Towards Semantic Web Mining

Bettina Berendt[1], Andreas Hotho[2], and Gerd Stumme[2]

[1] Institute of Information Systems, Humboldt University Berlin
Spandauer Str. 1, D–10178 Berlin, Germany
http://www.wiwi.hu-berlin.de/~berendt
berendt@wiwi.hu-berlin.de
[2] Institute of Applied Informatics and Formal Description Methods AIFB,
University of Karlsruhe, D–76128 Karlsruhe, Germany
http://www.aifb.uni-karlsruhe.de/WBS
{hotho, stumme}@aifb.uni-karlsruhe.de

Abstract. Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. The idea is to improve, on the one hand, the results of Web Mining by exploiting the new semantic structures in the Web; and to make use of Web Mining, on the other hand, for building up the Semantic Web. This paper gives an overview of where the two areas meet today, and sketches ways of how a closer integration could be profitable.

1 Introduction

Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. The idea is to improve the results of Web Mining by exploiting the new semantic structures in the Web. Furthermore, Web Mining can help to build the Semantic Web.

The aim of this paper is to give an overview of where the two areas meet today, and to sketch how a closer integration could be profitable. We will provide references to typical approaches. Most of them have not been developed explicitly to close the gap between the Semantic Web and Web Mining, but they fit naturally into this scheme. We do not attempt to mention all the relevant work, as this would surpass the paper, but will rather provide one or two examples out of each category.

In the next section, we start with a brief overview of the areas Semantic Web and Web Mining. The two areas can co-operate in various ways: First, Web mining techniques can be applied to help create the Semantic Web. A backbone of the Semantic Web are ontologies, which at present are often handcrafted. This is not a scalable solution for a wide-range application of Semantic Web techologies. The challenge is to learn ontologies, and/or instances of their concepts, in a (semi-)automatic way. A survey of these approaches is contained in Section 3.

Conversely, background knowledge — in the form of ontologies, or in other forms — can be used to improve the process and results of Web Mining. Existing techniques are investigated in Section 4.

I. Horrocks and J. Hendler (Eds.): ISWC 2002, LNCS 2342, pp. 264–278, 2002.

Recent developments have included the mining of sites that become more and more Semantic Web sites, and the development of mining languages that can tap the expressive power of Semantic Web knowledge representation. Section 5 discusses them and shows how they make the Semantic Web and Web Mining grow closer to each other.

In Section 6, we then sketch how the loop can be closed: from Web Mining to the Semantic Web and back. We believe that a tight integration of these aspects will greatly increase the understandability of the Web for machines, and will thus become the basis for the development of further generations of intelligent Web tools.

2 The Semantic Web and Web Mining

In the first part of this section, we briefly recall our understanding of the Semantic Web. In the second part, we give an overview of Web Mining approaches by classifying them into three categories: Web content mining, Web structure mining, and Web usage mining. In the remainder of the paper, we will then discuss how to bring together these different domains.

2.1 Semantic Web

The Semantic Web is based on a vision of Tim Berners-Lee. The great success of the current WWW leads to a new challenge: a huge amount of data is interpretable by humans only; machine support is limited. Berners-Lee suggests to enrich the Web by machine-processable information which supports the user in his tasks. For instance, today's search engines are already quite powerful, but still return too often too large or inadequate lists of hits. Machine-processable information can point the search engine to the relevant pages and can thus improve both precision and recall. To reach this goal the Semantic Web will be built up in different levels: Unicode/Unified Resource Identifiers, XML, RDF, ontologies, logic, proof, trust.[3]

The main focus of our research is on RDF, ontologies, and logic. We consider the content of the Semantic Web as being represented by ontologies and metadata. This approach is reflected by the Karlsruhe Ontology framework KAON[4] which is based on a formal definition of our understanding of what an ontology is [46]. It is built in a modular way, so that different needs can be fulfilled by combining parts.

This definition constitutes a core structure that is quite straightforward, well-agreed upon, and that may easily be mapped onto existing ontology representation languages. Step by step the definition can be extended by taking into account axioms, lexicons, and knowledge bases [46].

