WEBKDD 2002 – Web Mining for Usage Patterns & Profiles

Brij M. Masand
Data Miners, Inc.
76 Summer Street
Boston, MA 02110-1225 USA
brij@redwood.com

Myra Spiliopoulou
Department of E-Business
Handelshochschule Leipzig (HHL)
Jahnallee 59, D-04109 Leipzig
myra@ebusiness.hhl.de

Jaideep Srivastava
University of Minnesota
4-192 EECS Bldg, 200 Union Street
Minneapolis, MN 55455
srivasta@cs.umn.edu

Osmar R. Zaiane
Department of Computing Science
University of Alberta
Edmonton, Alberta
Canada T6G 2E8
zaiane@cs.ualberta.ca

ABSTRACT

In this paper, we provide a summary of the WEBKDD 2002 workshop, whose theme was 'Web Mining for Usage Patterns and Profiles'. This workshop was held in conjunction with the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002).

Keywords

Web mining, profiling, personalization, click-stream analysis, usage patterns.

1. THEME

Though "E-commerce" may not be a popular word these days, Web usage has continued to grow, both in terms of the size of the user base, and the rate of usage. In addition, its low cost and flexibility is leading to increasing adoption by organizations as the preferred customer contact channel. On the user side, as usage experience and sophistication grows, there is an increasing demand for Web sites to be responsive to the needs of users. The truth of this statement is supported by the immense popularity of sites that offer personalization, e.g. Yahoo (www.yahoo.com) – which offers the MyYahoo service, Amazon (www.amazon.com) – which offers dynamic personalization, and Yodlee (www.yodlee.com) – which offers personalization of password protected sites.

Knowledge about the user is fundamental for the establishment of effective personalized services. *Web mining for Usage Patterns and User Profiles* is the application of web mining techniques to acquire this knowledge. Typical concerns in e-commerce include improved cross-sells, up-sells, personalized ads, targeted assortments, improved conversion rates, and measurements of the effectiveness of actions.

WEBKDD 2002 was the fourth in a series of very successful workshops on knowledge discovery from Web data. The strong interest for KDD in the Web has been manifested since the first WEBKDD workshop in 1999. The WEBKDD'02 workshop brought together practitioners on web-commerce, portals and application service providers (ASPs), decision-makers in non-commercial institutions that exploit web technologies to optimize their services, technology providers and data mining researchers to foster the exchange of ideas and the dissemination of emerging solutions on user and usage modeling for web-based applications.

In response to call for papers, WEBKDD 2002 received 23 contributions. We would like to thank the authors for their efforts, since it is their submissions that laid the foundations of a strong technical program. Each submission was reviewed by at least three program committee members. Ten submissions were selected for presentation. The main selection criterion was the quality of the idea. We would like to thank the members of the program committee for taking time to provide insightful critique, and thus ensuring the high quality of the workshop.

2. WORKSHOP

The KDD community responded very enthusiastically to the WEBKDD 2002 workshop, and about 50 people attended the workshop, which brought together e-commerce practitioners, tool vendors and data mining researchers. The paper presentation was divided into four sessions. Described below are the contents of each of the sessions.

2.1 Sessions 1: Categorization of Users and Usage

This session focused on how web mining can address one of the fundamental issues of Web usage, namely how to classify the user population into various categories, so that the experience can be made better. In addition, further classification based on 'usage patterns' can help gauging the 'mood' of the user. The two papers in this session presented new ideas in this direction.

In their paper titled *"Intelligent Discovery and Analysis of Web User Traffic Composition"*, Chi, Rosien, Heer show how Web Usage Mining enables new understanding of user goals on the Web. This understanding has broad applications, and traditional mining techniques such as association rules have been used in business applications. They have developed an automated method to directly infer the major groupings of user traffic on a Web site [Heer01]. They do this by utilizing multiple data features in a clustering analysis. They have performed an extensive, systematic evaluation of the proposed approach, and have discovered that certain clustering schemes can achieve categorization accuracies as high as 99%. This paper describes the further development of this work into a prototype service called LumberJack, a push-button analysis system that is both more automated and accurate than past systems.

Shah, Joshi, Wurman, in their paper titled *"Mining for Bidding Strategies on e-Bay"*, illustrated how data mining can be used to better understand auctions – a fast emerging approach to consumer e-commerce. Millions of people participate in online auctions on websites such as eBay. The data available in these public markets offer interesting opportunities to study Internet auctions. This paper explored techniques for identifying common bidding patterns on eBay using data from eBay videogame console auctions. The analysis reveals that there are certain bidding behaviors that appear frequently in the data, some of which have been previously identified and others which are new. The authors proposed new attributes of bidding engagements and rules for classifying strategies. In addition, they suggest economic motivations that might lead to the identified behaviors.

