WEBKDD'99: Workshop on Web Usage Analysis and User Profiling

Brij Masand*
Redwood Investment Systems
brij@redwood.com

Myra Spiliopoulou
Institute of Information Systems
Humboldt-University Berlin
http://www.wiwi.hu-berlin.de/~myra

ABSTRACT

The WEBKDD'99 workshop on "Web Usage Analysis and User Profiling" took place at Aug. 15, 1999 under the auspices of the SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99). We report on the topics addressed in the workshop, the contributions and the discussions that took place in its framework.

Keywords

web usage mining

1. INTRODUCTION

Three years ago, Opher Etzioni posed the question of whether the web is a quagmire or a gold mine [5]. He showed that there are ways of obtaining gold, the gold of knowledge, out of web usage data — with appropriate effort. A recent business week article highlighted the commercial interest in this topic [1]. Mining the web for knowledge presents new and unprecedented challenges for the miners. The valuable fragments of information that can be turned into knowledge are incorporated into volumes of large scale web usage data which are often incomplete and error-prone. The desirable knowledge can include user demographics, correlations between web content, be it products or documents, but also patterns of user behavior that reflect the acceptability of and satisfaction with a web site. The discovery of this knowledge needs to not only address the challenges posed by the special nature of the data but as well the concerns for privacy, shared among many web users, and the legislatory and anonymization efforts undertaken in this direction.

The WEBKDD'99 workshop, organized in conjunction with the KDD99 ACM/SIGKDD conference in San Diego, brought together a community of practioners and researchers who face the challenges of mining web usage data. The extraction of user profiles from anonymous web usage data was the explicit subject of three contributions, and the motivating force for most of the others, too. The assessment of knowledge from the navigation behaviour of the users was addressed in five of the papers, while the discovery of association rules over web usage data was investigated in three studies. When searching for knowledge, we need methods of evaluating the usefulness of this knowledge and of concentrating on what is interesting and important; three of the contributions discussed this issue.

WEBKDD'99 had 23 contributions, from which 10 were accepted for presentation after refereeing by 3 reviewers. The accepted contributions were organized into 3 sessions, on user modelling, on discovery of association rules and navigation patterns and on interestingness measures. A panel discussion and the invited talk complemented the presentations by addressing open and emerging issues such as privacy and the practical challenges of developing and deploying applications.

The participation in WEBKDD'99 well exceeded our expectations. More than 100 KDD participants, comprised of roughly equal representation from industry and academia, gave rise to a lively workshop, through questions, initiating discussions and bringing forward open issues during the panel session.

2. THE INVITED TALK ON THE IBM SURFAID PROJECT

The opening event for webkdd99 was an invited talk by David C. Martin,(Global Business Intelligence Solutions of IBM). The subject of the talk was the SurfAid Project of IBM, focussing particularly on transactive analysis and prediction [10]. The talk also reflected the evolution of the nature of the project from the early concepts investigated in the first phase to the incorporation of the results in the IBM product/service palette in its third phase.

The first phase explored the utilization of data mining and machine learning to assist content providers and consumers in mutually beneficial transactions. The second phase concentrated on covering all activities of the users, including access, inspection and selection of a product, and retrieval of text, and on establishing a collection of users' demographics. The mining kernel was comprised of association rules' discovery, clustering and sparse predictive modelling. Particular emphasis was given to the analysis of text, in order to build topic hierarchies: for this, concepts were extracted from tags in web pages and combined with product correlations and with feedback from the customer companies. On this basis, the activities of users related to those topics were modelled as interaction patterns. Next to the analysis of interaction patterns for a particular timespot, the evolution of the patterns in a temporal vector space was made possible.

*Workshop organization while with GTE Labs
[1] http://www.businessweek.com/1999/99_30/b3639018.htm

The third phase focussed technically on the incorporation of the new services into a one-tier warehouse environment. The key observation in this context is that the internet is only *one* of the channels used for merchandizing. Companies with web presence need an integration of *all* their merchandizing channels and analysis of the whole integrated dataset.

3. SESSION 1: MODELLING USERS

The first session contained four papers investigating issues of user modelling in the web. Murray and Durrel addressed the problem of inferring demographic attributes, like gender and sex, for the users that anonymously access a web site [12]. To this purpose, the Latent Semantic Analysis (LSA) vector space model is used in conjunction with neural networks trained on data from survey responses.

Philip Chan investigated the problem of establishing user profiles for an adaptive personalized web browser [3]. The browser monitors the access behaviour of the user, gathers the pages retrieved by her and assigns an interestingness value on each visited page, according to a new interestingness measure. From the user's access behaviour and the contents of the visited pages, a user profile is built, describing the user's interests. When a user poses a query to the browser, the browser uses this profile to rank the results it obtains by forwarding the query to multiple search engines.

In [6], access patterns are used to build groups of web users. Fu, Sandhu and Shih organize the activities of each user in the web site into sessions and then apply attribute-oriented induction to generalize the sessions on the basis of a concept hierarchy on the pages' contents. Hierarchical clustering is applied on the generalized sessions to build groups of users that perform similar activities within a session.