The inference engine behind our implementation relies on F-Logic [26], but there are many other approaches. A complete overview would be a paper on

[3] see http://www.w3.org/DesignIssues/Semantic.html
[4] http://kaon.semanticweb.org

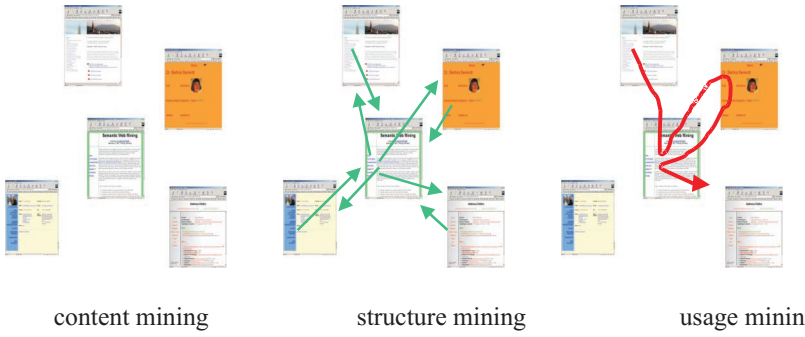

Fig. 1. The three areas of Web Mining.

its own. Hence we will only mention one, which is currently heavily discussed: DAML+OIL, a description logics formalism adapted to the Semantic Web.[5] We will not go into further detail here, but will rather discuss the topic of Web Mining and its relations to the Semantic Web in more depth.

2.2 Web Mining

Web mining is the application of data mining techniques to the content, structure, and usage of Web resources. This can help to discover global as well as local structure ("models" or "patterns" [19]) within and between Web pages. Like other data mining applications, Web mining can profit from given structure on data (as in database tables), but it can also be applied to semi-structured or unstructured data like free-form text. This means that Web mining is an invaluable help in the transformation from human understandable content to machine understandable semantics.

Three areas of Web mining are commonly distinguished: content mining, structure mining, and usage mining (see Fig. 1).

Content/text of Web pages. *Web content mining* is a form of text mining (for an overview, see [3]). The primary Web resource that is being mined is an individual page. Web content mining can take advantage of the semi-structured nature of Web page text. The HTML tags of today's Web pages, and even more so the XML markup of tomorrow's Web pages, bear information that concerns not only layout, but also logical structure.

Web content mining can be used to detect co-occurrences of terms in texts. For example, co-occurrences of terms in newswire articles may show that "gold" is frequently mentioned together with "copper" when articles concern Canada, but together with "silver" when articles concern the US. Trends over time may also be discovered, indicating a surge or decline in interest in certain topics such as the programming languages "Java". Another application area is event detection: the identification of stories in continuous news streams that correspond to

[5] http://www.daml.org

new or previously unidentified events (all examples from [7]). Further examples that allow the reconstruction of page content, and the discovery of relations in the domain under description, will be described in Section 6, where we will set them in relation to the Semantic Web.

Structure between Web pages. *Web structure mining* usually operates on the hyperlink structure of Web pages. The primary Web resource that is being mined is a set of pages, ranging from a single Web site to the Web as a whole. Web structure mining exploits the additional information that is (often implicitly) contained in the structure of *hyper*text. Therefore, an important application area is the identification of the relative relevance of different pages that appear equally pertinent when analyzed with respect to their content in isolation.

For example, hyperlink-induced topic search [27] analyzes hyperlink topology by discovering authoritative information sources for a broad search topic. This information is found in *authority* pages, which are defined in relation to hubs as their counterparts: *Hubs* are pages that link to many related authorities. The search engine Google, for instance, owes its success to the PageRank algorithm, which states that the relevance of a page increases with the number of hyperlinks to it from other pages, and in particular of other relevant pages [38].

Single pages too can be analyzed with respect to their structure, which gives information about their function, e. g., their function in the search for other pages. Cooley, Mobasher, and Srivastava [11] distinguish, based on [40], five types of Web pages: (i) "head" pages are entry points for a site, (ii) "navigation" pages contain many links and little information, (iii) "content" pages contain a small number of links and are visited for their content, (iv) "look-up" pages have many incoming links, few outgoing ones and no significant content, such as pages used to provide a definition or acronym expansion, and (v) "personal" pages have very diverse characteristics and no significant traffic.

Usage of Web pages. In *Web usage mining*, the primary Web resource that is being mined is a record of the requests made by visitors to a Web site, most often collected in a Web server log [43]. The content and structure of Web pages, and in particular those of one Web site, reflect the intentions of those who have authored and designed those pages, and their underlying information architecture. The actual behavior of those who use these resources may reveal additional structure.

First, relationships may be induced by usage where no particular structure was designed. For example, in an online catalog of products, there is usually either no inherent structure (different products viewed as a set), or one or several hierarchical structures given by product categories, manufacturers, etc. Mining the visits to that site, however, one may find that most (e. g., 80%) of those users who were interested in product A were also interested in product B. Here, "interest" may be measured by requests for product description pages, or the placement of that product into the shopping cart (indicated by the request for the respective pages). The identified association rules are at the center of cross-selling and up-selling strategies in E-commerce sites: When a new user shows

interest in product A, she will receive a recommendation for product B (cf. [34, 28]).

