Enhancing the Web Experience by Freely Available Metadata

Peter BUGÁŇ*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
bugisoft@gmail.com

Much of information available on the Internet can be easily understood by people, but not at all by computers, which present the actual content of the web pages. The problem is often solved by adding machine-understandable metadata. In addition to "lightweight" metadata in the form of 'meta' html tag, which is devoted mainly to the web search engines, we know also more formal metadata defined within Semantic Web initiative. The vision is a Web, in which the meaning (semantics) of information and services is shared across the web applications. Metadata contained within the Semantic web create so called Linking Open Data cloud, that is growing every year, although this growth is slow. Only interesting Semantic Web applications can persuade users to participate more actively in the Semantic Web initiative, which would hopefully increase the growth of the available metadata until a self-strengthening threshold is reached.

When we are reading articles on the web, e.g., online newspapers, we are often unable to understand it quickly (especially longer sections) and we need to re-read it several times. Therefore, we underline the most relevant keywords on the web pages and annotate them with automatically generated content. Underlined keywords should help readers to quickly recognize the main words of an article while the actual content of annotations should allow for better understanding of the underlined word.

Process of annotation consists of these parts:

1. Web document processing – as we annotate only the content part of a web document, we can ommit unnecessary parts such as menus, advertisements and focus only on segments which are of user's interests.
2. Lookup of words, that should be annotated
3. Annotations filtering
4. Creation of annotations content
5. Visualisation of annotations

For keyword search we use OpenCalais web service, which extract instances like well-known people, companies, organizations, geographical indications (states, cities,

* Supervisor: Ing. Michal Barla, Institute of Informatics and Software Engineering

Spring 2010 PeWe Workshop, April 18, 2010, pp. 57–58.

rivers …) from the given text or URL. The important feature of OpenCalais is that it provides also metadata for the selected keywords such as additional facts about retrieved keywords (i.e., that Robert Hughes is a reporter from BBC) or relevance of the retrieved keyword according to the rest of the document. The metadata help us to reduce irrelevant results in search for content of annotation.

After keyword extraction, we decide which words are to be annotated – which are in our case the most relevant words found in the article. As we already mentioned, the relevance is acquired from OpenCalais.

Content of annotations is tailored according to type of a particular article. For instance, if an article is about sport, the annotation will contain information relevant to the sport topic. This allows us to provide different annotations of the same word used in different context. A word 'Lisbon' within a political article would be annotated with information about mayor of the Lisbon, in case of a sport article, the annotations will contain information about sport events which were or will be hosted in Lisbon or about athletes coming from this city.

Annotation could contain:

- Textual information (e.g. population of a country)
- Hyperlinks to the external resources (e.g., reference to a photography of a monument in the city, or reference to the article on Wikipedia)
- Reference to other entities (e.g., when information in the note is that, Bratislava is the capital of Slovakia, by clicking on Slovakia, we will view information about that instance)

As we have indicated, the information source of our annotations is Semantic web. For geographical instances we use specific resources like Dbpedia.org, Wikitravel.org and Factbook, which we evaluated to be the most relevant resources for this kind of information. When we search for metadata about people, we use semantic search engine Sindice.org.

The last part of the annotation process is actual visualization of annotations. We decided to visualize annotated words by underlining them. There is a possibility to setup different colours for different types of instances. Clearly, it is not necessary to underline and attach an annotation to each and every occurrence of an instance within a document, which would for sure overload the readers and would not be very comfortable. Therefore, we underline only first occurrence of the word. The contents of the notes are shown in tool-tips, which appear after hovering the mouse over the underilned word.

From the technical point of view, our solution is based on enhanced proxy server being developed at Institute of Informatics and Software Engineering (peweproxy.fiit.stuba.sk).

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.