Portál pre odborné publikovanie ISSN 1338-0087

Evaluation model for ontology-based information extraction

Šmiraus Michal · Informačné technológie

25.08.2013

Web pages currently includes a large amount of information and documents, which are indeed largely comprehensible to people, but less so clear for the automated search engines, which currently cannot accurately identify what the content of this page conveys. Along with the increasing amount of information available on the web there is a need to effectively identify, recognize, and process the relevant information not only on the basis of a simple full-text search by key words, but also on the basis of knowledge bases using ontology (an explicit description of a concept), the object on one hand, the general development languages, methodologies and software tools and on the other hand also design their own ontologies describing different substantive areas, as well as applications that will use them.

1. Introduction

Idea of semantic integration was first described by Tim Berners-Lee (2001), where the computer sees as a personal assistant who confidentially knows its owner, and can for example recommend and plan his entire vacation (including hotel reservations etc.) in accordance with its capabilities and time preferences. These options in the field of artificial intelligence have always been promised, but it never occurred to their fulfillment. Semantic Web relies on advanced artificial intelligence that can meaning of words and claims process itself, but it is recommended to enrich the classical web of signs and statements written in special languages (such as RDF and OWL).

Smooth transition from the current to the WWW Semantic Web is to be realized through the systematic creation and insertion of metadata. In order to uniquely express the semantics of the terms used is necessary to use the languages stems right from research in the field of ontology. As the main application areas of ontologies are currently understood: knowledge management, e-commerce, natural language processing, intelligent information integration of distributed resources, information retrieval, semantic web portals and intelligent learning systems.

2. Semantic web architecture

The whole concept of the Semantic Web is based on public URI strings of characters that can help identify the source of information. At the very bottom of the pyramid we find the imaginary XML. Markup language with which we can create a structured

document with custom tags is followed by a layer RDF, which allows us to define relationships between objects (resources). The following layer, which allows capture complex ontological structures, is realized through language OWL. The logical layer allows us to describe the relationships between objects and complex application logic due to the application performs the derivation of implicit information. The last layer enables the Trust to ensure the reliability of information.

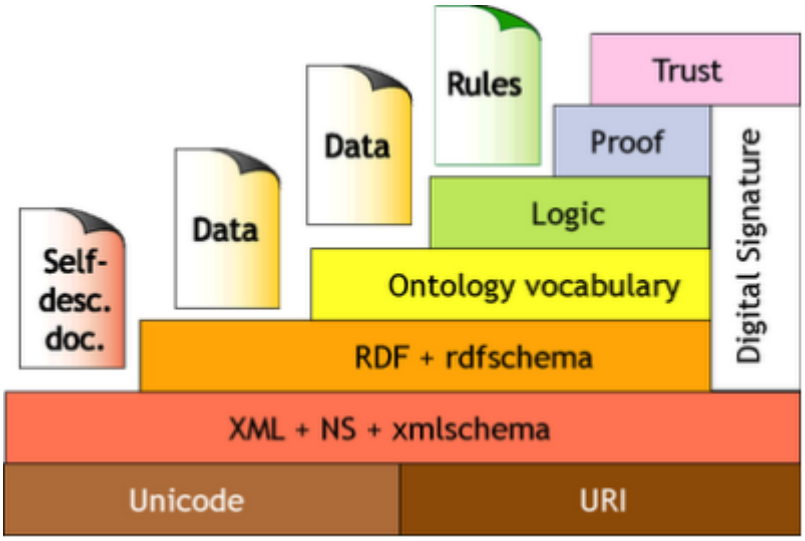

Fig. 1 – Architecture layers of the Semantic Web. (Source: [w3.presentation])

Semantic information woven into normal site allows the computer to manipulate data intelligently. For example, the word 'school' on the current Web site is for the computer only a string of five characters. On the Semantic Web, it is possible to mark the word "school" identifier (URI) of the term school in a much broader description of the concepts and their relationships, which is commonly called ontology.

Then computer in ontology finds that the school is an educational institution that accepts students and that student is a person who has a student ID. In case that text includes information "Paul is studying in this school", it is easier for the computer to infer that Paul is a student who has a student ID. But first there must be an ontology that describes relationships and the text should be annotated (complete with the tags). Creating ontologies is usually done by hand. Automatic derivation of ontologies is still the subject of active research. It is similar for tagging which is often done by hand, but there are also plenty of automatic and semi-automatic tools.