Getoor and Sahami addressed the issue of inferring user preferences for recommendation and prediction purposes [7]. While traditional approaches infer user preferences from dyadic and crisp relationships between a user and a flat object, the proposed model permits probabilistic relationships manifested over objects with richer structure. By modelling uncertainty for a relationship and by allowing object properties to depend probabilistically on other properties of the same or related objects, powerful assessments on the interests of the users on an object can be inferred from their interests to other objects or object components.

4. SESSION 2: DISCOVERING ASSOCIATION RULES AND NAVIGATION PATTERNS

The three papers of the second session investigated the mining methodologies appropriate for knowledge discovery on web usage. Büchner et al proposed MiDAS, a new algorithm for navigation pattern discovery in the web [2]. They address the problem of web usage for commercial sites, where knowledge about the customers is a prerequisite for successful marketing. MiDAS is accompanied by a template-based language, on which the background knowledge of the analyst can be expressed to guide the mining process. Navigation patterns thus extracted can be further categorized into three different types of browsing behaviour.

The discovery of navigation patterns was also the subject of the work of Borges and Levene [1]. They use a probabilistic grammar to model user sessions, whereby higher probability strings express the trails preferred by the users. A new mining algorithm is proposed for the discovery of patterns thus modelled. A new measure, entropy, is proposed as an estimator of the statistical properties of the grammer.

Bin Lan et al proposed a model for the optimization of web server load based on association rule discovery. [9]. They apply association rules to discover correlations between documents and thus estimate the probability of documents being requested together. Then, various schemes for pushing documents that are likely to be requested next are explored for different server architectures and for different push lengths.

5. SESSION 3: INTERESTINGNESS MEASURES

The last session of WEBKDD'99 concentrated on methods for measuring the interestingness of the mining results. Gomory et al presented a new set of metrics for on-line merchandizing, called "micro-conversion rules" [8]. They showed how these metrics provide quantitative and qualitative insight into shopper behavior, by applying measures like look-to-click rate, click-to-basket rate, basket-to-buy rate etc. To assist the inspection of measurements with these metrics, a juxtaposed visualization of the results was proposed.

Measures for the effectiveness of a web site in e-commerce applications were also the subject of Spiliopoulou et al [13]. They propose a methodology of assessing the quality of a web site in turning its users into customers. Based on the template-based navigation pattern discovery mechanism of their miner WUM, they compare the navigation patterns of customers to those of non-customers and extract rules on how the site should be improved. They finally propose a technique for dynamically adapting the site according to those rules.

The interestingness of mining results against a set of beliefs was the topic of Cooley et al [4]. Their WebSIFT system for web data preparation and mining exploits content and structure information on a web site in order to identify potentially interesting web usage patterns. Based on available domain knowledge and the web site structure itself, a belief set is defined that helps identify patterns of web usage that are/are not expected to ocurr. Then, discovered frequent itemsets from web usage data are matched against the expected beliefs results into a filtered set of novel and 'interesting patterns' that are unexpected.

6. THE PANEL: OPEN ISSUES IN WEB USAGE MINING AND USER PROFILING

The panelists were Doug Beeferman (Lycos), Alan Broder (White Oak Technologies), Coyne Gibson (Knowledge Discovery One), Brij Masand (GTE Labs.) and Alex Tuzhilin (Stern School of Business, NY University). Myra Spiliopoulou (Humboldt-University Berlin) moderated the discussion.

Each panelist gave a short presentation on major issues that must be resolved or are expected to emerge in the near future. Alex Tuzhilin elaborated on the need for integrated solutions that exploit the demographics, location and activities of the users, consider their motivation and perceptions and also take the temporal dimension into account. Alan Broder addressed the issue of privacy, which is a concern for a lot of web users, and he discussed different degrees of anonymity that can be enforced by technology, ranging from instrumented browsers to anonymizers. Coyne Gibson also discussed ongoing anonymization efforts as a means for ensuring privacy, but observed that the deliberate usage of anonymizing technologies by web users is low and impacts only a fraction of traffic and log data. Brij Masand and Doug Beefermann addressed the issue of sessionizing, i.e. distinguishing among users and correctly attributing activities to them, as a basis for reliable analysis. Doug Beeferman juxtaposed the need for accurate sessionizing with the initiatives on anonymization. Brij Masand discussed the reliability of discovered findings from web logs. In the context of estimating the number of unique visitors, which turns out to be surprisingly non-trivial, errors may be as high as 50%-100%, depending on the mix of cookie and non-cookie based traffic. Thus one needs to exercise some caution when interpeting web statistics from logs and related mined knowledge.

After the presentations, an open discussion followed. The first discussion topic was data quality, which is a source of major difficulties for web usage mining, and particularly for the sessionizing problem. The solution of a dedicated server recording all activities of each user individually was put forward as a tested and successful solution. In the absence of such a server, cookies and scripts can be used to distinguish among unique users. However, they are not always feasible or popular, due to privacy considerations. In addition caching and proxy servers can make recontructing reliable sessions difficult.

