



International Journal of Computing and Corporate Research

Specialized and Refereed Journal for
Research Scholars, Academicians, Engineers and Scientists



DEVELOPMENT AND ENRICHMENT OF RESEARCH COMMOTION USING WEB BASED RESEARCH SUPPORT SYSTEMS

SHEILINI JINDAL

Department of Computer Science & Engineering

Chitkara University

Rajpura, Punjab, India

GAURAV KUMAR

Department of Computer Applications

Chitkara University

Rajpura, Punjab, India

ABSTRACT



The aim of Research Support Systems (RSS) is to enhance, develop and support research, which is a major and important part of decision support systems (DSS) for scientists and researchers. Scientists are helped by Web based RSS (WRSS) in their research processes which is carried on the Web platform. WRSS are based on the assembling, integration, adaptation and continuing advancement of existing computer technology and information systems for the purpose of research in the field of computers and technology. A framework of WRSS is presented by focusing on research activities and its various phases, as well as the technology support needed. The emphasis is on the conceptual formulation of WRSS and extracting semantics from the web. Different systems are linked to various research activities, and a mass of support sub-systems is established. The results of WRSS may lead to new and viable research tools.

KEYWORDS

Data mining, Decision support system, Framework, Information retrieval support system, Research support system, Semantic web.

WEB BASED RESEARCH SUPPORT SYSTEMS

The World Wide Web provides a new medium for storing, presenting, gathering, sharing, processing, and using information. The impacts of the Web can be felt in most aspects of our life. The impacts are two fold: Web technology provides us with more opportunities in terms of information availability, accessibility, and flexibility. However, more challenges are in front of us. We have to find the right information and tools from largely available resources. We have to learn to use the existing tools that keep changing all the time. The study of WSS aims to take the opportunities of the Web, to meet the challenges of the Web, and to extend



International Journal of Computing and Corporate Research

Specialized and Refereed Journal for
Research Scholars, Academicians, Engineers and Scientists



the human physical limitations of information processing. We define WSS as a multidisciplinary research field that focuses on supporting human activities in specific domains based on computer science, information technology, and Web technology. One of the goals is to find out how applications and adaptations of existing methodologies on Web platforms benefit our decision making and other various activities. The following are some potential benefits of Web technology:

- The Web provides a distributed infrastructure for information processing.
- The Web delivers timely, secure information and tools with a user friendly interface.
- The Web has no time or geographic restrictions. Users can access systems at any time and any place.
- Users can control and retrieve results remotely and instantly.

A TWO DIMENSIONAL VIEW OF WSS

Application domain	Technology	
	<i>Computer technology</i>	<i>Web technology</i>
Decision making	DSS	WDSS
Business application	BSS	WBSS
Information retrieval	IRSS	WIRSS
Scientific research	RSS	WRSS
Teaching	TSS	WTSS
Knowledge management	KMSS	WKMSS



Data mining	DMSS	WDMSS
-------------	------	-------

TABLE 1: A Two dimensional view of WSS (*Source: An introduction to Web Based Support Systems, J.T.Yao, Department of Computer Science, University of Regina Regina, Saskatchewan, Canada S4S 0A2*)

THE ARCHITECTURE OF WEB-BASED SUPPORT SYSTEMS

Interface, functionality, and databases are some of the components that are needed to be considered when we design a system. The architecture of WSS can be viewed as a (thin) client/server structure. The users, including decision makers and information seekers, are clients on the top layer. They access the system with browsers via the Web and Internet. The interface that is designed on the server side will be presented on the client's side by browsers. The lower layers and components encapsulated by the oval dotted line are very similar to conventional computerized support systems. In other words, a Web-based support system can be viewed as a support system with the Web and Internet as the interface. The architecture is presented from a usage point of view and is logical but not physical.

