Web Semantization - a process of automated annotation

Jan Dedek
Department of Software Engineering
Charles University in Prague
Malostranske namesti 25
Prague, Czech Republic
jan.dedek@mff.cuni.cz

Alan Eckhardt
Institute of Computer Science,
Academy of Sciences of the Czech Republic
Pod Vodarenskou vezi 2
Prague, Czech Republic
eckhardt@cs.cas.cz

Peter Vojtas
Institute of Computer Science,
Academy of Sciences of the Czech Republic
Pod Vodarenskou vezi 2
Prague, Czech Republic
vojtas@cs.cas.cz

ABSTRACT

We understand Web Semantization as an automated process of increasing degree of semantic content on the web. We are convinced that Web Semantization can be a way of gradual realization of the Semantic web – an attractive vision of the web future, which is changing the way people think and act. Our idea is supported by models, methods, prototypes and experiments with a web repository, WIE with assisted learning, automated annotation tools producing third party semantic annotations, a semantic repository serving as a sample of semantized web and a proposal of an intelligent software agent. We are working on a proof of concept that even today it is possible to develop a semantic search engine designed for software agents.

Keywords

Semantic Web, Web Content Mining, Linguistic Analysis

1. INTRODUCTION

Recently, Lee Feigenbaum, Ivan Herman, Tonya Hongsermeier, Eric Neumann and Susie Stephens in their Scientific American 2007 article [6] conclude that "Grand visions rarely progress exactly as planned, but the Semantic Web is indeed emerging and is making online information more useful as ever". They support their claim with success of semantic web technology in drug discovery and health care (and several other applications). By our opinion, these are mainly corporate applications with data annotated by humans.

Ben Adida when bridging clickable and Semantic Web with RDFa ([1]) assumes also human (assisted) activity - annotations of newly created web resources.

What to do with the content of the web of today or of pages published without annotations? The content of the web of today is too valuable to be lost for emerging semantic web applications. We are looking for a solution how to make it accessible in semantic manner.

We would like to address the problem of semantization (enrichment) of current web content as an automated process of third party annotation for making at least a part of today's web more suitable for machine processing. It would hence allow intelligent tools to search and recommend things on the web (as advocated by Tim Berners-Lee [2]). Our approach is not limited only to one specific domain of application, it can be used across all the disciplines presented on the web. Fro example in [3] we have demonstrated how our approach can help with environmental issues.

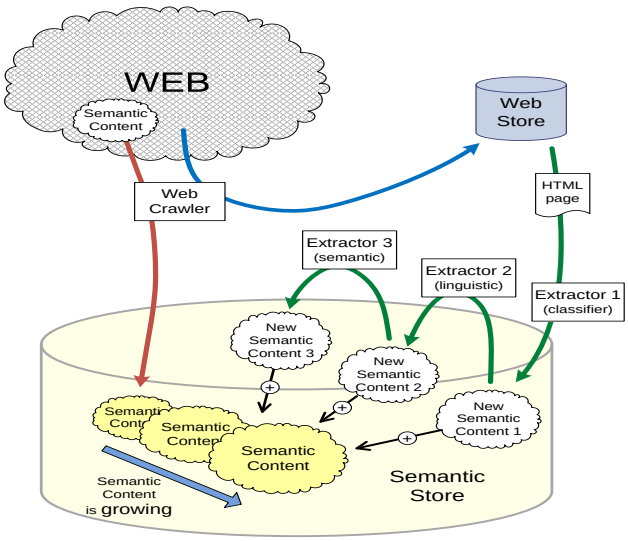

Figure 1: The process of semantization of the Web

2. PROTOTYPES AND EXPERIMENTS

Our main idea is to fill a semantic repository with information that is automatically extracted from the web and make it available to software agents. We are working on a proof of concept that this idea is realizable and we give results of several experiments in this direction.

Our web crawler (see Fig.1) downloads a part of the web to the web repository (web archive). Page classifier selects those parts of web archive which are suitable for further semantic enrichment (we are able to enrich only a part of resources). More semantic content is created by several extractors and annotators in several phases.

(1) The idea of a web repository.

The web repository is a temporal repository of web documents crawled by a crawler. The repository supports document's metadata, e.g. modification and creation dates, domain name, ratio of HTML code/text, content type, language, grammatical sentences etc. It keeps track of all changes in a document and simplifies access to and further processing of Web documents. We are experimenting with the web crawler Egothor[1] 2.x and it's web repository. For pages from hidden web we have used RSS feeds.

[1]http://www.egothor.org/

(2) The second idea is to split annotation process into two parts, the first is domain independent intermediate annotation and the second is domain dependent user directed annotation of a smaller portion of the web. Semantic enrichment is in fact a data mining task (although a special one) - to add to web documents a piece of knowledge, which is obvious for human perception and hard for a machine. That means to annotate data by concepts from an ontology which is the same as to map instances to ontology. Such a data mining task will be easier to solve when there is a sort of a repetition (modulo some similarity).