The second critical issue was the increasing concern for privacy. As opposed to other channels of activity, in which privacy can be violated, the web was found particular for many reasons, including: (a) the absensce of laws restricting or regulating the usage of private data; (b) the absence of a single point of control, given the international nature of the web, and phenomena such as company mergers, which often make the privacy policies of the former companies obsolete; (c) the immediate actionability of the private data, of which the user becomes aware very soon; (d) the often annoying ways in which actionability becomes transparent and (e) the lack of control on the way private data will be exploited by the companies gathering them. Then the efforts of W3C in this regard were mentioned: the format of web data *and* the methods of mining them are going to be standardized, thus offering some transparency about web usage data capture and use.

Finally, the particularity of WEBKDD itself as an application area of data mining was discussed. Next to the aforementioned issues, which make data preprocessing considerably more difficult and critical for the success of the analysis, the volatility of the web was mentioned as a problem: data become outdated in each site redesign. Foremost, though, it is the very nature of the web that makes WEBKDD particular: In the web, there is a direct and immediate contact between author (page owner) and reader (user). This has no parallel to other paradigms like book publication and other forms of information dissemination. Hence, there is possibility *and demand* for immediate feedback and an unpreceded wide range for the design of personalization services.

7. CONCLUSIONS

The WEBKDD'99 workshop is over, but activity in this research area and in the framework it posed is still continuing. The full version of the contributions will appear in a book volume to be published by the Springer Verlag within the LNCS series. We expect the book to appear in March 2000. The short paper versions, as presented in the workshop, are available as part of the ACM archive at `http://www.acm.org/sigkdd/proceedings/webkdd99/`.

The intensive discussions, the many open issues, the involvement of the PC members and of the authors, and the encouraging words from all participants motivate us to pursue a continuation of the WEBKDD as a meeting point for scientists on web usage mining, and we look forward to an opportunity of establishing a WEBKDD workshops' tradition.

Acknowledgments

We would like to express our gratefulness to the Program Committee members Peter Brusilovsky (Human-Computer Interaction Institute, Carnegie Mellon Univ., USA), Ronny Kohavi (Blue Martini Software, USA), David C. Martin (IBM Global Business Intelligence,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), Steve Smith (Optas Inc., USA) and Alex Tuzhilin (Stern School of Business, New York University) for all the effort and time they devoted in ensuring the selection of the best quality contributions and to the success of this workshop.

We thank the panelists Doug Beeferman (Lycos), Alan Broder (White Oak Technologies), Coyne Gibson (Knowledge Discovery One), Brij Masand (GTE Labs.) and Alex Tuzhilin (Stern School of Business, NY University) for presenting their visions on the future of the WEBKDD community and initiating a lively and fruitful discussion. 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.

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 also acknowledge GTE Labs and the Humboldt-University Berlin for allowing Brij Masand and Myra Spiliopoulou to organize the workshop.

8. REFERENCES

[1] Jose Borges and Mark Levene. Data mining of user navigation patterns. In *[11]*, 1999.

[2] A. G. Büchner, M. Baumgarten, S. S. Anand, M. D. Mulvenna, and J. G. Hughes. Navigation pattern discovery from internet data. In *[11]*, 1999.

[3] Philip K. Chan. A non-invasive learning approach to building web user profiles. In *[11]*, 1999.

[4] Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. WebSIFT: The web site information filter system. In *[11]*, 1999.

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

[6] Yongjian Fu, Kanwalpreet Sandhu, and Ming-Yi Shih. Clustering of web users based on access patterns. In *[11]*, 1999.

[7] Lise Getoor and Mehran Sahami. Using probabilistic relational models for collaborative filtering. In *[11]*, 1999.

[8] Stephen Gomory, Robert Hoch, Juhnyoung Lee, Mark Podlaseck, and Edith Schonberg. Analysis and visualization of metrics for online merchandizing. In *[11]*, 1999.

[9] Bin Lan, Stephane Bressan, and Beng Chin Ooi. Making web servers pushier. In *[11]*, 1999.

[10] David Martin. IBM SurfAid project: Transactive analysis and prediction. Invited Talk in [11], 1999. see also http://surfaid.dfw.ibm.com/.

[11] Brij Masand and Myra Spiliopoulou, editors. *KDD'99 Workshop on Web Usage Analysis and User Profiling WEBKDD'99*, San Diego, CA, Aug. 1999. ACM. Online archive of the extended abstracts at http://www.acm.org/sigkdd/proceedings/webkdd99/. Long version of the contributions in preparation for the Springer Verlag LNCS series.

[12] Dan Murray and Kevan Durrel. Inferring demographic attributes of anonymous internet users. In *[11]*, 1999.

[13] Myra Spiliopoulou, Carsten Pohle, and Lukas C. Faulstich. Improving the effectiveness of a web site with web usage mining. In *book [11]*, 1999.