Second, relationships may be induced by usage where a different relationship was intended. For example, sequence mining may show that most of those users who visited page C later went to page D, along paths that indicated a prolonged search (frequent visits to help and index pages, frequent backtracking, etc.) [10, 25]. This can be interpreted to mean that visitors wish to reach D from C, but that this was not foreseen in the information architecture, hence that there is at present no hyperlink from C to D. This insight can be used for static site improvement for all users (adding a link from C to D), or for dynamic recommendations personalized for the subset of users who go to C ("you may wish to also look at D").

It is useful to combine Web usage mining with content and structure analysis in order to "make sense" of observed frequent paths and the pages on these paths. This can be done using a variety of methods. Some methods classify pages in terms of a pre-defined ontology, while others rely on the extraction of keywords found in these pages, and subsequent human naming of the keyword clusters represented by frequent paths. The ontology itself can be hand-crafted or (semi-)automatically learned, and the classification of pages in terms of the ontology can also be (semi-)automated in various ways.

In the following section, we will first look at how ontologies and their instances can be learned. We will then go on to investigate how the use of ontologies, and other ways of identifying the meaning of pages, can help to make Web Mining go semantic.

3 Extracting Semantics from the Web

The effort behind the Semantic Web is to add semantic annotation to Web documents in order to access knowledge instead of unstructured material, allowing knowledge to be managed in an automatic way. Web Mining can help to learn definitions of structures for knowledge organization (e. g., ontologies) and to provide the population of such knowledge structures.

All approaches discussed here are semi-automatic. They assist the knowledge engineer in extracting the semantics, but cannot completely replace her. In order to obtain high-quality results, one cannot replace the human in the loop, as there is always a lot of tacit knowledge involved in the modeling process. A computer will never be able to fully consider background knowledge, experience, or social conventions. If this were the case, the Semantic Web would be superfluous, since then machines like search engines or agents could operate directly on conventional Web pages. The overall aim of our research is thus not to replace the human, but rather to provide him with more and more support.

3.1 Ontology Learning

Extracting an ontology from the Web is a challenging task. One way is to engineer the ontology by hand, but this is quite an expensive way. In [32], the expression

Ontology Learning was coined for the semi-automatic extraction of semantics from the Web in order to create an ontology. There, machine learning techniques were used to improve the ontology engineering process. An example is given in Section 6.

Ontology learning exploits a lot of existing resources, like text, thesauri, dictionaries, databases and so on. It combines techniques of several research areas, e. g., from machine learning, information retrieval (cf. [29]), or agents [47], and applies them to discover the 'semantics' in the data and to make them explicit. The techniques produce intermediate results which must finally be integrated in one machine-understandable format, e. g., an ontology.

3.2 Mapping and Merging Ontologies

With the growing usage of ontologies, the problem of overlapping knowledge in a common domain occurs more often and becomes critical. Domain-specific ontologies are modeled by multiple authors in multiple settings. These ontologies lay the foundation for building new domain-specific ontologies in similar domains by assembling and extending multiple ontologies from repositories.

The process of *ontology merging* takes as input two (or more) source ontologies and returns a merged ontology based on the given source ontologies. Manual ontology merging using conventional editing tools without support is difficult, labor intensive and error prone. Therefore, several systems and frameworks for supporting the knowledge engineer in the ontology merging task have recently been proposed [24, 6, 36, 33]. The approaches rely on syntactic and semantic matching heuristics which are derived from the behavior of ontology engineers when confronted with the task of merging ontologies, i. e., human behavior is simulated. Another method is FCA-Merge which merges ontologies following a bottom-up approach, offering a global structural description of the process [44]. For the source ontologies, it extracts instances from a given set of domain-specific text documents by applying natural language processing techniques. Based on the extracted instances it uses the Titanic algorithm [45] to derive a concept lattice. The concept lattice provides a conceptual clustering of the concepts of the source ontologies. It is explored and interactively transformed to the merged ontology by the ontology engineer.

3.3 Instance Learning

It is probably reasonable to expect users to manually annotate new documents to a certain degree, but this does not solve the problem of old documents containing unstructured material. In any case we cannot expect everyone to manually mark up every produced mail or document, as this would be impossible. Moreover some users may need to extract and use different or additional information from the one provided by the creator. For the reasons mentioned above it is vital for the Semantic Web to produce automatic or semi-automatic methods for extracting information from Web-related documents, either for helping in annotating new documents or to extract additional information from existing unstructured or partially structured documents.