International Journal of Computing and Corporate Research



Specialized and Refereed Journal for
Research Scholars, Academicians, Engineers and Scientists

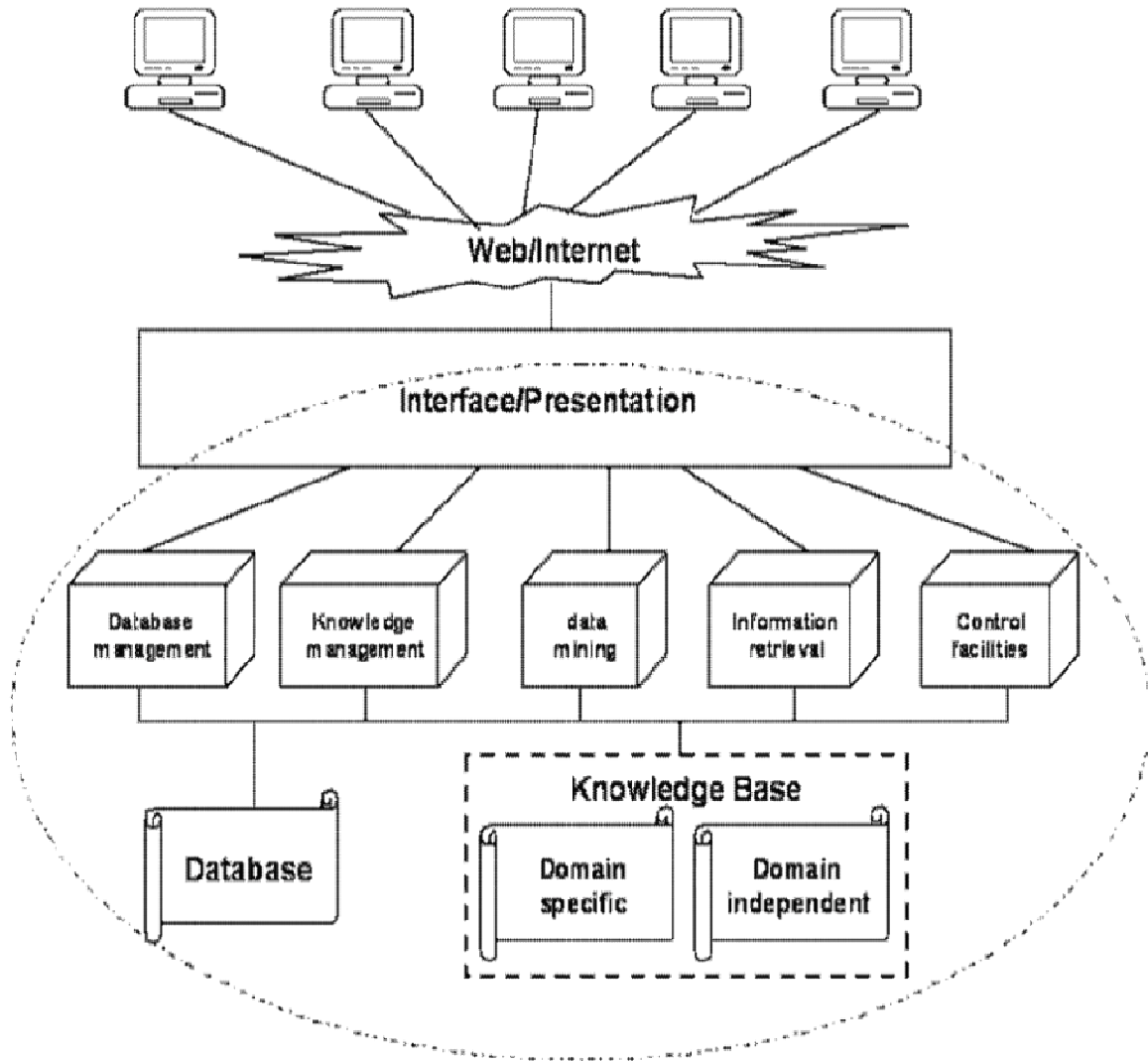


Figure 1: An architecture of web based research support system (Source: <http://citeseerx.ist.psu.edu>)



