Web Information Extraction by Semantic Tagging

Mirel Cosulschi[1], Roberto De Virgilio[2], Tommaso Di Noia, Roberto Mirizzi[3]

1 Motivation and Goal

An important aspect of research for Web information extraction relates to the inference of complex reasoning and correlation based on distributed information available in many different Web data sources. By defining the semantics of information and services available on the Web, the World Wide Web becomes a vast store of information that can be easily processed by computer applications. *Semantic Web* aims at creating a universal medium where data and knowledge can be exchanged between applications. In this framework, we extend a structure discovery technique [1] that: **(i)** identifies blocks grouping semantically related objects occurring in Web pages, and **(ii)** generates a logical schema of a Web site by semantic tagging support. Very often a Web page does not relate to a single semantic topic. The decomposition of a Web page into smaller semantic annotated fragments would surely help in supporting more accurate results for semantic Web searches, richer data integration and better navigation experience. On the one hand, in case the original Web page is already annotated, the annotation can be used by the Web page segmentation process together with the visual and structural information. On the other hand, when no annotation is available (this is the most frequent case in the current Web), the page has to be decomposed via a segmentation process and then each extracted page-block has to be annotated. In case of automatic annotation, the latter approach could facilitate this process due to the reduction in the size of input data. Using a Data Reverse Engineering process [1], the logical schema of the Web site can be obtained from the conceptual representation and then one can label the extracted HTML blocks using *RDFa* (e.g. an HTML block with *paper details* as *title* and *author*). Figure 1 sketches the process. The semi-automated annotation of a page fragment might exploit both techniques for named entities extraction and the availability of a huge base of shared and inter-linked data, the so called *Web of Data*. The first step towards the interpretation of the information contained within an extracted Web block passes through the identification of named entities [3] in the block-body. These entities may be automatically classified with respect to a shared generic ontological schema (e.g. *DBpedia.org*, *OpenCyc.org* or *FreeBase.com*) or highly specialized and contextualized ones (see *MusicBrainz.org* as an example). Actually, information within a block might not refer only to named entities but also to generic concepts. In these cases, the use of semantic-enhanced vocabulary such as *WordNe* might help in the identification of the main concepts

[1] University of Craiova, Romania {mirelc@central.ucv.ro}
[2] Università Roma Tre, Rome, Italy {devirgilio@dia.uniroma3.it}
[3] Politecnico di Bari, Bari, Italy {t.dinoia@poliba.it, mirizzi@deemail.poliba.it}

representing the information in the block. Once the concepts are identified, the use of ontologies allows to semantically enrich such classes. Similarly to [2], the computed set of both named entities (if any) and words' senses allow to find the most appropriate RDF resources to annotate identified pieces of information within the HTML blocks. The relevance of a set of resources with respect to a single piece of information in the extracted HTML block may be computed taking into account: **(i)** domain information from the selected (portion of the) ontology, **(ii)** annotations related to all the other blocks extracted from the same HTML page, and **(iii)** annotations either from similar pages or pages from the main Web site. Once the RDF resources have been retrieved and ranked, they can be used to semi-automatically embed RDFa annotations within HTML blocks.

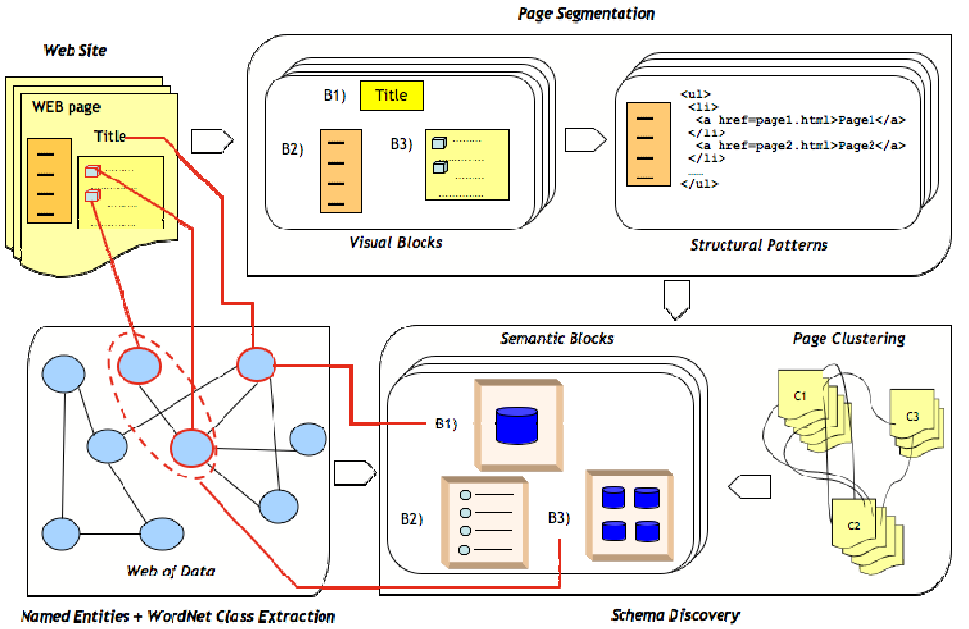

Figure 1. Web Information Extraction Process

References

[1] R. De Virgilio, R. Torlone. A Structured Approach to Data reverse Engineering of Web Applications. *In Int. Conf. on Web Engineering (ICWE'09)*, 2009.

[2] J. Garcia, M. DAquin, E. Mena. Large Scale Integration of Senses for the Semantic Web. In *18th International World Wide Web Conference (WWW'09)*, 2009.

[3] T. Poibeau, L. Kosseim. Proper Name Extraction from Non-Jurnalistic Texts. In *Computational Linguistics in the Netherlands*, 2000.