In this context, Information Extraction from texts (IE) is one of the most promising areas of Human Language Technologies. IE is a set of automatic methods for locating important facts in electronic documents for subsequent use, e. g. for annotating documents or for information storing for further use (such as populating an ontology with instances). IE as defined above is the perfect support for knowledge identification and extraction from Web documents as it can — for example — provide support in documents analysis either in an automatic way (unsupervised extraction of information) or semi-automatic way (e. g. as support for human annotators in locating relevant facts in documents, via information highlighting). One such system for IE is FASTUS (cf. [21]). Another is the OntoMat Annotizer [20], which also supports authoring. The approach of [12] is discussed in Section 6.

4 Exploiting Semantics for Web Mining

Semantics can be exploited for Web Mining for different purposes. The first major application area is Web content mining, i. e., the explicit encoding of semantics for mining the Web content.

4.1 Web Content Mining

In [22], we propose an approach for applying background knowledge in the form of ontologies during preprocessing in order to improve clustering results and allow for selection between results. We preprocess the input data (e. g. text) and apply ontology-based heuristics for feature selection and feature aggregation. Based on these representations, we compute multiple clustering results using k-Means. The results can be characterized and explained by the corresponding selection of concepts in the ontology.

In another current project, we are working on facilitating the customized access to courseware material which is stored in a peer to peer network[6] by means of conceptual clustering. We will make use of techniques of Formal Concept Analysis, which have been applied successfully in the Conceptual Email Manager CEM [9]. Based on an ontology, it generates a search hierarchy of concepts (clusters) with multiple search paths.

4.2 Web Structure Mining

Web structure mining can also be improved by taking content into account. The PageRank algorithm mentioned in Section 2.2 co-operates with a keyword analysis algorithm, but the two are independent of one another. So PageRank will consider any much-cited page as 'relevant', regardless of whether that page's content reflects the query. To improve search results, however, it is desirable to consider this content. By also taking the hyperlink anchor text and its surroundings into account, CLEVER [4] can more specifically assess the relevance

[6] http://edutella.jxta.org/

for a given query. The Focused Crawler [5] improves on this by integrating topical content into the link graph model, and by a more flexible way of crawling. Ontology-based focused crawling is proposed by [30].

4.3 Web Usage Mining

Exploiting the semantics of the pages visited along user paths can considerably improve the results of Web usage mining, since it helps the analyst understand what users were looking for, what content co-occurred, etc. The most basic form is again to use hand-crafted ontologies, in combination with automated schemes for classifying the large number of pages of a typical Web site according to an ontology of the site. For many current Web sites, this classification will be *ex post* and operate on pages that have been designed independently of an overall ontological schema (cf. [12]). However, a growing number of sites deliver pages that are generated dynamically in an interaction of an underlying database, information architecture, and query capabilities.

As an example, we have used an ontology to describe a Web site which operates on relational databases and also contains a number of static pages, together with an automated classification scheme that relies on mapping the query strings for dynamic page generation to concepts [2]. Pages are classified according to multiple concept hierarchies that reflect content (type of object that the page describes), structure (function of pages in object search), and service (type of search functionality chosen by the user). A path can then be regarded as a sequence of (more or less abstract) concepts in a concept hierarchy, allowing the analyst to identify strategies of search. This classification can make Web usage mining results more comprehensible and actionable for Web site redesign or personalization: The semantic analysis has helped to improve the design of search options in the site, and to identify behavioral patterns that indicate whether a user is likely to successfully complete a search process, or whether he is likely to abandon the site [42]. The latter insights could be used to dynamically generate help messages for new users.

In [1], we extend this approach by using the ontology to semi-automatically generate interesting queries for usage mining, and to create meaningful visualizations of usage paths. The classification scheme can easily be generalized to a wide range of other sites, in particular if these also operate on one or several underlying relational databases.

The more structured the underlying model is, and the more pages in a site are generated exclusively based on it, the more closely pages correspond to well-defined ontological entities (e. g., [15]). And the smaller the gap between the model generating the pages and the model analysing requests for those pages, the better semantics can be exploited in Web usage mining. At this level, the distinction between the use of semantics of Web Mining (as described in this section) and the mining of the Semantic Web itself (as described in the next section) starts to blur. An outlook on semantic usage mining that also evaluates the query strings, but operates on pages generated from a full-blown ontology (a "knowledge portal" in the sense of [23]) will be given in the following section.

The approaches discussed so far associate pages with an ontology and thus make their semantics explicit. An alternative, recurring on the semantics of pages that are implicitly contained in their text, is the automatic extraction of content by keyword analysis using standard Information Retrieval techniques (e. g., TF.IDF). Usage paths can then be clustered according to common content. This may help the analyst understand what kind of information users were seeking along frequently travelled paths [8]. It may also be used to identify content that co-occurred frequently in user histories, and to generate recommendations on the basis of these co-occurrences. Using a common representation of feature vectors, [35] show how clustering can use and combine usage, content, and structure similarities.

