SEMI STRUCTURED DATA INFORMATION RETRIEVAL USING ONTOLOGIES

Vladimír Smataník, Karol Matiaško

Institute of Information and Communication Technologies
University of Žilina, Žilina, Slovakia

E-mail: vladimir.smatanik@uniza.sk

Abstract: Vast amounts of data are being stored on many servers, in databases, files or different structures. The capability of storing everything, however, does not provide good results from queries and information retrieval [1]. Data organized differently require specific approaches to allow more sophisticated information retrieval. One of the interesting ways of storing the data is ontology. This article suggests an ontology structure for storing some of semi-structured data types.

Key words: ontology, RSS, semi structured data, text mining

Introduction

Most of the time usually spent on the internet is searching for information. Depending on the task, we need to find high quality sources, images or web pages between millions of results. Any general term put into a search engine yields thousands of pages, images and sorted based on the relevance to the given query.

Sorting is usually made with the aid of statistics and the behaviour of previous users, who searched the same or similar term. However, there is usually a load of misleading sources which need to be skipped in order to access the desired information.

The transition from the idea and image in our minds into a query understood by the search engine is still a big deal, especially with growing volumes of all types of data. Any meta information is merely enough to make the results more accurate [2].

This article deals with a specific part of these data stored on the internet and describes possible approaches towards retrieving information from such sources. It is structured as follows. In the Section 1, definition of structured, unstructured and semi-structured data is expressed, as well as the issue of data and text mining and information retrieval. Section 2 introduces the open world assumption and ontology. Section 3 suggests the transformation of one specific type of semi-structured data – RSS feeds. Section 4 offers the structure of ontology for storing the RSS feeds. The conclusion sums up the advantages and disadvantages of the given approach and offers future challenges in this area of research.

Semi-Structured Data and Information Retrieval

In this article, facts and statistics gathered for reference and analysis will be considered as data. Digital data are quantities, characters and symbols with which computers performs operations. They are stored on magnetic, optic or mechanic storages and transferred using electrical signals [3].

Data can be divided into 2 basic categories, based on their organisation and pre-defined model of behaviour:

1. **Structured data** are stored in a field inside of a record or file. Structured data are usually stored in a database. Although the structure is known, accessing the desired information requires the right querying [4].

2. **Unstructured data** are missing known organisation or structure and mostly contain unnecessary data along with crucial information. Their size requires automatic processing, which is however difficult to achieve due to their structure [5].

In some sources, there is one more category of data, called semi-structured data. This category is sometimes put inside the structured category, because of its known structure, in other cases into the unstructured part because of the difficulties of extracting valuable information from such sources. Thus, retrieving information from this category of data requires methods from dealing with both, structured and unstructured data [6].

Semi-structured data does not fit with the formal structure of relational database or different models, but still contain markers or other elements for separating different data structures, as well as to construct semantic hierarchies, known as self-describing structures [7].

There are many issues concerning information retrieval within this group. Entities are grouped together, even though they have different attributes. Attributes may be in different order with the same meaning, entities can contain a lot of textual data (like in RSS feeds), which still requires further processing.

Basically, xml and json are the most known types of semi-structured data. In this article, we will mainly focus on xml and json information retrieval and semi-structured data will make up individual category of data, based on the differences described in previous text. From this rather immense group of different data structures we will mainly focus on widely used RSS feeds.

Ontology and Open World Assumption

The most widely used storage engines use one way of assuming for handling queries and extracting information. This way is called closed world assumption (CWA). In CWA is assumed that any statement, which is currently not known, is not valid, because all relevant information concerning queries should be stored in a complete knowledge base of the application. Any statement concerning the modelled objects, which is not explicitly stated, can be formally reasoned from those present in the knowledge base. This assumption is very strong and can leave out some interesting results of queries.

Open world assumption (OWA) is an opposite of CWA. The knowledge base does not contain every statement, so a missing statement is not considered to be false. There is an area in between those statements considered to be true and statements considered to be false. If there is no information available, the value of statements cannot be determined [8].

The most widely used instance of CWA are databases – relational or object databases. A good example of OWA based technology is ontology, an explicit specification of shared conceptualization [9]. Ontology is the formal representation of knowledge within a domain and is defined as formal specification of shared conceptualization (abstract, simplified view of world). Ontology can be defined as tuple $O = (C, H, I, R, P, A)$, where:

C is set of entities (concepts),
H is set of taxonomical relationships (which define the concept hierarchy),
R is set of non-taxonomical relationships,
P is set of attributes,
I is set of individuals,
A is set of axioms [10].

In this article, we would like to suggest ontology for storing extracted values from semi structured data and utilize its potential in inferring the use of reasoners.

Transformation of Semi-structured Data

As mentioned in previous sections, our mainly focus will be on one particular category of semi structured data – RSS feeds. RSS stands for Really Simple Syndication and allows you an easy way of syndicating site contents, sharing headlines and news which are automatically updated and written in XML.

