Preface

1 Data Mining for the Web

The web has revolutionized our conception of communication and interaction. It offers new ways of business-to-business and business-to-customer transactions, new mechanisms for person-to-person communication, new means of discovering and obtaining information, services and products electronically. The volume of web data increases daily and so does its usage. Both *web contents* and *web usage data* are potential bearers of precious knowledge.

Some years ago, Oren Etzioni questioned whether the web should be observed as a quagmire or a gold mine [Etz96]. Indeed, the web grows freely, is not subject to discipline and contains information whose quality can be excellent, dubious, unacceptable or simply unknown. However, a carefully designed vehicle for the analysis of data can discover gold in the web [Etz96]. More recently, Richard Hackathorn proposed a methodology called ”web farming” for acquiring business-related information from the web, maintaining it and turning it into useful and actionable knowledge [Hac99].

After the advent of data mining and its successful application on conventional data, web-related information has been an appropriate and emerging target of knowledge discovery. Depending on whether the data used in the knowledge discovery process concerns the web itself in terms of *content* or the *usage of this content*, we can distinguish between “web content mining” and “web usage mining” [CMS99]. However, the two areas overlap, as is shown in Fig. 1.

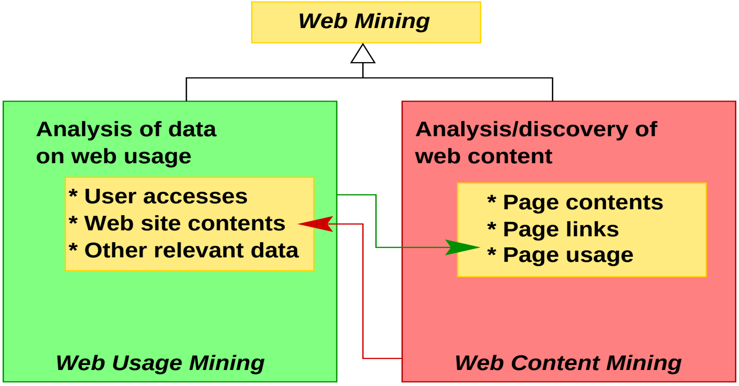

Fig. 1. Data mining for the web

Web content mining concentrates on discovering useful information in the web and on analyzing, categorizing and classifying documents. For document analysis, not only the document contents are taken into account, but also the

M. Spiliopoulou and B. Masand (Eds.): WEBKDD’99, LNAI 1836, pp. 1–6, 2000.

links connecting the web pages and potentially reflecting semantic relationships among them. Moreover, the access to the pages reflect how people conceive and interrelate them.

In contrast, web usage mining focusses on the discovery of knowledge about the people using the web, their interests and expectations, the problems they face and the implicit requirements they carry with them. Knowledge about the users forms the basis for dynamically adjustable sites, navigation assistants, customizable search engines, recommender systems and any other personalized services that seek to optimize the interaction between the person and the web. This knowledge is acquired by combining information on the page accesses, as recorded by the site servers, with information on page contents and with external data, such as user demographics, warehouse data on company customers etc.

2 Web Usage Mining at KDD'99

The 1999 ACM/SIGKDD International Conference of Knowledge Discovery and Data Mining hosted its first workshop concentrating on the challenges posed by mining the usage data from the web. WEBKDD'99, the Workshop on Web Usage Analysis and User Profiling brought together the communities of mining researchers, tool vendors and web usage data holders. The establishment of user profiles from anonymous web usage data, the extraction of knowledge from the navigation behaviour of the users and the assessment of the *usefuleness* of this knowledge were the main axes of investigation identified during the workshop. Open issues concerning data quality, scalability and privacy were vividly discussed during the panel session.

A report on the activities of the WEBKDD'99 workshop can be found in [MS00]. Briefly, WEBKDD'99 had 23 submissions, from which 10 were accepted for presentation after refereeing by 3 reviewers. The accepted workshop contributions, in the form of long abstracts, are in the on-line archive of ACM under
http://www.acm.org/sigkdd/proceedings/webkdd99/.

3 Web Usage Mining in this Volume

The collection of papers in this volume was established as follow-up of WEBKDD'99. The workshop contributions were revised and expanded to incorporate some of the open issues brought forward during the workshop, especially in the panel session. The following provides a tour of the different contributions organized roughly according to the workshop sessions.