Web usage mining that is semantic in this sense is not only helpful for an ex post understanding of the paths users took through a site, but can also be used to aid users on-line, e. g. to improve their queries in a search engine. [39] use a combination of IR techniques analyzing single pages, ontologies, and the mining of a user's previous search history to make recommendations for query improvement. The basic idea is to (a) offer terms that are shown in the hierarchy as related, and to (b) infer from terms that occurred frequently in previous search histories a relative weighting on the set of pages that are described only coarsely by the few terms of the initial current query.

5 Mining the Semantic Web

As the Semantic Web enhances the first generation of the WWW with formal semantics, it offers a good basis to enrich Web Mining: The types of (hyper)links are now described explicitly, allowing the knowledge engineer to gain deeper insights in Web structure mining; and the contents of the pages come along with a formal semantics, allowing her to apply mining techniques which require more structured input. In the previous section, we have already seen that the distinction between the exploitation of semantics for 'standard' Web Mining on one side and the mining of the Semantic Web on the other side is all but sharp. Anyway, in this section we study those approaches which belong more to the latter.

5.1 Semantic Web Content and Structure Mining

In the Semantic Web, content and structure are strongly intertwined. Therefore, the distinction between content and structure mining vanishes. However, the distribution of the semantic annotations may provide additional implicit knowledge. We discuss now first steps towards semantic Web content/structure mining.

An important group of techniques which can easily be adapted to semantic Web content/structure mining are the approaches discussed as *Relational Data Mining* (formerly called *Inductive Logic Programming* (*ILP*); see [14] for an introductory collection of articles). Relational Data Mining looks for patterns that involve multiple relations in a relational database. It comprises techniques for

classification, regression, clustering, and association analysis. It is quite straightforward to transform the algorithms so that they are able to deal with data described in RDF or by ontologies. There are two big scientific challenges in this attempt. The first is the size of the data to be processed (i. e., the scalability of the algorithms), and the second is the fact that the data are distributed over the Semantic Web, as there is no central database server. Scalability has always been a major concern for ILP algorithms. With the expected growth of the Semantic Web, this problem increases as well. Therefore, the performance of the mining algorithms has to be improved, e. g. by sampling (see for instance [41]). As for the problem of distributed data, it is a challenging research topic to develop algorithms which can perform the mining in a distributed manner, so that only (intermediate) results have to be transmitted, and not whole datasets.

5.2 Semantic Web Usage Mining

Usage mining can also be enhanced further if the semantics are contained explicitly in the pages by referring to concepts of an ontology. Semantic Web usage mining can for instance be performed on log files which register the user behavior in terms of an ontology. A system for creating such semantic log files from a knowledge portal [23] has been developed at the AIFB [37]. These log files can then be mined, for instance to cluster users with similar interests in order to provide personalized views on the ontology.

6 Closing the Loop

In the previous three sections, we have analyzed how to establish Semantic Web data by data mining, how to exploit formal semantics for Web Mining, and how to mine the Semantic Web. In this section, we sketch one out of many possible combinations of these approaches. We will first *learn an ontology* using Web Mining, then *fill the ontology* with instances by again using Web Mining, and finally *mine the resulting data* in order to gain further insights. We will only give a rough sketch in order to illustrate our ideas. The example is taken from the Getess project[7] which provides ontology-based access to tourism Web pages in Mecklenburg-Vorpommern[8], a region in north-eastern Germany.

One may split the first step, *ontology learning*, in two sub-steps. First a concept hierarchy is established using the knowledge acquisition method OntEx (Ontology Exploration, [17]). It relies on the knowledge acquisition technique of Attribute Exploration [16] as developed in the mathematical framework of Formal Concept Analysis [18]; and guarantees that the knowledge engineer considers all relevant combinations of concepts while establishing the subsumption hierarchy. OntEx takes as input a set of concepts, and provides as output a hierarchy on them. This output is then the input to the second sub-step, together with a set of Web pages. [31] describes how association rules are mined