In practice, data and control components may not necessarily sit physically on the same point of the network, which is one of the major differences between WSS and traditional computerized support systems. System components may be spread all over the network. Users of the systems are located globally. Agent, grid computing, and Web services play important roles in WSS implementation. The data layer comprises two components. A database is a basic component in any modern system. WSS is not an exception. Another major component is the knowledge base. The knowledge base stores rules, principles, and guidelines used in supporting activities. We intend to divide the knowledge base into two parts: a domain-specific knowledge base and a domain independent knowledge base. The former is the knowledge specific to the domain that is supported. The latter involves general knowledge for all support systems. Knowledge management, data management, information retrieval, data mining, and other control facilities form the management layer. These serve as middleware for the three- tier client/server architecture and as the intermediaries between the interface and data layers. Reasoning, inference, and agent technologies play important roles on this layer. The separation between the management of data and user profiles results in a secure and standardized system. To take advantage of Web technology, these processes are distributed over the Internet to form a virtual server. In fact, databases and knowledge bases on the lower tier are also distributed. The WSS can be classified into three levels. The first level is support for personal activities. An example of such support is research support for individuals. Personal research activities such as search, retrieval, reading, and writing are supported. The second level is organizational support, such as research support on an institutional level. The top level is the network level. The collaborations between organizations or decision making by a group of people like in group decision support systems fall in this level. The group-decision support room may be a virtual room on the Web.



RESEARCH SUPPORT SYSTEM FRAMEWORK

In order to explore web data, we construct a research support system framework for web data mining, consisting of four phases: source identification, content selection, information retrieval and data mining. Different phases can be explained as follows:

In the first phase, proper web sites should be chosen according to research needs. This includes identifying availability, relevance and importance of web sites. Key words searching by using search engine can be used to find appropriate web sites. After finding all web sites identified by the first phase.

The second phase is to select appropriate contents on those web sites, such as documentation, newsgroups, forums, mailing lists, etc. Usually, a web site contains many web pages, including relevant and irrelevant information. This phase is important because it decides which web information should be extracted. The selection of web pages is based on research purpose and a researcher's experience. In the information retrieval phase, a crawler is designed to automatically extract information selected during the selection phase. Specific tools and techniques are employed to effectively retrieve useful knowledge/information from web sources. Additional effort may be required for dynamic content retrieval and specific data sources such as newsgroup, forum, etc.

The final phase is to conduct data mining on extracted web data. It includes preparing data for analysis. An extracted web page may contain missing data, extraneous data, wrong format and unnecessary characters. Furthermore, some



data should be processed in order to protect privacy. Advanced data mining techniques are employed here to help analyzing data.

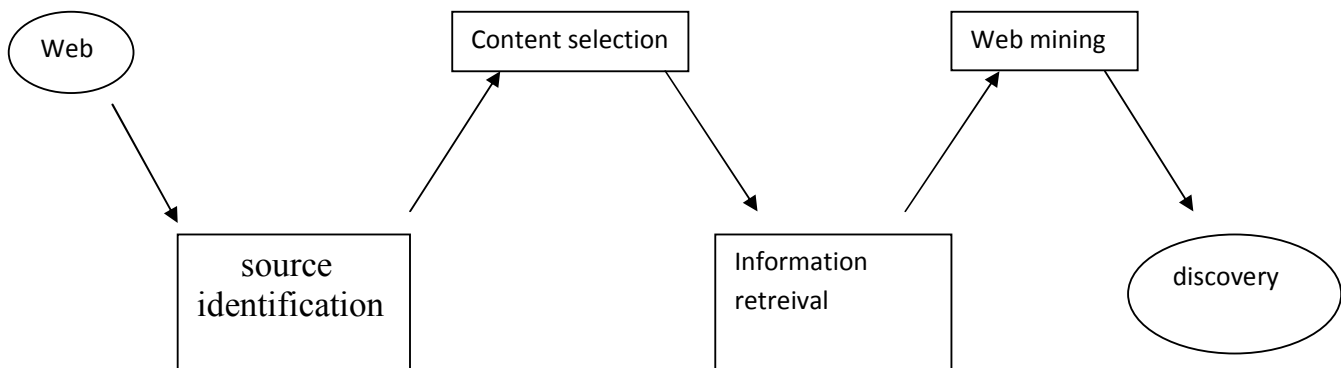


Figure 2: Research support system framework for web mining

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.

Ontology Learning

Extracting 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 the expression *Ontology Learning* was coined for the semi-automatic extraction of semantics from the Web in order to create ontology. There, machine learning techniques were used to improve the ontology engineering process. 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, or agents 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.

