Abstract: Automatic Metadata Analysis for Environmental Information Systems

Jens Hartmann
Center for Computing Technologies
University of Bremen
e-mail: jhart@tzi.de

Heiner Stuckenschmidt
AI Department
Vrije Universiteit Amsterdam
e-mail: heiner@cs.vu.nl

February 26, 2002

Introduction

Metadata plays an important role in web-based environmental information systems. It structures existing information and provides background information about technical issues as well as the context in which information has been generated or should be interpreted. Further, many systems such as the BUISY (Bremer Umweltinformationssystem) use metadata in order to provide content-based search facilities [VSV00]. Such methods, however, depend on the correctness and completeness of the metadata. Therefore, techniques for analyzing and verifying metadata are of great importance [Sv01]. In this paper, we discuss an approach for automatically analysing the metadata of web-based information systems that is based on machine learning techniques and its application to different environmental information systems.

The Approach

Based on a representation of HTML and XML documents as a *Document Object Model (DOM)* we use a *formal representation* to describe the logical structure of a document. We then identified a subset of structures occuring in web we want to analyse, in our case metadata structured that occur on web pages in terms of HTML meta tags. The structural decsriptions of web pages are devided into a Training and a Test Set. Using the Inductive Logic Programming System Progol [Mug95] we can now generate classifiers for differnet topic area in an environmental system. These classifiers consist of logical rules describing what all pages of a category have in common [SHv02]. For example we discoverd the following rule for the environmental information system of Bavaria that classifies pages from the area of waste management:

```
document(A) :- relation(A,B), metatag(B,keywords,abfall).
```

the rule states that all pages on this topic have a link to another pages that contains the word "abfall" in its keyword list. In the same way we can also analyze other properties of a web page that may contain information about its content. An example is the page title, where we were able to find characteristic word patterns for different types of pages. The example below shows the result for specific pages from the BUISY system.

```
document(A) :- doctitle(A,bodenschutz).
document(A) :- doctitle(A,boden).
document(A) :- doctitle(A,bo).
```

Rules like these can be used if no real metadata is available.

Experiments

We used the method to validate the metadata of three environmental information information systems:

- Bremen: `www.umwelt.bremen.de`
- Vienna: `www.ubavie.gv.at`
- Bavaria: `www.umweltministerium.bayern.de`

We topic areas that are normally used in order to structure information within these systems (waste management, soil protection, nature conservation, air- and water pollution and tried to sort pages into these topic areas based on their metadata. In general this should be possible with an accuracy of 100% because all three information systems contain content related metadata.

The results of the experiments shown in table 1 revealed that these systems expect for some parts of the bavarian systems have a good metadata infrastructure. We could classify almost all pags making use of the meta information provided by the systems.

Conclusions

The approach that will be presented in more detail in the full paper enables us to analyse and validate metadata of web-based environmental information system almost completely automatically. This does not allow us to handle large websites more efficiently, it also paves the path for advanced information processing based on metadata. On the other hand, the methods can be used by web masters in order to detect missing or incorrect metadata. The areas "Natur" and "Wasser" in the bavarian system aer a good example. The method clearly revealed the lack of appropriate metadata with respect to the content of the pages.

	Kategorie	correct	incorrect	*Accuracy.*
1.	Bremen - Abfall	39	0	100
2.	Bremen - Boden	33	0	100
3.	Bremen - Luft	84	0	100
4.	Bremen - Natur	62	1	98,41
5.	Bremen - Wasser	164	10	94,25
	Breman - Subtotal			98,53
1.	Vienna - Abfall	66	0	100
2.	Vienna - Boden	117	0	100
3.	Vienna - Luft	21	0	100
4.	Vienna - Natur	99	0	100
5.	Vienna - Wasser	360	0	100
	Vienna - Subtotal			100
1.	Bavaria - Abfall	18	0	100
2.	Bavaria - Boden	21	0	100
3.	Bavaria - Luft	14	1	93,33
4.	Bavaria - Natur	-	-	-
5.	Bavaria - Wasser	-	-	-
	Bavaria - Subtotal			58,67
	Total			85,73

Table 1: Classification Results

References

[Mug95] S. Muggleton. Inverse entailment and Progol. *New Generation Computing, Special issue on Inductive Logic Programming*, 13(3-4):245–286, 1995.

[SHv02] Heiner Stuckenschmidt, Jens Hartmann, and Frank van Harmelen. Learning structural classification rules for web-page categorization. In *Proceedings of FLAIRS 2002, special track on the Semantic Web*, 2002. to appear.

[Sv01] Heiner Stuckenschmidt and Frank van Harmelen. Knowledge-based validation, aggregation and visualization of metadata: Analyzing a web-based information system. In Y.Y. Yao, Ning Zhong, Jiming Liu, and Setsuo Oshuga, editors, *Web Intelligence 2001*, number 2198 in Lecture Notes in Artificial Intelligence. Springer, 2001.

[VSV00] Thomas Voegele, Heiner Stuckenschmidt, and Ubbo Visser. Buisy - using brokered data objects for environmental information systems. In Klaus Tochtermann and Wolf-Fritz Riekert, editors, *Hypermedia im Umweltschutz*. Metropolis, 2000.