[7] http://www.getess.de/index_en.html
[8] http://www.all-in-all.de/

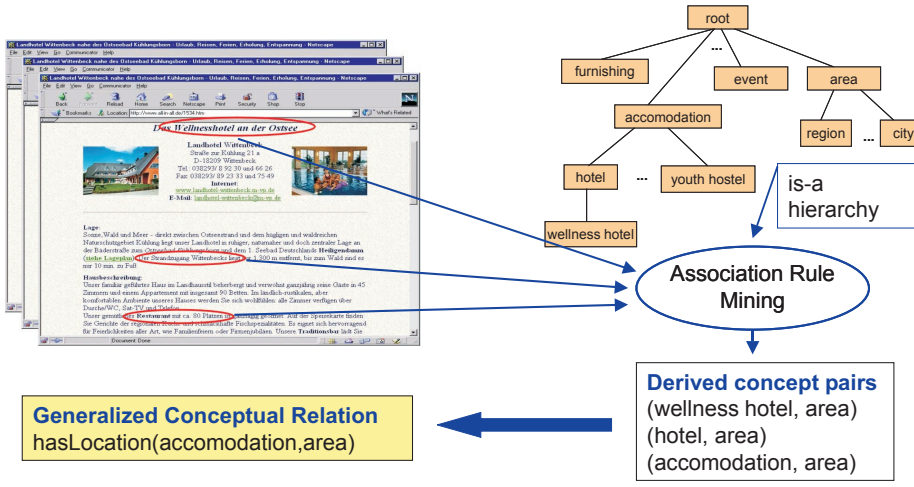

Fig. 2. Step 1: Mining the Web for learning ontologies.

from this input, which lead to the generation of relations between the ontology concepts (see Fig. 2). The association rules are used to discover combinations of concepts which frequently occur together. These combinations hint at the existence of conceptual relations. They are suggested to the user. As the system is not able to derive automatically names for the relations, the user is asked to provide them.

In the example shown in the figure, automatic analysis has shown that three concepts frequently co-occur with the concept "area". Since the ontology bears the information that the concept "wellness hotel" is a subconcept of the concept "hotel", which in turn is a subconcept of "accommodation", the inference engine can derive that only one conceptual relation needs to be inferred based on these co-occurrences: the one between "accommodation" and "area". Human input is then needed to identify that an accommodation "hasLocation" that is an area, i. e., to specify a name for the generalized conceptual relation.

In the second step, *the ontology is filled*. In this step, instances are extracted from the Web pages, and the relations from the ontology are established between them using techniques described in [12] (see Fig. 3), or any other technique described in Section 3.3. Beside the ontology, the approach needs tagged training data as input. Given this input, the system learns to extract instances and relations from other Web pages and from hyperlinks.

In the example shown in the figure, the relation "belongsTo" between the concepts "golf course" and "hotel" is instantiated by the pair (SeaView, Wellnesshotel), i. e., by the fact derived from the available Web pages that the golf course named "SeaView" belongs to the Wellness Hotel.

After the second step, we have an ontology and a knowledge base, i. e., instances of the ontology concepts and relations between them. These data are now input to the third step, in which *the knowledge base is mined*. Depending on the purpose, different techniques may be applied. One can for instance derive

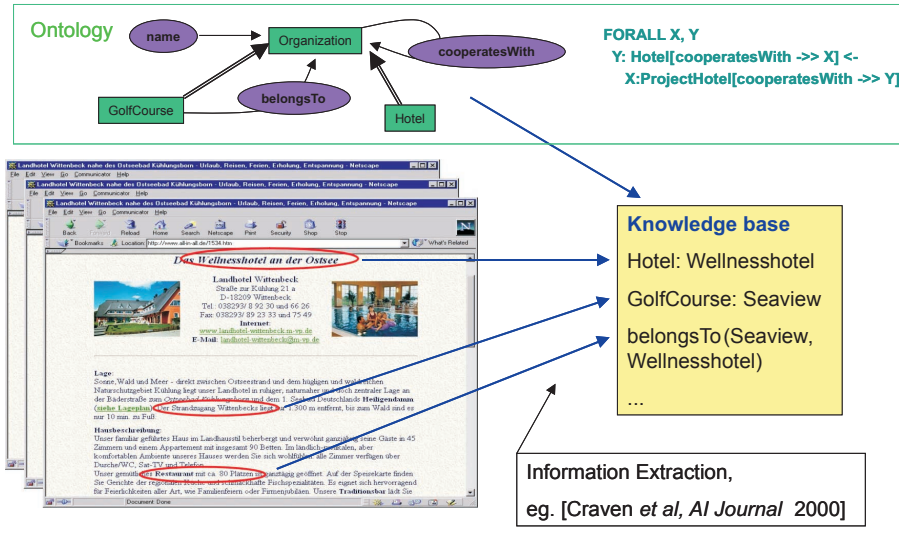

Fig. 3. Step 2: Mining the web for filling the ontology.

