Ontology based Extraction of Company Profiles

Sylke Krötzsch, Dietmar Rösner
Otto-von-Guericke-University Magdeburg
Institute Knowledge and Language Engineering
P.O.box 4120, 39106 Magdeburg, Germany
{kroetz | roesner}@iws.cs.uni-magdeburg.de

Motivation

We report about a research project, which uses XML technologies for the semantic web. The WWW is a broad source with heterogeneous information, so that with help of different methods data extraction is possible. We choose for our work the example "casting technology" in view of an industrial application. There exists a lot of HTML documents which are to be evaluated both with respect to document structure as well as data content. This includes not only structural tagging in XML but also linguistic analysis of texts.

Casting Technology and Company Profiles

In oder to find information in the WWW we can use several techniques. As application field we have choosen the creation of company profiles. In this paper we use the example of foundries. Already carried out analyses comprehends the workout of offers for castings in foundries [3]. We determined that a multitude of information has to be considered in order to create an offer. This enfolds both customer information (e.g. customer master data) and product specifications (e.g. size, material, weight). A possible scenario is the following: a product can not completely be produced in a company. We propose in this case the use of company profiles for choosing a supplier company.
Analyses have shown, that the WWW is used in order to get information about potential suppliers. In this way the foundry itself acts as the customer because the company searches suppliers. But also companies of other industrial sectors (e.g. motor industry) use the WWW in order to inform themselves about possible suppliers (e.g. foundries). This proceeding is necessary, for instance if a company can't produce specific products because of preciseness requirements. Here the company depends on other companies. Especially in our application field "foundry industry" several domain specific aspects are to be considered. These includes on the one side product related points, like size and weigth of the casting, material, moulding processes and quality assurance resp. certificates, and on the other side kinds of production (e.g. single-piece work, small, middle, and mass production). In addition the location of the potential supplier has to be considered, in order to reduce transport costs.
Based on information about the concerns user specific company profiles can be created. In various industrial sectors approaches based on Semantic Web exist, however not for casting industry. Thus, the results of our work are utilized for the creation of company webpages, in order to describe implicitly available information concise and semantic annotated.
A company profile includes different information about a specific company. In the context of our application field the key aspects of the company profile are production plants and production processes, quality certificates, producible materials and producible dimensions for products.

Another interesting aspect are data about the previous customers of the companies. For the inquiring concern is it important to know, which products (e.g. boxes, engine blocks, cylinder heads, axles or rotors) for which industrial sectors (e.g. motor industry, wind power industry, maschine building industry, shipbuilding industry, ...) are produced by the "individual" company. These information allows inferences with respect to performed quality requirements of the company during production processes. Likewise the company profile contains data of contact persons as well as address and additional contact information.
The targeted evaluation of existing HTML pages in the WWW is very important for the user. With the help of company profiles of different selected foundries the user can well directed compare in order to find potential suppliers.

Creation of Company Profiles

In the initial stage of tool development for the creation of company profiles fixed criteria are used, in order to filter important information from webpages. These critera are:

- dimensions and weight,
- material,
- customers,
- quality certificates as well as
- products and industrial sectors.

It must be pointed out that in webpages on the one hand dimensions of the casting and on the other hand dimensions of the moulding boxes are denoted. For this reason the evaluation must be designed flexible for adapting the company profile. In future work these criteria should be arbitrary in order to design the custom-made company profiles.
Several techniques are utilizable for the generation of company profiles, for instance the tagging of the specification with XML tags. Further on is to note that the WWW (resp. company specific webpages) is to search not only with help of conventionally search engines for specific words or word combinations. At beginning we use a tool [1] in order to search for specific terms in the internet. Presently the tool use a conventional search engine and evaluates the located webpages and transform these into text files. Future planned analyses will expand the tool with various functionalities such as:

- classification of company specific webpages,
- link tracing of coherent webpages (e.g. webpages of one company),
- evaluation of phrasal patterns and
- evaluation of HTML tables and transformation in the XML tagging.

This includes also the consideration of these creteria referring to "all HTML webpages " of one company. The information of the company profiles are tagged in XML, in order to generate in a favored format of an profile with the help of XSL. In the following we explain the functionalities mentioned in more detail.

Classification of Webpages

An aspect of our task is the detection of all relevant webpages of one company. A result of precursory analyses is that the structure of webpages with respect to the available information differs from company to company. For this reason it is necessary to classify the webpages and individual webpages. Company specific webpages of our application field “casting technology“ can be divided in:

- *Information Pages:*
  Information pages are text pages which have a central continuous text. These “simple“ webpages contains both texts and graphics. Likewise links to other company internal or external webpages are possible.

- *Lead Pages:*
  Lead pages very often based on frames and are used for a central structuring. Frames also refer to further company internal webpages. The navigation through the webpages of a company is to be made by the lead page resp. the frame. In addition links to external company independent webpages are possible, too. The company internal webpages are very often classified as information pages.