3.1 User Modelling

The assessment of user profiles is an issue of major interest for web applications. User modelling is being investigated for a long time in different contexts and for a variety of domains, ranging from intelligent tutoring systems to recommendation agents in e-commerce. As merchants and service providers shift from mass products/services to customized offerings, the demand for knowledge about the individual interests and needs of each potential customer becomes paramount. The preferences of and further information on web site visitors can be obtained either by requesting input from the users or by drawing conclusions based on the users' observed behaviour.

The second approach fits ideally to the domain of data mining and is the subject of the first three papers in this chapter. Murray and Durrel investigate the issue of inferring demographic attributes, like gender and age, for users anonymously accessing a web site. The data on their activities are preprocessed and summarized using the the Latent Semantic Analysis vector space model and then classified by a neural network. Pre-classified data from demographic surveys are used to train the network, so that the demographics of the anonymous users can be derived with a certain degree of accuracy.

Fu, Sandhu and Shih investigate the discovery of user groups characterized by similar access patterns. They apply attribute-oriented induction to generalize from the concrete page accesses into concepts describing the page contents. Then, user groups are formed by hierarchical clustering on the concepts appearing in the sessions.

While these two studies focus on the acquisition of user profile data by the site provider, Philip Chan considers user profiling in the context of building a personalized browser of web documents. His adaptive personalized web browser forwards the user's queries to multiple search engines. It analyzes the result pages actually read by the user and ranks them according to an "interestingness" measure. From this ranking, themes interesting the user can be assessed; they form her profile. When the user issues a new query, the browser applies this derived profile to rank the results.

The gathering of personalized information and the establishment and exploitation of user profiles by the site providers is raising privacy concerns. The juxtaposition of privacy and data mining in the web is the subject of Alan Broder's invited paper. HTTP services do not preserve identity and can thus blurr the mining results on web usage. Hence, web site providers apply more advanced techniques, such as smart cookies, to acquire reliable data on the site visitors. At the same time, anonymization services emerge to better shield the user against unwelcome knowledge acquisition. So, the user can exploit modern technology to protect her privacy, albeit both engagement and effort are required.

3.2 Extraction of Knowledge in the Form of Rules and Patterns

The papers in this chapter discuss data mining algorithms that can extract rules or more general patterns from web usage data. Researchers in web usage analysis mostly take one of two approaches. In the first approach, data are analyzed using general-purpose algorithms, mainly for discovery of association rules, sequence mining and clustering, whereby the research concentrates on appropriate data modelling and preparation and on measures to evaluate the results for a particular problem setting. The second approach contributes to web usage mining with algorithms dedicated to cope with the particularities of knowledge discovery in the web.

Baumgarten et al adhere to the second approach. They propose M*i*DAS, a sequence miner for the discovery of navigation patterns. Taking web marketing as an example application domain, they investigate how the background knowledge of a company can drive the process of knowledge discovery. This knowledge is reflected in the establishment of the web site topology, in the construction of concept hierarchies over the data being mined and in the formulation of query templates that guide the mining process performed by M*i*DAS.

A mining algorithm for the discovery of navigation patterns is also proposed by Borges and Levene. Their approach is based on probabilistic grammars, whereby highly preferred trails are modelled as strings with high probability values. The theory of probabilistic grammars is combined with a new measure: entropy is used as an aggregate estimator of the statistical properties of each result grammar. Since the set of all grammar strings with probability above a threshold can be quite large, the authors investigate heuristics that compute a selected high-quality subset of the result set.

Lan, Bressan and Ooi adhere to the first approach of web pattern analysis. They use discovery of association rules for effective document pushing in a two tier (web server/browser) and a three tier (web server/proxy/browser) architecture. The conventional miner returns rules describing the support and confidence with which one document will be requested immediately after another one. The challenge lays then in devising measures and heuristics that select which document should be pushed among those suggested by a set of association rules. The authors propose a weighting scheme for the rules taken into account by the prefetching mechanism, derive heuristics based on this scheme and apply them on different two-tier and three-tier settings.

3.3 Determining the Interestingness of Knowledge Findings

The articles in the last chapter propose mechanisms for identifying *interesting* findings from the data. Lee et al investigate the domain of web merchandizing. They first distinguish among different areas of analysis in this domain. They then propose a set of metrics, called *micro-conversion rates*, with which they measure the effectiveness of merchandizing policies in real applications. For the analysis of merchandizing data with the new metrics, they use a grouping & visualization mechanism, with which they model data on promotion, offerings and on-line purchases of products across different dimensions and project them into a two-dimensional trajectory for display.

