Web Semantization

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

Web Semantization is a concept we introduce in this paper. We understand Web Semantization as an automated process of increasing degree of semantic content on the web. Part of content of the web is further usable, semantic content (usually annotated) is more suitable for machine processing. The idea is supported by models, methods, prototypes and experiments with a web repository, automated annotation tools producing third party semantic annotations, semantic repository serving as a sample of semantized web and a proposal of an intelligent software agent. We present a proof of concept that even today it is possible to develop a semantic search engine designed for software agents.

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing; H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval; I.2.4 [**ARTIFICIAL INTELLIGENCE**]: Knowledge Representation Formalisms and Methods

General Terms

Web Semantization

Keywords

Semantic Web, Web Content Mining, Linguistic Analysis

1. INTRODUCTION

In their Scientific American 2001 article [3], Tim Berners-Lee, James Hendler and Ora Lassila unveiled a nascent vision of the semantic web: a highly interconnected network of data that could be easily accessed and understood by a desktop or handheld machine. They painted a future of intelligent software agents that would "answer to a particular question without our having to search for information or pore through results" (quoted from [7]). Lee Feigenbaum, Ivan Herman, Tonya Hongsermeier, Eric Neumann and Susie Stephens in their Scientific American 2007 article [7] 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". L. Feigenbaum et al. support their claim with success of semantic web technology in drug discovery and health care (and several further applications). 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 by annotations of newly created web resources.

But 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 sulution how to make it accessible in semantic manner.

In this paper 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 and hence enabling it intelligent tools for searching and recommending things on the web (see [2]).

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 give 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 Store). Resources with semantic content can be uploaded directly to the semantic repository (Semantic Store). Extractor 1 (classifier) extracts those parts of Web Store which are suitable for further semantic

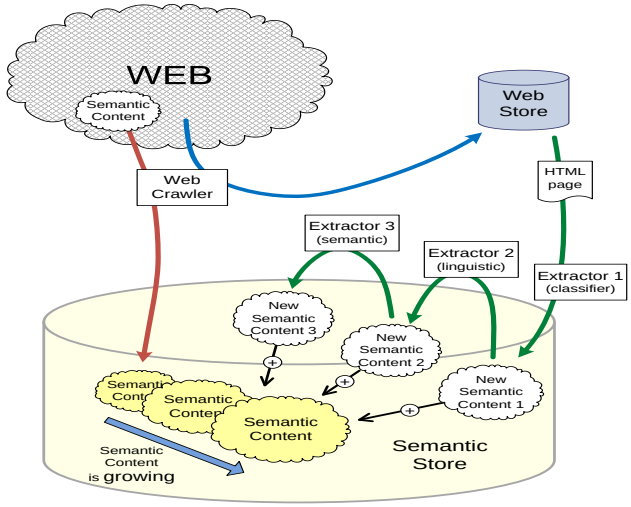

Figure 1: The process of semantization of the Web

WWW2009, April 20-24, 2009, Madrid, Spain.
.

enrichment (we are able to enrich only a part of resources). More semantic content is created by several extractors and annotators in several phases. The emphasis of this paper is on the automation and the usage of such extracted/enriched data.

2. IDEA OF WEB SEMANTIZATION

First idea is the idea of a web repository. It develops as follows in details. 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 not 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).

We decided to enrich only textual content for the present, no multimedia (this substantially reduces the size of information we have to store). Especially we restrict to the pages with content consisting either dominantly of grammatical sentences (let us call them textual pages) and those containing large number of table cells (let us call them tabular pages). Of course this division need not be disjoint, and will not cover the whole web.

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 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. We have been filled this repository with several terabytes of textual part of Czech web (domain *.cz) and it very simplified access to this data.

Second idea is to split annotation process to two parts, one domain independent intermediate annotation and second domain dependent user directed annotation. Both should be automated, with some initial human assisted learning. This 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.

The first domain independent annotation will serve as an intermediate annotation supporting further processing. This will run once on the whole suitable data. Hence initial necessary human assistance in training can be done by a highly skilled expert (indeed it can be a longer process).

The second domain dependent annotation part can consists of a large number of tasks with different domains and ontologies. This should be possible to be executed very fast and if possible with assistance of an average computer skilled user. We can afford this having intermediate annotation.

Domain independent intermediate annotation can be done with respect to general ontologies. First ontology is the general PDT tectogrammatical structure [8] (it is not exactly an ontology written in ontology language, but could be considered this way) which captures semantic meaning of a grammatical sentence. This is the task for computational linguistics; we make use of their achievements which were originally focused on machine translation. Of course tectogrammatical structure is not the only option for this purpose. English language for example can be parsed in many different ways (most often according to some kind of grammar). All the other possibilities are applicable, but in our current solution we make 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.