3. Aplication of the ontology spectrum

The very concept of ontology is usually defined as an explicit description of the conceptualization. That is, records concepts and the relationships between them in any language (Smrz, 2004). These languages tend to have great expressive power and often require expert knowledge. But there are also less strong and more widespread means for the description of conceptualization. Is sometimes used almost every user today's web – they're tagging, taxonomies and thesauri. Tagging or assigning labels (where the label is a common character string) has the least expressive power – meaning mediated tagging is small. Tagging on the Web helps users organize information primarily for later retrieval.

Taxonomy is a tree hierarchy of concepts. Usually describes only one type of relationship (eg relationship "is a subclass"), but it may implicitly be more types of relationships, such as is the case of directories in file systems. Subdirectory P may be in the parent directory N, because P (Dachshund) is a type of N (dog). Another subdirectory Q may be N in the parent directory, because Q (head) is a part of the N (dog). Thesaurus also describes a hierarchy of concepts, but clearly states that describes the relationships between concepts. Most are concepts "is more general than," "is less general than" and "is related to".

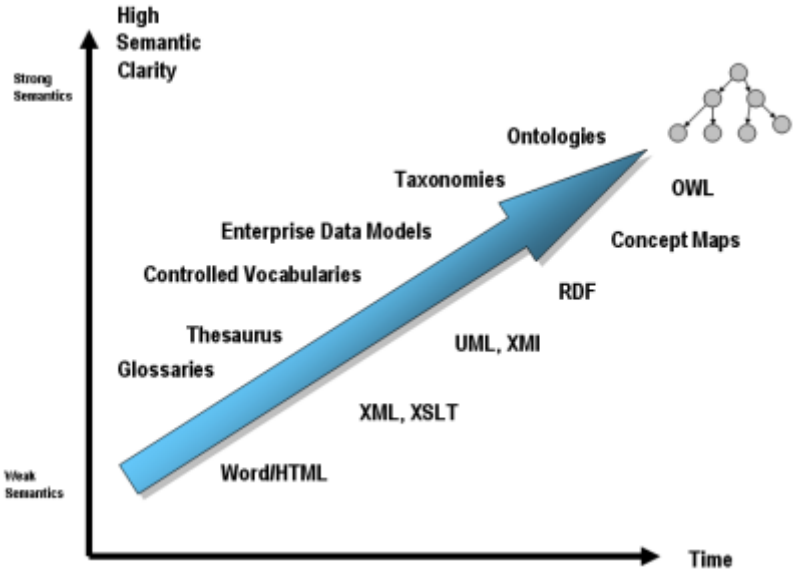

Fig. 2 – Architecture layers of the Semantic Web. (Source: [mkbergman.com])

Ontology is the richest way to describe conceptualization. The ontology language, such as OWL, it is possible to introduce concepts and new relationships, which are then used for further description of terms. Tagging, taxonomy, thesaurus and ontology consists of so-called semantic spectrum (sometimes instead of tagging provides a common and controlled vocabulary).

Although the semantic web is developed for over ten years and fulfill his original vision is still far away, there is already a lot of ideas and specific technologies that could be realized in practice. One example is the so-called faceted search, which in the form of "whispers" has long been used by Google. Among others, such as displaying a "micro formats", which are bits of information complemented with semantic annotations that search engines tell you that the information given is a description of the product, business cards, event, or structure of a site, and interpersonal relationships in the case of social networks.

Support for micro formats and RDF in its core search algorithm joined Google in May 2009 and since then it has been extended to support the entire search index, which motivates the authors of pages to get to their sites started some information semantically mark (this concerns mainly reviews, persons, firms and products) because Google bot with them longer able to recognize them and will continue to offer users as "rich snippets" (Hansson 2009).

Very valuable assistant is also called Linked Data Cloud. It is the ever-expanding set of semantic vocabularies (ontologies) that are publicly accessible, reusable and created

by the principles of Linked Data (Bizer, 2009). These are a few points about how we should properly connect information on the Web, so they went well browse, search and find others sources. Each object should have URL on which applicants can find information about the object and links to other relevant resources. Information is provided on the basis of who is asking. If the computer asks then gets machine-readable data and if a man asks then gets answer in human form (Petrak, 2007).

4. Evaluation of ontology-based extraction

Common approaches to information extraction have been developed at the level of formal models or solely focus on metrics for result quality. Even the presence of ontologies in extraction is only reflected in scoring formulae modified so as to handle taxonomic similarity instead of exact in/correctness of results (Maynard, 2006).