Spiliopoulou, Pohle and Faulstich address the problem of assessing the effectiveness of a web site in turning its users into customers. They generalize the concept of ”customer” from web marketing and study the contribution of each web page into the overall *contact and conversion efficiency* of the site, as reflected in the navigation patterns of its visitors. The navigation patterns of customers and non-customers are discovered by the web utilization miner WUM. Then, selected customer patterns are juxtaposed to comparable non-customer patterns, in order to identify differences and detect pages that are responsible for the different behaviour among customers and non-customers. Such pages should be redesigned, either statically for all users or on-the-fly by dynamic links. A mechanism assessing such links is suggested in the last part of the paper.

Cooley, Tan and Srivastava study the general problem of interestingness in the context of web usage mining. They consider a set of beliefs and propose a quantitative model based on *support logic* to determine the interestingness of a pattern. They assert that important domain information on the web is encapsulated in the content, structure and usage of the web pages, and exploit the last two factors to derive the evidence provided by a discovered pattern for or against a belief. Their quantitative model is embedded in the WebSIFT system, which is comprised of tools for data preprocessing, pattern discovery and pattern analysis. Pattern discovery is undertaken by conventional mining algorithms, while pattern analysis concentrates on identifying interesting patterns among the results of the miners.

4 Open Issues

Knowledge discovery using web data is an emerging research area with still many open issues. E-commerce is a motivating force in this area, since companies place high investments in the electronic market and seek to maximize their revenues and minimize their risks. Further applications, like tele-learning and tele-teaching, service support and information dissemination for the citizen, are also flourishing in the web. In all these application areas, there is a need for understanding and supporting the user, by means of recommender systems, dynamically adjustable information findings and personalized services. Knowledge about the user is indispensable for the design of such services.

In web usage mining, the discovery of patterns is not the sole issue to be addressed. The data acquisition itself is not a universally acceptable activity; it raises privacy concerns and active resistance from users and institutions. The acquired data are often noisy and must be preprocessed in special ways to preserve validity. Access logs can grow by gigabytes daily and become outdated soon, due to factors internal and external to the site. The web usage findings are also large in number, partly due to the large amount of quickly changing data. They require new innovative solutons to detect and summarize important findings. Hence, further new issues emerge, including: the validation of the reliability of the data, the scalability of the mining algorithms, the volatility of data and the short life expectation of patterns derived from them, the interpretation of the discovered patterns. Also, the issue of compatibility between data acquisition and privacy remains, asking for technical and political solutions.

Acknowledgements

We would like to express our gratefulness to the Program Committee members of WEBKDD'99 who invested their time and mental effort to the reviewing process for the preparation of this volume: Peter Brusilovsky (Human-Computer Interaction Institute, Carnegie Mellon Univ., USA), Bamshad Mobasher (DePaul University, USA), Christos Papatheodorou (NCSR Demokritos, Greece), John Roddick (Univ. of South Australia), Ramakrishna Srikant (IBM, Almaden Research Center, CA, USA) and Alex Tuzhilin (Stern School of Business, New York University).

We would like to thank all people that contributed to the WEBKDD'99 workshop, including the original Program Committee that performed the first selection and reviewing of papers and the workshop panelists who brought forward their visions on the future of the WEBKDD community and motivated many thoughts that found expression in this volume. We are grateful to the KDD99 organizing committee, especially Rakesh Agrawal (workshops chair) and Usama Fayyad, for helping us in bringing the WEBKDD community together. We would also like to thank the many participants for their active involvement in the discussions, their comments and their ideas on the evolution of the WEBKDD research area. They motivated us to publish this volume on research in this domain and to engage on the further work and prospering of the new area.

Myra Spiliopoulou Brij Masand

References

CMS99. Robert Cooley, Bamshad Mobasher, and Jaidep Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems*, 1(1), 1999. 1

Etz96. Oren Etzioni. The World-Wide Web: Quagmire or gold mine? *CACM*, 39(11):65–68, Nov. 1996. 1

Hac99. Richard D. Hackathorn. *Web Farming for the Data Warehouse*. Morgan Kaufmann Publishers, Inc., 1999. 1

MS00. Brij Masand and Myra Spiliopoulou. Webkdd'99: Workshop on web usage analysis and user profiling. *SIGKDD Explorations*, 2, 2000. to appear. 2