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

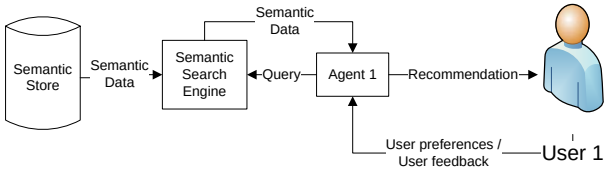

Figure 2: Process of querying Semantic Search Engine by an user agent

For structured survey or product summary pages (we call them "tabular 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 values from detailed product pages. Here annotation tools will be also trained by humans – nevertheless only once for the annotation of the whole repository.

We can see an interesting inversion in the use of similarity and repetition.

While for tabular pages we use similarities in the intermediate phase, for textual pages we use similarities in the domain dependent phase where similar sentences often occur.

Domain (task) dependent (on demand) annotation is concerning only pages previously annotated by general ontologies. This makes second annotation faster and easier. An assistance of a human is assumed here for each domain and new ontology. For textual pages repetitions make possible to learn a mapping from structured tectogrammatical instances to an ontology. This domain (task) dependent task can be avoided by a collaborative filtering method, assuming there is enough users' acquaintance.

Third important idea is to **design semantic repository**. It should store all the semantic data extracted by extraction tools and accessed thorough a semantic search engine. Of course many different problems (e.g. querying and scalability) are connected with this but they are at least partially solved now days. Let us mention work of our colleges [5] that is available for use.

The semantic repository should also contain some sort of uncertainty annotation besides above mentioned ontologies. The main reason is that annotation process is error prone and we can have in future different alternative annotation tools and aggregate results. This aspect is not further described in the paper but can be found with many details in [4] and [6].

Last, but also important idea is **(4) to design at least one agent** which will give evidence that our semantization really improved general web search. Besides using annotated data it should also contain some user dependent preference search capabilities.

The process of a user agent searching and making use of

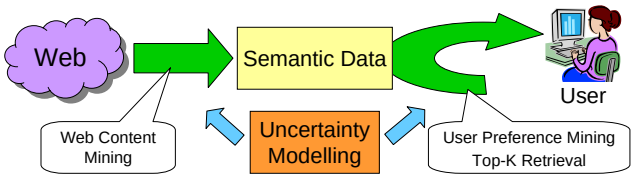

Figure 3: Connecting Web and User

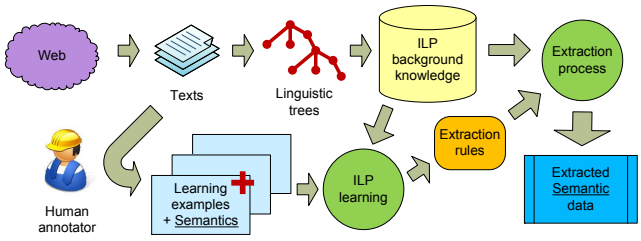

Figure 4: ILP Learning of Extraction Rules

semantic search engine is represented in Figure 2.

Our main focus in this paper is on the agent and on the extractors.

3. ACKNOWLEDGMENTS

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 - Special: The Future Web*, 72:24–25, January 2008.

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

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web, a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):34–43, May 2001.

[4] J. Dědek, A. Eckhardt, L. Galamboš, and P. Vojtáš. Discussion on uncertainty ontology for annotation and reasoning (a position paper). In P. C. G. da Costa, editor, *URSW '08 Uncertainty Reasoning for the Semantic Web - Volume 4*. The 7th International Semantic Web Conference, 2008.

[5] J. Dokulil, J. Tykal, J. Yaghob, and F. Zavoral. Semantic web infrastructure. In P. Kellenberger, editor, *First IEEE International Conference on Semantic Computing*, pages 209–215, Los Alamitos, California, 2007. IEEE Computer Society.

[6] A. Eckhardt, T. Horváth, D. Maruščák, R. Novotný, and P. Vojtáš. *Uncertainty Issues and Algorithms in Automating Process Connecting Web and User*, volume 5327 of *Lecture Notes in Computer Science*. Springer Verlag, 2008.

[7] L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neumann, and S. Stephens. The semantic web in action. *Scientific American*, 297:90–97, December 2007.

[8] M. Mikulová, A. Bémová, J. Hajič, E. Hajičová, J. Havelka, V. Kolářová, L. Kučová, M. Lopatková, P. Pajas, J. Panevová, M. Razímová, P. Sgall, J. Štěpánek, Z. Urešová, K. Veselá, and Z. Žabokrtský. Annotation on the tectogrammatical level in the prague dependency treebank. annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep., 2006.