- *Overview Pages:*
  These pages are e.g. offered by consolidated companies and contain links to the associated concerns. For instance consolidated companies may consist of 8 foundries and refer to the webpages of the individual foundries on the main webpage resp. overview page. These referred webpages may be classified both as information pages and as lead pages. They contain information, for instance specifications of the company, like name, producible materials or producible weight and contact persons.

General fundamentals for document classification are analysed in [6]. Information pages are very often used by companies for their own representation in the WWW. Lead pages are only used for navigation through the company internal webpages. For this reason we will discuss only casting specific information pages. For instance these pages comprise the special tags <META> and <TITLE>. These tags contain very often information about the company, like keywords, name of the company, or additionally specifications. An evaluation of these tags is necessary in order to draw conclusions from the other company internal webpages.

Example 1 (<TITLE>-Tag)

```
<TITLE>Foundry XXX GmbH</TITLE>
```

Lots of foundry webpages use the <TITLE>-tag in order to specify the name of the foundry (Example 1). In this case all company internal information pages use this <TITLE>-tag, so the tag can be utilized for comparisons. By comparison of the company name (as part of the link in a webpage) and the <TITLE>-tag of a company internal webpage an affiliation to the company webpages can be inferred.
Another indication for company internal webpages is the aspect, that links not only contain complete WWW addresses but also refer to single secondary webpages. In this connection the “directory structure“ is to be evaluate, too.

Evaluation of HTML Tables

All already analysed webpages of different companies use HTML tables. It is no matter if we have information pages or lead pages. For instance some information pages misuse tables in order to arrange information on the webpage (instead of frames). However during the utilization and evaluation of tables we have to consider, that for instance table items are in a specific relationship to other table items. Especially foundries very often use tables in order to correlate the described information items. Thus the tables are to differ from format (layout) and semantic.
In planned future work developments HTML tables are to be transformed into a representation with XML tags in order to make possible semantic relations explicit. This has also to be consider that tables can comprise tables again.
In our application tables with semantic relations contains very often values and numbers. For instance an assignment *weight* $\Rightarrow$ *10 till 50 kg* is possible. The example shows that ranges of numbers shall be evaluated, too. Especially the foundry industry uses various possibilities for representation of ranges of numbers (e.g. 10...50 kg, 10kg till 50 kg, 10-50 kg, till 50 kg, and so on).
Another possible semantic relation can be that several nouns (N) or noun phrases (NP) are assigned to one noun (e.g. industrial sector $\Rightarrow$ product). Thereby it must be considered that a list is possible in a table item, which is separated through <UL>-, <LI>-,
-tags or a comma.
Another important aspect is the analysis of graphics in tables. These graphics can be present on the on hand as pictures of company specific products or on the other hand as clues (tips) for navigation. Evaluation of HTML tables and the transformation into XML must consider these aspects, too. In [2] analyses for table recognition are carried out.

Evaluation of Phrasal Pattern

In addition to the structured information in a webpage also unstructured texts do exist. These texts must be evaluated in reference to their relevance for company profiles. For this we suggest the introduction of specific categories of phrasal pattern. Examples are in 2 and 4, which were extracted from information pages of various companies. The original examples were in german because we use for analyses a tool which is specified for german language.

Example 2 (category: industrial sector/products)

- *We produce* casings, bed plates, ...
- *Our products are* coupling disks, rings, ...
- *Manufacturer of* gate valve, lift table, ...
- *Our customers are* shipbuilding industry, foundry machine industry, ...
- *The important industrial sectors are* motor industry, wind power industry, ...

Here we differenciated into the categories “product related“ and “industry related“ depending on the underlying semantic. Very often a list of products or industry sectors separated with <UL>- or <LI>-tags follows the phrase. But also a list of nouns, or noun phrases separated through comma or without comma (with
-tag) is possible. A potential tagging in XML as basis for follow up processing is presented in Example 3 (also in german).

Example 3 (german phrase: *unsere Kunden sind* **(english phrase:** *our customers are***))**

```
<MS-CL V-ORDER="LAST" NP-ARG1="OBJ" RULE="MS-CL32">
    <NP CAS="_" NUM="_" GEN="_">
      <POSPRO>unsere</POSPRO>
      <N SRC="UC1">Kunden</N>
    </NP>
    <HV ROOT="sei" FLEX="FIN">sind</HV>
</MS-CL>
```

The XML tagging is created with the help of the tool “Document Suite“, which is used for a semantic and linguistic analysis of texts [5]. For instance our application uses the semantic relation *produce* <*company*> <*product*>. This semantic relation is expressible through a number of recurrent phrasal patterns (stereotypical linguistic realizations).

Example 4 (phrasal pattern - category: quality)

- We are certified for ...
- Certificates ...
- Factory certification ...
- We produce according to ...