In reality, however, the result quality (typically measured by extraction accuracy) is only one factor of the overall cost; another one is the cost of procurement of extraction knowledge. An exception is the extraction of notorious types of generic named entities (such as peoples' names or locations in English) for which reasonably performing, previously trained tools already exist. However, in most cases, the potential user has to deal with a specific task for which no extraction model exists yet. The extreme alternatives now are 1) to let humans manually label a decent sample of the corpus and train a model, or 2) to prepare the extraction patterns by hand, e.g. in the form of an extraction ontology. Various middle ways are of course possible.

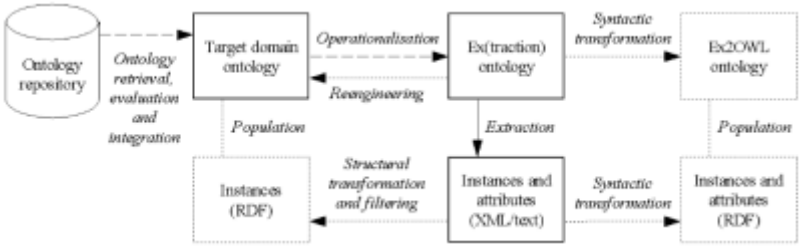

Fig. 3 – High-level schema of ontology-based extraction. (Source: [semanticweb.cz])

Let me sketch a very simple evaluation model that would allow to compare dissimilar extraction methods including the model-building context. Instead of directly comparing the accuracy of different methods, we can declare the minimal accuracy value required for the target application (target accuracy – TA). Then the overall cost (in terms of the human power consumed) will be calculated which is required by those different methods in order for the TA to be reached. For a purely inductively-trained model, the cost amounts to

$$C_I = c_{annot} n_I \tag{1}$$

where c_{annot} is the cost of annotating one elementary unit (such as ontological instance) and n_I is the number of annotations needed to learn a model reaching the TA. Similarly, for an extraction ontology that only uses manual extraction evidence, the cost is

$$C_O = c_{inspect} n_O + C_{ODesign} \tag{2}$$

where $c_{inspect}$ is the cost of merely inspecting (viewing) one elementary unit and n_O is the number of units that had to be viewed by the extraction ontology designer to build a model reaching the TA; $C_{ODesign}$ then is the cost of designing the actual extraction ontology. It is important to realise that $c_{inspect}$ << cannot (among other, $c_{inspect}$ does not have to deal with exact determination of entity boundaries, which is a well-known problem in creating the ground truth for IE) and most likely also $n_O < n_I$; what now matters is whether this lower cost in C_O is/not outweighed by the relatively high cost of $C_{ODesign}$.The model can be arbitrarily extended: e.g. for hybrid approaches and also the cost of deciding which attributes are to be extracted using which method could be consider – inductive vs. manual.

5. Conclusion

Vision and technical details of the semantic web is still the subject of debate, particularly among software and knowledge engineers, which relates to the implementation of most. Furthermore the semantic web discussion even among the creators of web pages and applications. Most of them expect the results of developments with interest, although attitudes are not unambiguously positive. Someone expect increase in complications when publishing on the web and the orientation of the semantic web for the benefit of users (for which the site is a source of information, knowledge and services) sees as a dangerous change of condition. They are probably those web users for whom the primary site service options as simply publish.

It is clear that the consistency is easier to achieve in a closed area. That is why today there are a relatively high number of domain ontologies. Current research question is whether the successful implementation of information systems of the future is sufficient to increase the number of pebble mosaics of semantic web or the need to involve a general, broad-based “embracing” ontology. The most important is probably the time intensity of such a resource. To overcome the vulnerability may result in the traditional way – try to use what is already done. The aim of such an approach to building ontologies is the integration of existing databases, extensive knowledge base constructed for other purposes, and other sources of “clean” refine stored and eventually integration into a single ontological structure.

References

1. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.
2. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), 1-22.
3. Hansson, O. (2009, May 12). Google engineering explains microformat support in searches. Retrieved from http://radar.oreilly.com/2009/05/google-adds-microformat-parsin.html
4. Maynard, D., Peters, W., Li, Y. (2006). Metrics for Evaluation of Ontology-based Information Ex-traction. Workshop EON’06 at WWW’06.
5. Petrak, J.: (2007, July) Vse o propojovani dat (Linked Data). June 2007 Retrieved from: http://zapisky.info/?item=vse-o-propojovani-dat-linked-data&category=vse-o-semantickem-webu

6. Smrz, P., & Pitner, T. (2004). Semanticky web a jeho technologie (3). Zpravodaj UVT MU. ISSN, 1212-0901.

Ing. Michal Šmiraus, Thomas Bata University in Zlín, Faculty of Applied Informatics, Nad Stráněmi 4511, Zlín, 760 01