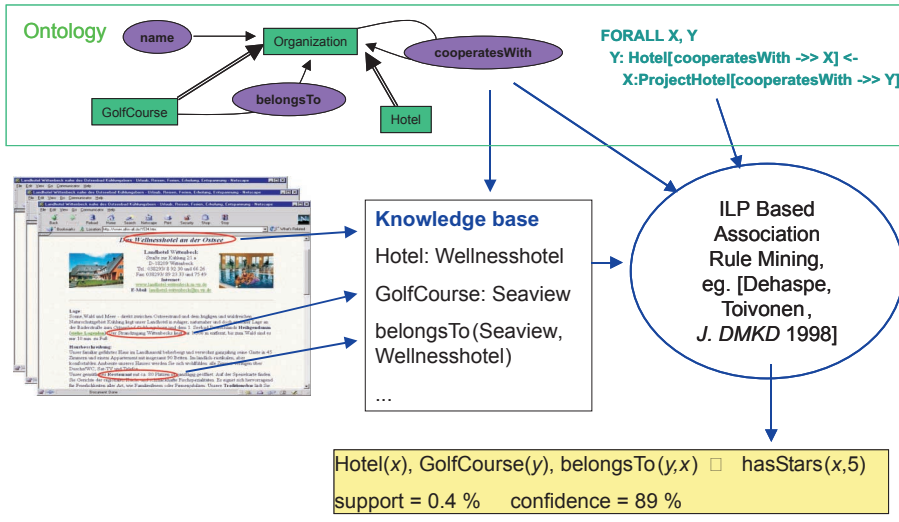

Fig. 4. Step 3: Using the ontology for mining again.

relational association rules, as described in detail in [13] (see Figure 4). Another possibility is to conceptually cluster the instances, e. g. using [45].

In the example shown in Figure 4, a combination of knowledge about instances like the Wellnesshotel and its SeaView golf course, with other knowledge derived from the Web pages' texts, produces the rule that hotels with golf courses often have 5 stars. More precisely, this holds for 89 % of hotels with golf courses, and 0.4 % of all hotels in the knowledge base are five star hotels owning a golf course.

The results of the last step may lead to further modifications of the ontology and/or knowledge base. When new information is gained, it may be used as input to the first steps in the next turn of the ontology life cycle.

7 Conclusion

In this paper, we have studied the combination of the two fast-developing research areas Semantic Web and Web Mining. We discussed how Semantic Web Mining can improve the results of Web Mining by exploiting the new semantic structures in the web; and how the construction of the Semantic Web can make use of Web Mining techniques. The example provided in the last section shows the potential benefits of further research in this integration attempt. The research questions arising from this interplay are likely to stimulate further research both in the Semantic Web as also in Web Mining.

References

1. B. Berendt. Using site semantics to analyze, visualize and support navigation. *Data Mining and Knowledge Discovery*, 6:37–59, 2002.
2. B. Berendt and M. Spiliopoulou. Analysing navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal*, 9(1):56–75, 2000.
3. S. Chakrabarti. Data mining for hypertext: A tutorial survey. *SIGKDD Explorations*, 1:1–11, 2000.
4. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th World-wide web conference (WWW7),30(1-7)*, pages 65–74, 1998.
5. S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings of the 8th World-wide web conference (WWW8),31(11-16)*, pages 1623–1640, Toronto, May 1999.
6. Hans Chalupsky. Ontomorph: A translation system for symbolic knowledge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000)*, pages 471–482, 2000.
7. G. Chang, M.J. Healey, J.A.M. McHugh, and J.T.L. Wang. *Mining the World Wide Web. An Information Search Approach.* Boston: Kluwer Academic Publishers, 2001.
8. E.H. Chi, P. Pirolli, and J. Pitkow. The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*, pages 161–168, Amsterdam: ACM Press., 2000.
9. Richard Cole and Gerd Stumme. Cem - a conceptual email manager. In Bernhard Ganter and Guy W. Mineau, editors, *Proc. ICCS 2000*, volume 1867 of *LNAI*, pages 438–452. Springer, 2000.
10. R. Cooley. *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data.* PhD thesis, University of Minnesota, Faculty of the Graduate School, 2000.
11. R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems*, 1(1):5–32, 1999.