These phrases are very often followed by lists, too. Here special expressions (terms) follow the phrases, for instance DIN EN ISO 9002, TDR 100, DIN EN 10204-2.3 and so on. Also names of industrial sectors may follow these phrases, for example the production depends on a special certificate (e.g. manufacturer qualification for German Railway, Lloyds Register of Shipping).

Ontology for Industrial Sectors

In our application we exploit several ontologies designed for different domains. In (Fig. 1) an ontology for possible industrial sectors is shown. Such data are necessary to evaluate the information of Example 3. For instance an analysis of the phrases contains an assignment to concepts. Here two versions are possible (Examples 5 and 6).

Example 5 (Ontology: IS-Part relation)

```
<CONCEPT TYPE="industry" IS-PART="Nutzfahrzeugbau">
    Schiffbau
</CONCEPT>
```

In Example 5 we see, that *Schiffbau* (shipbuilding industry) is part of *Nutzfahrzeugbau* (commercial vehicle industry). *Industry* is the concept type (type of ontology). A morphosyntactic tagger is used in order to get the correct word class (e.g. noun). With help of the IS-PART relation it is possible to decide that the phrase “Unsere Kunden sind der Schiffbau, ...“ (english: Our customers are shipbuilding industry, ...) provide information about a special topic. For instance we can use this information in order to create the aspect “customer“ in a company profile. In this example the data are classified as *industry related*.

Example 6 (Ontology: Assignment)

```
<CONCEPTS>
  <CONCEPT TYPE="industry">
    <WORD>Schiffbau</WORD>
    <DESC>branche</DESC>
  </CONCEPT>
</CONCEPTS>
```

If we need detailed information about the used concepts, our tool “Document Suite“ [4] gives also possible semantic relations between the concepts (Example 6). In this case more information can be derived about the term. That means, there is not only a concept type but also a description of the word (e.g. the classification of “Schiffbau“ as name of an industrial sector with <DESC> branche <DESC> in Example 6), which can be used to assign the term *Schiffbau* to the aspect “industrial sector“ in the main point “customers“ of the company profile.

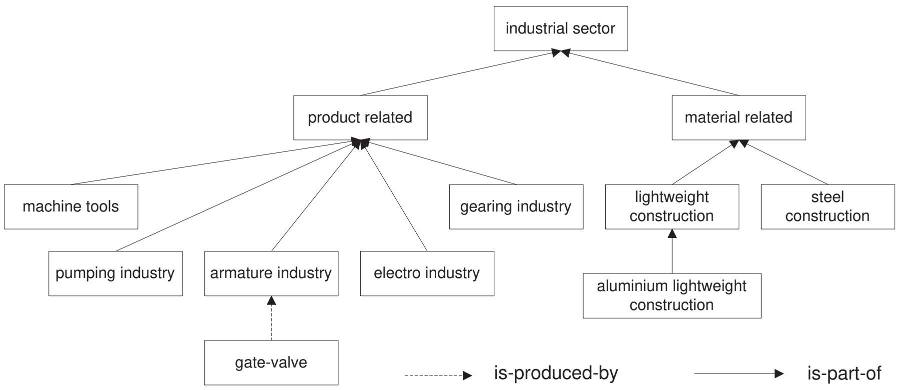

Figure 1: Ontology of industrial sectors

With the help of the XML tagged phrases a company profile can be created. The company profile is based on a defined XML structure, in order to form the basics for a use in connection with the semantic web.

Conclusion

In this paper we try to analyze the available information in the WWW depending on user specific requirements. We introduce our application field “casting technology“. But the results of our work are also to be extended to further application fields. Our research field comprehends both analyses of the semantic web and the use of XML technologies and analyses of knowledge akquisition.

References

[1] Y. Ackermann, C. Becker, and S. Danzmann. Erstellung eines Web-Dokumentagenten. *Abschlußbericht Softwarepraktikum, Fakultät für Informatik, Otto-von-Guericke-Universität Magdeburg*, 2001 (in german).

[2] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Experiments in Table Recognition. 1st International Workshop on Web Document Analysis, Seattle, USA, 2001.

[3] S. Krötzsch. *Angebotserarbeitung in Gießereien mit Hilfe einer strukturabhängigen Steuerung*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, 2001 (in german).

[4] M. Kunze and D. Rösner. Eine XML-basierte Werkbank für das Document mining. Sprach- und Texttechnologie im digitalen Medium; 12. Jahrestagung der GLDV, Giessen, Deutschland, 2001 (in german).

[5] M. Kunze and D. Rösner. XDOC-Extraktion, Repräsentation und Auswertung von Informationen. In GLDV-Workshop: Werkzeuge zur automatischen Analyse und Verarbeitung, Trier, Deutschland, 2001 (in german).

[6] G. Nagy and S. Seth. Twenty Questions for Document Classification. 1st International Workshop on Web Document Analysis, Seattle, USA, 2001.