2.2 Session 2: Predictions and Recommendations - I

This session addressed the important issue of personalized recommendations. Personalization has clearly been one of the success stories of Web usability and user experience. The three papers in this session presented new approaches to this important problem.

In *"Categorization of web pages and user clustering with mixtures of hidden Marko models",* Ypma, and Heskes propose mixtures of hidden Markov models for modeling click streams of web surfers. Hence, the page categorization is learned from the data without the need for a (possibly cumbersome) manual categorization Theye provide an EM algorithm for training a mixture of HMMs and show that additional static user data can be incorporated easily to possibly enhance the labeling of users. Furthermore, they use prior knowledge to enhance generalization and avoid numerical problems. They use parameter tying to decrease the danger of over fitting and to reduce computational overhead. They put a flat prior on the parameters to deal with the problem that certain transitions between page categories occur very seldom or not at all, in order to ensure that a nonzero transition probability between these categories nonetheless remains. In applications to artificial data and real-world web logs we demonstrate the usefulness of our approach. They train a mixture of HMMs on artificial navigation patterns, and show that the correct model is being learned. Moreover, they show that the use of static 'satellite data' may enhance the labeling of shorter navigation patterns. When applying a mixture of HMMs to real-world web logs from a large Dutch commercial web site, they demonstrate that sensible page categorizations are being learned.

Hay, Wets, Vanhoof, in their paper *"Web Usage Mining by means of Multidimensional Sequence Alignment Methods",* present a new algorithm called Multidimensional Sequence Alignment Method (MDSAM) is illustrated for mining navigation patterns on a web site. MDSAM examines sequences composed of several information types, such as visited pages and visiting time spent on pages. Besides, MDSAM handles large databases and uses heuristics to compute a multidimensional cost based on one-dimensional optimal trajectories. Empirical results show that MDSAM identifies profiles showing visited pages, visiting time spent on pages and the order in which pages are visited on a web site.

In *"A Prediction Model for User Access Sequences"* Frias-Martinez, Karamcheti address one of the important Internet challenges in coming years, namely the introduction of intelligent services and a more personalized environment for users. Analysis of Web server logs has been used in recent years to model the behavior of web users in order to provide intelligent services. In this paper we propose a model for predicting sequences of user accesses that is distinguished by two elements: sequentiality and personalization. The concept of sequentiality in the model possesses three characteristics: (1) preservation of the sequence of the click stream in the antecedent, (2) preservation of the sequence of the click stream in the consequent and (3) a measure of the time gap between the antecedent and the consequent using the number of user clicks. In order to improve its prediction ratio, the model includes a personalization scheme in which each frequent user of a web site has a personal prediction system. The model has been defined as a black box that can be used as part of any intelligent service. As an example, they present a cache prefetching system based on the prediction model. The hit ratio of the cache is highly satisfactory.

2.3 Session 3: Predictions and Recommendation – II

The third session continued the theme of the second session, and presented four papers that describe new directions in some of the foundational issues of recommendation systems.

Bergholz, in his paper*, "Coping With Sparsity In A Recommender System"*, reports experiments on using an implementation of a recommender system called ``Knowledge Pump'' (KP) developed at Xerox. He repeats well-known methods such as the Pearson method, but also addresses common problems of recommender systems, in particular the sparsity problem. The sparsity problem is the problem of having too few ratings and hence too few correlations between users. He addresses this problem in two different manners. First, he introduces ``transitive correlations'', a mechanism to increase the number of correlations between existing users. Second, he adds ``agents'', artificial users that rate in accordance with some predefined preferences. He shows that both ideas pay off, albeit in different ways: Transitive correlations provide a small help for virtually no price, whereas rating agents improve the coverage of the system significantly but also have a negative impact on the system performance.