12. M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, 118(1-2):69–113, 2000.
13. L. Dehaspe and H. Toivonen. Discovery of frequent datalog patterns. *Data Mining and Knowledge Discovery*, 3(1):7–36, 1999.
14. Saso Dzeroski and Nada Lavrac, editors. *Relational Data Mining*. Springer, 2001.
15. M. Fernández, D. Fiorescu, A. Levi, and D. Sucin. Declarative specification of web sites with strudel. *The VLDB Journal*, 9:38–55, 2000.
16. B. Ganter. Attribute exploration with background knowledge. *TCS*, 217(2):215–233, 1999.
17. B. Ganter and G. Stumme. Creation and merging of ontology top-levels. In *Proc. ECAI02*. submitted, 2002.
18. B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin – Heidelberg, 1999.
19. D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. Cambridge, MA: MIT Press, 2001.
20. Siegfried Handschuh and Steffen Staab. Authoring and annotation of web pages in CREAM. In *Proc. Of WWW11*. to appear, 2002.
21. Jerry Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. Fastus: A cascaded finite-state transducer for extracting information from natural-language text. In E. Roche and Y. Schabes, editors, *Finite State Devices for Natural Language Processing*. MIT Press, Cambridge MA, 1996.
22. A. Hotho, A. Maedche, and S. Staab. Ontology-based text clustering. In *Proceedings of the IJCAI-2001 Workshop "Text Learning: Beyond Supervision", August, Seattle, USA*, 2001.
23. A. Hotho, A. Maedche, S. Staab, and R. Studer. SEAL-II — the soft spot between richly structured and unstructured knowledge. *Journal of Universal Computer Science (J.UCS)*, 7(7):566–590, 2001.
24. E.H. Hovy. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proc. 1st Intl. Conf. on Language Resources and Evaluation (LREC)*, Granada, 1998.
25. H. Kato, T. Nakayama, and Y. Yamane. Navigation analysis tool based on the correlation between contents distribution and access patterns. In *Working Notes of the Workshop on Web Mining for E-Commerce - Challenges and Opportunities (WebKDD 2000) at the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 95–104, Boston, MA, 2000.
26. Michael Kifer, Georg Lausen, and James Wu. Logical foundations of object-oriented and frame-based languages. *Journal of the ACM*, 42:741–843, 1995.
27. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
28. W. Lin, S.A. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6:83–105, 2002.
29. A. Maedche. *Ontology Learning for the Semantic Web*. Kluwer, 2002.
30. A. Maedche, M. Ehrig, S. Handschuh, L. Stojanovic, and R. Volz. Ontology-focused crawling of documents and relational metadata. In *Proceedings of the Eleventh International World Wide Web Conference WWW-2002*, Hawaii, 2002.
31. A. Maedche and S. Staab. Discovering conceptual relations from text. In *ECAI-2000 - European Conference on Artificial Intelligence. Proceedings of the 13th European Conference on Artificial Intelligence*, pages 321–325. IOS Press, Amsterdam, 2000.

32. A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72 –79, 2001.
33. D. McGuinness, R. Fikes, J. Rice, and S. Wilder. An environment for merging and testing large ontologies. In *In the Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000)*, pages 483–493, Breckenridge, Colorado, USA, 2000.
34. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.
35. B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu. Integrating web usage and content mining for more effective personalization. In *Proceedings of the International Conference on E-Commerce and Web Technologies (ECWeb2000)*, pages 165–176, Greenwich, UK, 2000.
36. N. Noy and M. Musen. Prompt: Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 450–455, Austin, Texas, 2000.
37. D. Oberle. *Semantic Community Web Portals - Personalization, Studienarbeit.* Universität Karlsruhe, 2000.
38. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.
39. S. Parent, B. Mobasher, and S. Lytinen. An adaptive agent for web exploration based of concept hierarchies. In *Proceedings of the 9th International Conference on Human Computer Interaction*, New Orleans, LA, 2001.
40. Ramana Rao Peter Pirolli, James Pitkow. Silk from a sow's ear: Extracting usable structures from the web. In *Proc. ACM Conf. Human Factors in Computing Systems, CHI*, pages 118 – 125, New York, NY, 1996. ACM Press.
41. Tobias Scheffer and Stefan Wrobel. A sequential sampling algorithm for a general class of utility criteria. In *Knowledge Discovery and Data Mining*, pages 330–334, 2000.
42. M. Spiliopoulou and C. Pohle. Data mining for measuring and improving the success of web sites. *Data Mining and Knowledge Discovery*, 5:85–14, 2001.
43. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and application of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.
44. G. Stumme and A. Maedche. FCA–Merge: Bottom-Up Merging of Ontologies. In *IJCAI-2001 – Proceedings of the 17th International Joint Conference on Artificial Intelligence, Seattle, USA, August, 1-6, 2001*, pages 225–234, San Francisco, 2001. Morgen Kaufmann.
45. G. Stumme, R. Taouil, Y. Bastide, N. Pasqier, and L. Lakhal. Computing iceberg concept lattices with titanic. *J. on Knowledge and Data Engineering (in print)*, 2002.
46. Gerd Stumme. Using ontologies and formal concept analysis for organizing business knowledge. In *Proc. Referenzmodellierung 2001 (in print)*, 2002.
47. A.B. Williams and C Tsatsoulis. An instance-based approach for identifying candidate ontology relations within a multi-agent system. In *Proceedings of the First Workshop on Ontology Learning OL'2000*, Berlin, Germany, 2000. Fourteenth European Conference on Artificial Intelligence.