to that subject and the object part is some other resource, being pointed to using a URI (Uniform Resource Identifier). A resource can be a document, a literal, or another object. And they work by extending the URIs currently used on the World Wide Web to not just link directly to webpages, but also individual objects on those webpages.

A good data interchange model needs a query language, and for RDF the query language is SPARQL (SPARQL Protocol and Standard Query Language). It has similarities to SQL, usually taking the form shown below:

```
PREFIX …     SELECT …     FROM …     WHERE …
```

The final 'semantic technology' to be mentioned here is ontologies. One way to describe an ontology is as an "explicit specification of a conceptualization" [5]. In other words, it is a list of set definitions of concepts and the relationships between them, that can be referred to across different data sources to force a consistent structure. This is seen as the technology with the "highest degree of semantic richness" [6], in other words it is the one best suited for use on the Semantic Web.

These technologies are all related in the way that they use Linked Data (discussed in the next section) to help create a ‘web of data’ for the semantic web, in a consistent structure that can be interacted with by both human users and automated computer agents.

Linked Data vs. Semantic Web

‘Linked Data (LD)’ and ‘Semantic Web (SW)’ are two very closely related terms. Think of the SW as an umbrella term, which covers a range of techniques for making the world wide web (WWW) machine-readable, or a framework that defines and promotes standard data formats for sharing and reusing across sources on the WWW. Linked Data is member of the SW and a tool used to create *relationships* between data. These relationships can be across any collection of data sources that have been created using SW standards. And it is only when Linked Data is in use that the Semantic Web can really become interesting.

According to Tim Berners-Lee, the creator of the WWW and the man behind the ideas of the SW, “Linked Data is the semantic web done right” [7, 8]. Linked Data is where the term ‘web of data’ really comes in, and the overall idea of LD is to have all objects on all areas of the WWW or SW interlinked and related, where they can be easily queried upon by machines, but also still understandable by humans. The main principles of Linked Data are as follows [9]:

1. Use URIs (Uniform Resource Identifiers) to name and identify things.
2. Use HTTP URIs so that these things can be looked up (just as you would look up a webpage on the current WWW).
3. Provide useful information about what a name identifies when it is looked up, using open standards (for example RDF and SPARQL).
4. Refer to other things using their HTTP URI-based names when publishing data on the Web.

Linked Data makes the Semantic Web much more **connected**, and thus much more fit for purpose.

Areas of use

The semantic technologies described above can be employed in a huge (and growing) number of areas. For example, in life sciences they can be used for genetic sequencing – in other words, cataloguing human genetic variations and the relationships between these differences. They can also be used to combine different sources of research into one, so people can combine results for the greater good. For example, in research on Alzheimer's disease, data can come from all sorts of disciplines including psychiatry, neurology, microscopic anatomy, and more. It makes sense to have systems in place to combine these sources so everyone concerned has access to them.

OpenBiodiv-O is an example ontology that aims to make "biodiversity data and information accessible worldwide" [7] – another example of combining usually separate sources to help the common cause.

And finally, the Semantic Web is already being employed through search engines. Often, when searching through Google for a movie title, for example, you will be provided with an overview box which tells you the name of the movie as well as other information such as the director and cast. These are often being combined and being summarised from completely seperate sources. A member of the cast can be selected to link you to a completely new resource describing that actor or actress. This is an example of Linked Data.

Ultimately, having these systems in place allows the human user to do less work to find the information we want, and the more automated the Web becomes, the less work we will have to do to find new information.

There is a long way to go before the World Wide Web can be completely transitioned into the Semantic Web, but the foundations are there, and technologies such as RDF, SPARQL and Ontologies are the building blocks we need to make this transition.

Word Count (not including title page or references): 1494 words

References

[1]. H. Wu & A. Yamaguchi, 2014. “Semantic Web technologies for the big data in life sciences” in Bioscience Trends, 8(4), pp.192–201.

[2]. T. Berners-Lee, J. Hendler, O. Lassila “Scientific American: Feature Article: The Semantic Web”, May 2001. Available: https://www.scientificamerican.com/magazine/sa/2001/05-01/#article-the-semantic-web.

[3]. I. Szilagyi and P. Wira, "Ontologies and Semantic Web for the Internet of Things - a survey," *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, Florence, 2016, pp. 6949-6954.
doi: 10.1109/IECON.2016.7793744.

[4]. World Wide Web Consortium. (2018, March, 22) *Resource Description Framework* [online]. Available: https://www.w3.org/RDF/.

[5]. T. Gruber “Toward Principles for the Design of Ontologies Used for Knowledge Sharing.” Int. J. Hum. Comp. Stud. 1995, 43, 907–928.

[6]. Bastian Eine, Matthias Jurisch & Werner Quint, 2017. Ontology-Based Big Data Management. Systems, 5(3), p.45.

[7]. V. Senderov et al. “OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system”, *Journal of Biomedical Semantics*, 2018.

[8]. P. Ristoski, H. Paulheim (2016). “Semantic Web in Data Mining and Knowledge Discovery”. *Data and Web Science Group, University of Mannheim, B6, 26, 68159 Mannheim.*

[9]. *T. Berners-Lee (2006-07-27). "Linked Data". Design Issues. W3C.*