Towards the Discovery of Implicit Metadata in Commercial Web Pages

Vojtěch Svátek, Martin Kavalec and **Jiří Klemperer**
Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
{svatek,kavalec,klemper}@vse.cz

In addition to 'cropping' explicit metadata, Semantic Web agents should also be able to 'excavate' *implicit* metadata hidden in the content and structure of web pages, to situate them with respect to standards, and record in explicit form. Hot candidates for implicit metadata extraction are the websites of small organisations, which are are poorly equipped with explicit metadata, relatively closed (with few important external links), and their scope of semantic 'messages' is limited to those most important for the customer. To attack so general a class of pages, we need *cross-domain* extraction patterns, relying on generic keywords within simple, micro-level structures: *sentences* and low–level *HTML mark-up*.

The text represented with valid *sentences* (in paragraphs) typically amounts to the description of the company profile. A frequent structure is that of syntactic predicate containing an *indicator* term, and the syntactic *object* (depending on the predicate) containing *target metadata*. To obtain the indicators, we relied on frequency analysis of terms occurring in the syntactical neighborhood of target metadata in the training set of pages. To elide the tedious task of training data acquisition and labeling, we have developed a novel method reusing the information contained in a *public WWW directory*; Open Directory (`www.dmoz.org`) has been used. Since the hierarchical headings of the directory nodes often coincide with target metadata (i.e. generic names of products and services, as well as 'domains of competence') in the pages referenced by these nodes, we obtain a large sample of *labeled* training cases (sentences with 'heading' terms marked up as 'target metadata') with no human effort. The result of the first learning experiment was a collection of approx. 20 'indicator' verbs (see Table 1) with high coverage: within randomly selected pages, one or more of them occurred in 70-80%. While the present method does not make difference between different types of metadata (such as products vs. areas), we plan to enhance it towards separate induction of different types; we have constructed a (meta)ontology of directory headings for this purpose. For more details see [Kavalec, 2001].

Subject	**Predicate-object pair**	**Examples of predicate indicators**
Homepage	Predicate and object in sentence parse tree	*supply, offer, specialize, sell, manufacture, buy, provide*
Homepage	two adjacent HTML structures	*address, projects, references*
Page or section	two adjacent HTML structures	*price, code, shipment*

Table 1: Typical RDF-like structures of implicit metadata

Furthermore, low–level *HTML structures* together with a few *special symbols* (such as colon) may provide similar discourse structuring to a text *below the sentence level*. By generalization of approx. 60 pages, we have observed that the 'predicate-object' pair may be expressed, for example, by a heading (or font-emphasized) element followed by a paragraph or list, as well as by the row of a two-column table, in which the first column is font-emphasized. It seems that the variety of 'messages' in structured HTML code is greater than for sentences, since there are several types of commercial information that are inherently structured or quantitative (addresses, prices, product codes); it would be impractical to embed them into sentences. The second and third row of Table 1 attempt to distinguish the metadata related to the company and to a particular product.

The declared goal of many Semantic Web activities is the *harmonization* of the metadata content, including *product classifications* [Guarino, 2001]. We hope that our effort is complementary: for example, automated extraction of *ad hoc* generic *product names* could collect material for updating the product ontologies with respect to the dynamics of new product development. One step further to *RDF representation* of discovered metadata would be storing the predicate part of these triples (in particular, concepts abstracted from generic indicator terms) as RDF resources rather than as literals.

Acknowledgments

The research is partially supported by grant 201/00/D045 of the Grant Agency of the Czech Republic.

References

[Kavalec, 2001] Martin Kavalec, Vojtěch Svátek, and Petr Strossa. Web Directories as Training Data for Automated Metadata Extraction. *Semantic Web Mining*, Workshop at ECML/PKDD-2001, pages 39-44. Freiburg 2001.

[Guarino, 2001] Nicola Guarino et al. Focused Clusters of Content Standards. OntoWeb project, D3.2, 2001.