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. The approaches rely on syntactic and semantic matching heuristics which is derived from the behavior of ontology engineers when confronted with the task of merging ontologies, i. e., and 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. 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 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.

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. Another is the Onto Mat Annotizer, which also supports authoring.

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.

Web Content Mining

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 network⁶ 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.

Web Structure Mining

Web structure mining can also be improved by taking content into account. The PageRank algorithm 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 can more specifically assess the relevance for a given query. The Focused Crawler 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.

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. 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. 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. The latter insights could be used to dynamically generate help messages for new users. We extend this approach by using the ontology to semi-automatically 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. 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 and the mining of the Semantic Web itself 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 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 traveled paths . 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, 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. 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.

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.

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)*). 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. 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.

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 has been developed at the AIFB. These log files can then be mined, for instance to cluster users with similar interests in order to provide personalized views on the ontology.

CONCLUSION

This paper illustrates a framework for web mining research support system and describes its procedures. It then discusses implementing techniques on web data extraction and analysis. A sourceforge web mining case is presented as an example of how to apply this framework. This work is an exploratory study of web data retrieval and data mining on web data. We try to evaluate the data extraction process and data mining software which can be used to discover knowledge in the web data. The actual interesting discoveries are still in progress. We are expected to discover interesting patterns from the data.

REFERENCES

1. URL : <http://www.galeas.de/webmining.html>. Last Accessed : 22 November 2010
2. URL : <http://en.wikipedia.org/wiki/web-mining>. Last Accessed : 22 November 2010
3. Web Mining: Information and Pattern Discovery on the World Wide Web



Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, Department of Computer Science University of Minnesota, Minneapolis, MN USA, 2010

4. Design and Implementation of A Web Mining Research Support System, Research Proposal, University of Notre Dame, by Jin Xu, M.S., 25th November 2010

5. URL : <http://paginas.fe.up.pt/~ec/filesfi0405/slides/06%20WebMining.pdf> Last Accessed: 30 Nov. 2010

6. Web Mining: Accomplishments & Future Directions, Jaideep Srivastava

University of Minnesota, USA URL : <http://www.cs.umn.edu/faculty/srivasta.html>

7. Advanced AI Techniques for Web Mining, Ioan Dzitac. IT Department Agora University, Piata Tineretului, Oradea, Romania Ioana Moisil, Hermann Oberth, Faculty of Engineering - Computer Science and Automatic Control, Lucian Blaga, University of Sibiu

8. URL : <http://searchoracle.bitpipe.com/olist/Data-Mining.html> Last Accessed : December 2, 2010

9. URL : http://findarticles.com/p/articles/mi_m00OL/is_2_4/ai_99824637/ Last Accessed : December 2, 2010

10. URL : <http://maya.cs.depaul.edu/~mobasher/webminer/survey/node1.html> Last Accessed : December 2, 2010

11. URL : <http://www.mineit.com> Last Accessed : December 2, 2010

12. URL : <http://www.information-management.com/news/5458-1.html> Last Accessed : December 2, 2010



13. URL:

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.7261&rep=rep1&type=pdf> Last Accessed : December 2, 2010

14. URL: http://www2.cs.uregina.ca/~jtyao/Conf/WSS_WIC_Summerschool.pdf Last Accessed : December 2, 2010

15. URL: <http://www2.cs.uregina.ca/~wss/wss03/03/wss03-37.pdf> Last Accessed : December 2, 2010

16. A. Maedche and S. Staab. Ontology learning for the semantic web. IEEE Intelligent Systems, 16(2):72 –79, 2001.

17. 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.

18. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. Communications of the ACM, 43(8):142–151, 2000.

19. 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.

20. 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.



21. D. Oberle. Semantic Community Web Portals -Personalization, Studienarbeit. Universit"at Karlsruhe, 2000.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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. Morgan Kaufmann.



29. G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing iceberg concept lattices with titanic. J. on Knowledge and Data Engineering (in print), 2002.

30. Gerd Stumme. Using ontologies and formal concept analysis for organizing business knowledge. In Proc. Referenzmodellierung 2001 (in print), 2002.

31. 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.