For defining the standard structure, RSS version 2.0 will be used, as described in [11]. Root of the XML file is <rss> elements, which contains attribute version and one required child – <channel> element, which contains these required child elements:

1. <title> – the title of the channel (Example RSS Feed)
2. <link> – hyperlink to the channel (www.exampleRSS.com)
3. <description> – description of the channel (Example of RSS feed)
4. One or more <item> elements – defines the data itself

There are many more optional elements for a <channel>, for instance:

- <category> – category of the channel (news)
- <copyright> – copyright information
- <image> with required child elements <url>, <title> and <link>
- <language> – language of feed (en-us)

Every item contains the same required child elements as the <channel> element. The optional elements of <item> are for instance:

- <author> – author of the content
- <comments> – url for the comments
- <enclosure> – including media file with the item

This structure contains many markers with clear meaning; thus, it is possible to process this information automatically, for example store in the database or display on a mobile device. However, there is no definition of relationships between items themselves, no possibility of sophisticated search or to find and infer additional facts from given information. RSS feed contains mostly plain text without more meta information than category or author.

Our goal is to load RSS feed into ontology and run reasoner for getting additional information from the given information. The structure of the ontology will be suggested in the next chapter.

Suggested Ontology Model and Reasoning

For writing the ontology, Protegé from Stanford University will be used [12]. Elements from the RSS feed will be mapped in following way:

- <rss>, <channel>, <title>, <link>, <description> and optional elements will be classes;
- the hierarchy of the elements will be constructed using object properties (properties between classes) – hasChild and isChildOf. Class itself will be subclassOf axiom constructed from object properties, qualified cardinalities and other classes;

- instance of <channel>, <item> and other elements will be individuals with type of the classes (one article will be type item and so on);
- data will be set to an individual using data properties.

A semantic reasoner is used for inferring logical consequences from the set of facts – the structure suggested above. Every reasoner comes with a different mechanism for inferring, but mainly first-order logic is used. An inference usually proceeds with forward and backward chaining mechanisms [13]. Reasoners can be added to Protégé as plugins. Examples of semantic reasoners:

- Chainsaw
- Fact++
- JFact
- HermiT
- Pellet
- RacerPro

For better results, more than just 1 reasoners can be used, with compared results. The RSS feed data will be processed automatically and transformed into individuals in the .owl file of the ontology. Ontology can be written in different formats and syntaxes, which can be easily generated automatically from the application reading and transforming the RSS feeds.

Conclusion

Although the usage of ontologies and open world assumption looks promising, there are many pros and cons of this approach. Concerning information retrieval, loading RSS feed into ontology will not guarantee getting more relationships and hidden information. The main reason for that is textual character of data stored in RSS feeds – even though the ontology copies markers and the structure of feed itself are very good, meaning of the words and text semantics remain unknown to it.

This approach needs to be connected with text mining methods of the data stored in items and channels, so that the individuals stored in ontology can be connected by a reasoner. This remains the main goal of our future research in this area.

Using ontologies and OWA principle in this case can offer following advantages:

- Possibility of inferring additional information with aid of semantic reasoners;
- Detecting inconsistencies in facts of the ontology;
- Getting additional information and facts from the data;
- Usage of more sophisticated queries and semantic search, based on the text meaning.

However, there are many disadvantages as well:

- Possibility of inferring wrong information;
- Difficulty constructing the model, which corresponds to the modelled object (part of the world);
- Reasoning is very time consuming and the speed of ontology querying using SPARQL is generally slower than querying databases in CWA.

Even with all the performance issues, with growing hardware possibilities and coming of the big data, information retrieval is becoming a crucial task, when the quality of resulting information is taken into account. Ontology offers a promising approach to achieve it.

Literature

[1] D.C. Manning, P. Raghavan, H. Schütze. *An Introduction to Information Retrieval.* Cambridge UP, 2009.

[2] A. Gerber, A. Van der Merwe, A. Barnard. A functional semantic web architecture. In *European Semantic Web Conference 2008, Tenerife*, 2008.

[3] Oxford dictionaries. www.oxforddictionaries.com/definition/american_english/data. Accessed 6.10.2014.

[4] www.webopedia.com/term/s/structured_data.html. Accessed 6.10. 2014.

[5] www.webopedia.com/term/u/unstructured_data.html. Accessed 6.10.2014.

[6] D. Clegg. Semi-structured data analytics: Relational or hadoop platform? part 1. www.ibmbigdatahub.com/blog/semi-structured-data-analytics-relational-or-hadoop-platform-part-1. Accessed 24.11.2015.

[7] P. Bueman. *Semistructured Data.* Philadelphia, 1997.

[8] L. A., H. H., T. Z. and V. M. Formální reprezentace znalostí. In *Ostravská univerzita v Ostravě, Ostrava*, 2010.

[9] T. Gruber. What is an ontology? www.ksl.stanford.edu/kst/what-is-an-ontology.html. Accessed 20.10.2014.

[10] R. Girardi. Guiding ontology learning and population by knowledge system goals. *International Conference on Knowledge Engineering and Ontology Development*, pages 480–484, 2010.

[11] www.w3schools.com/xml/xml_rss.asp. w3schools.com Accessed 27.11.2015.

[12] protege.stanford.edu. S. C. f. B. I. Research Accessed 27.11.2015.

[13] M. Horridge, S. Bechhofer. The owl api: A java api for owl ontologies. *Semantic Web Journal, Special Issue on Semantic Web Tools and Systems*, 2(1):11–21, 2011.