In *"On the use of constrained association rules for web mining",* Yang, Parthasarty and Reddy explore recommendation system further. They observe that in recent years there has been an increasing interest and a growing body of work in web usage mining as an underlying approach to capturing and modeling the behavior of users on the web for business intelligence and browser performance enhancements. Web usage mining strategies range from strategies such as clustering and collaborative filtering, to accurately modeling sequential pattern navigation. However many of these approaches suffer problems in terms of scalability and performance (especially online performance) due to the size and sparse nature of the data involved and the fact that many of the methods generate complex models that are less amenable to an online decision making environment. In this paper, first an approach is present, which is based on association rule mining. Their algorithm discovers association rules that are constrained (and ordered) temporally. The approach relies on the simple premise that pages accessed recently have a greater influence on pages that will be accessed in the near future. The approach not only results in better predictions, it also prunes the rule-space significantly producing only rules that matter, enabling faster online prediction. Further refinements based on sequential dominance are also evaluated, and prove to be quite effective. Detailed experimental evaluation shows how the approach is quite effective in capturing a web user's access patterns; consequently, our prediction model not only has good prediction accuracy, but also is more efficient in terms of space and time complexity. The approach is also likely to generalize for e-commerce recommendation systems.

Oyanagi, Kubota, Nakase, in *"Mining WWW Access Sequence by Matrix Clustering"* explore the issues in sequence mining for Web data. Sequence pattern mining is one of the most important methods for mining WWW access log. The Apriori algorithm is well known as a typical algorithm for sequence pattern mining. However, it suffers from inherent difficulties in finding long sequential patterns and in finding interesting patterns among a huge amount of results. This article proposes a new method for finding sequence patterns by matrix clustering. This method decomposes a sequence into a set of sequence elements, each of which corresponds to an ordered pair of items. Then matrix clustering is applied to extract a cluster of similar sequences. The resulting sequence elements are composed into a graph. The method is evaluated with practical WWW access log, which shows that it is superior to the conventional methods in finding long sequences and in generating a sequence graph from the resulting cluster.

2.4 Session 4: Evaluation of Algorithms

Now that we have a number of years' worth of experience with Web mining, it is time to evaluate the proposed algorithms and see how effective they are. The third session, titled *Evaluation of Algorithms*, presented two papers in this important emerging area. We hope to see more submissions in this area in the future.

In "*Evaluation of Recommender Algorithms for an Internet Information Broker based on Simple Association-Rules and on Repeat-Buying Theory",* Gayer-Schulz, Hahsler present a novel approach to evaluating recommendation algorithms. Association-rules are a widely used technique to generate recommendations in commercial and research recommender systems. Since more and more Web sites, especially of retailers, offer automatic recommender services using Web usage mining, evaluation of recommender algorithms becomes increasingly important. This paper compares the performance of a recommender algorithm based on repeat-buying theory known from marketing research with a recommender algorithm that uses association-rules. For the evaluation we concentrated on how well the the patterns extracted from usage data match the idea of "useful recommendations" of users. They used usage data of an educational Internet information broker as input for the recommender algorithms and asked users from the target group of the broker to classify a sample of recommendations with regard to their usefulness. In this paper the authors present and discuss the results of the evaluation of the two recomender algorithms using standard performance measures.

Finally, in *"The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis*", Berendt, Mobasher, Nakagawa, Spiliopoulou present a framework for evaluating session reconstruction algorithms – which is a fundamental issue for Web Usage Mining. The analysis of user behaviour in the Web presupposes a reliable reconstruction of the users' navigational activities. Cookies and Web-server-generated session identifiers have been designed to allow a faithful session reconstruction. However, in the absence of reliable methods, analysts must rely on heuristics methods (a) to identify unique visitors to a site, and (b) to distinguish among the activities of such users during independent sessions. The characteristics of the site, such as the site topology and structure, as well as the methods used for data collection (e.g., the existence of cookies and reliable synchronization across multiple servers) may necessitate the use of different types of heuristics. In this study, they extend their past work on the reliability of sessionizing mechanisms, by investigating the impact of site structure on the quality of constructed sessions. Specifically, they juxtapose sessionizing on a frame-based and a frame-free version of a site. We investigate the behavior of cookies, Web-server-generated session identification, and heuristics that exploit session duration, page stay time and page linkage. Different measures of session reconstruction quality, as well as experiments on the impact on the prediction of frequent entry and exit pages, show that different reconstruction heuristics can be recommended depending on the characteristics of the site. They also present first results on the impact of session reconstruction heuristics on prediction applications, which indicate a high quality of dynamic recommendations for personalization.

3. CONCLUSION

WEBKDD 2002 turned out to be a very successful workshop by all measures. More than __ attended it. The quality of papers was excellent, the discussion was lively, and a number of interesting directions of research were identified. This is a strong endorsement of the level of interest in this rapidly emerging field of inquiry.

4. REFERENCES

[1] Web site of the WebKDD 2002 Workshop, http://db.cs.ualberta.ca/webkdd02/.