Both annotation tasks should be automated, with some initial human assisted learning. The first part of learning could require assistance of a highly skilled expert; the second (probably faster part) should be doable by an user with average computer literacy.

Domain independent intermediate annotation can be done with respect to general ontologies. We distinguish two cases: (1) textual resources and (2) structured resources. For *textual resources* the ontology we use is the general linguistic PDT tectogrammatical structure [7] which captures semantic meaning of grammatical sentences in Czech. For example English language can be parsed in many different ways (most often according to some kind of grammar). **Current solution** makes use of a tree structure of the annotations. In this paper we will present our experience with Czech language and tectogrammatical structure that we have used for domain independent intermediate annotation of pages dominantly consisting of grammatical sentences.

For *structured survey or product summary pages* we assume that their structure is often similar and the common structure can help us to detect data regions and data records and possibly also attributes and their values from detailed product pages. Annotation tools will be trained by humans here also – nevertheless only once for the annotation of the whole repository. **Current solution** uses similarities of DOM structure of pages.

Domain dependent annotation is concerning only pages previously annotated by general ontologies. This makes second annotation faster and easier. An assistance of a unskilled human is assumed here for each domain and a new ontology. Nevertheless we cannot expect an average user to annotate a big amount of resources, so we have to work with the realistic (30 to 40) size of test set.

Repetitions in *textual pages* make possible to learn a mapping from structured tectogrammatical instances to an ontology. **Current solution** uses ILP tool `PROGOL` over annotations obtained in the first domain independent part. For traffic accident reports, we were able to learn rules for finding sentences reporting on injuries (note that linguistic is need in sentences like 'nobody was injured', simple key word search does not work in this case)

```
Rule set1:
injured(A):-edge(A,C),edge(C,D),t_lemma(D,accident).
injured(A):-edge(A,C),t_lemma(A,injure),neg(C,false).
```

See `Rule set1` as an example of learned rules. Fig. 2 shows the 4-fold table where rows depict classification by rule set and columns classification by user. Results are promising, nevertheless tagging is done by unskilled user for whom the tagging is usually tedious.

For *structured pages* the domain dependent annotation **current solution** uses simple ontology in a form of relational schema provided by user. We experiment with several heuristics for learning regular expressions for attribute values.

	User tag	¬User tag
Rule set1	11	1
¬Rule set1	4	22

Figure 2: Results on the realistic test set

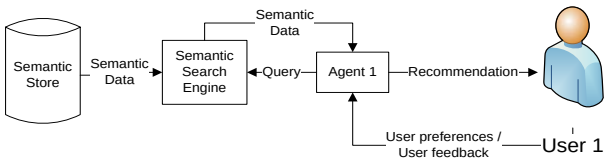

Figure 3: Querying Semantic Search Engine

(3) Next idea is to design a semantic repository. It should store all the semantic data extracted by extraction tools and accessed through a semantic search engine. **Current solution** uses [4]. It supports RDF storage and SPARQL querying.

(4) Design of an software agent, which will give the evidence that our semantization really improved general web search. Besides using annotated data it should also contain some user dependent preference search capabilities. **Current solution** exploits user preference modelling technique published in our previous work [5].

The process of a user agent searching and making use of semantic search engine is represented in Figure 3.

3. CONCLUSION

In this paper we have presented our work on web semantization – models and prototypes making the web to the web of things ([2]). Preliminary experiments are promising.

This work was partially supported by Czech projects 1ET100300517, 201/09/0990 GACR and MSM-0021620838.

4. REFERENCES

[1] B. Adida. Bridging the clickable and semantic webs with rdfa. *ERCIM News*, 72:24–25, January 2008.

[2] T. Berners-Lee. The web of things. *ERCIM News - Special: The Future Web*, 72:3, January 2008.

[3] J. Dědek and P. Vojtáš. Web information extraction for e-environment. In *TOWARDS eENVIRONMENT (accepted for publication)*. Prague, Czech Republic, http:/www.e-envi2009.org.

[4] J. Dokulil, J. Tykal, J. Yaghob, and F. Zavoral. Semantic web infrastructure. In *IEEE ICSC*, pages 209–215. IEEE, 2007.

[5] A. Eckhardt, T. Horváth, and P. Vojtáš. Learning different user profile annotated rules. In *SUM*, LNCS 4772, pages 116–130. Springer, 2007.

[6] L. Feigenbaum et al. The semantic web in action. *Scientific American*, 297:90–97, 2007.

[7] M. Mikulová et al. Annotation on the tectogrammatical level in the PDT. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep., 2006.