Recommendation Models for User Accesses to Web Pages (Invited Paper)

Şule Gündüz[1] and M. Tamer Özsu[2]

[1] Department of Computer Science, Istanbul Technical University
Istanbul, Turkey, 34390
gunduz@cs.itu.edu.tr
[2] School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
tozsu@db.uwaterloo.ca

Abstract. Predicting the next request of a user as she visits Web pages has gained importance as Web-based activity increases. There are a number of different approaches to prediction. Markov models and their variations, collaborative filtering models, or models based on pattern recognition techniques such as sequence mining, association rule mining, clustering users or user sessions have been found well suited to this problem. In this paper we review various proposed techniques and also summarize two new models that we have developed. They consider the user access patterns to the pages as well as the time spent on these pages. We report experimental studies that show that the proposed methods can achieve a better accuracy than the other approaches.

1 Introduction

Web mining is defined as the use of data mining techniques to automatically discover and extract information from Web documents and services [9]. With the rapid growth of the World Wide Web, the study of modelling and predicting a user's access to a Web site has become more important. It has been used to improve the Web performance through caching and prefetching, to recommend related pages, improve search engines and personalize browsing in a Web site. Given a user's (who may, for example, be a customer in an e-commerce site) current actions, the goal is to determine which Web pages (items) will be accessed (bought) in the near future.

In general, Web mining is a common term for three knowledge discovery domains that are concerned with mining different parts of the Web: Web Content Mining, Web Structure Mining, and Web Usage Mining [25]. While Web content and structure mining utilize real or primary data on the Web, Web usage mining works on the secondary data such as Web server access logs, proxy server logs,

browser logs, user profiles, registration data, user sessions (or transactions)[1], cookies, user queries, and bookmark data. Web usage mining refers to the application of data mining techniques to discover usage patterns from these secondary data, in order to understand and better serve the needs of Web-based applications. The usage data collected at different sources will represent the navigation patterns of different segments of the overall Web traffic, ranging from single-user, single-site browsing behavior to multi-user, multi-site access patterns. The information provided by the data sources can all be used to construct/identify several data abstractions, such as users, server sessions, episodes, click stream, and page views [13].

In this paper, we survey the research in the area of recommendation (or recommender) systems. These systems collect ratings from web users explicitly or implicitly. Most of the models based on implicitly gathered information are based on data mining methods, which attempt to discover patterns from a variety of data sources. Web usage mining is one of these methods for building recommender systems. In this paper, we attempt to put the research done in a way from the Web usage mining point of view.

The rest of the paper is organized as follows. Section 2 briefly reviews the work related to Web mining and recommendation systems. Section 3 presents the proposed models. Finally, in Section 4 we conclude and discuss future work.

2 Recommendation Systems for Internet

It is often necessary to make choices without sufficient personal experience of the alternatives. In everyday life, we rely on recommendations from other people either by word of mouth, recommendation letters, movie and book reviews, or general surveys. Recommender systems assist and augment this natural social process. Since World Wide Web serves as a huge, widely distributed, global information service center for every kind of information such as news, advertisements, consumer information, financial management, education, government, e-commerce, health services, and many other information services, it becomes more important to find the useful information from these huge amounts of data. Recommender systems on Internet help people make decisions in this complex information space where the volume of information is available to them is very large. This section describes some approaches used in recommender systems.

2.1 Collaborative Filtering

One of the most successful and widely used technologies for building recommendation systems is collaborative filtering (CF). The term collaborative filtering is first introduced by Tapestry [10]. Collaborative filtering systems collect visitor opinions on a set of objects, using ratings provided by the users or implicitly

[1] The term *server session* is defined as the click stream of page views for a single visit of a user to a Web site [25]. In this paper we will use this term interchangeably with "user session" and "user transaction".

computed, to form peer groups and that establishes the basis of a learning system to predict a particular user's interest in an item. It is often based on matching, in real-time, the current user's profile against similar records (nearest neighbors) obtained by the system over time from other users. The ratings collected by the system may be both implicit and explicit. Explicit voting refers to a user consciously expressing his or her preference for a title, usually on a discrete numerical scale. Some example of systems that use this approach include SIFT [26], Tapestry [10] and the system described in [17]. The GroupLens project [12] is a purely collaborative filtering approach that automates prediction by collecting explicit user ratings and employing statistical techniques. The lack of explicit user ratings as well as the sparseness and the large volume of data pose limitations to standard collaborative filtering. As a result, it becomes hard to scale collaborative filtering techniques to a large number of items, while maintaining reasonable prediction performance and accuracy. A number of optimization strategies have been proposed and employed to remedy this shortcoming. These strategies include similarity indexing and dimensionality reduction to reduce real-time search costs.

2.2 Hybrid approaches

Content-based approaches have been proposed to address some of the limitations of collaborative methods. These systems work by comparing text descriptions or other representations associated with an item. A hybrid approach to recommendations combines aspects of both content-based and collaborative filtering. Balabonovic and Shoham [2] describe a system that helps users to discover new and interesting sites that are of interest to them. The system uses artificial intelligence techniques to present users with a number of documents that it thinks the user would find interesting. Users evaluate the documents and provide feedback for the system. From the feedback, the system knows more about the users' areas of interest in order to better serve the users in subsequent searches. The hybrid approach used in this model retains the advantages of content-based and collaborative approaches while overcoming their disadvantages.

The system described in [3] aims at offering innovative on-line services to support the trade fair business processes among a great number of exhibitors organized in a Web-based virtual fair. In order to build user profiles and provide recommendations, a method has been implemented which is based on the integration of data the system (explicitly and implicitly) collects about users and a classical collaborative filtering technique. The system then provides appropriate recommendations to the user in any circumstances during the visit.

WebWatcher [11] is an assistant agent that helps the user by using visual representations of links that guide the user reach a particular target page or goal. It learns by creating and maintaining a log file for each user and from the user feedback it improves its guidance.

2.3 Automated Recommender Systems

Implicit rating used for collaborative filtering can be divided into three categories: rating based on examination, when a user examines an item; rating based on retention, when a user saves an item; and rating based on reference, when a user links all or part of an item into an other item. Most of the techniques for implicitly gathering user information are based on data mining methods, which attempt to discover patterns or trends from a variety of sources. Web usage mining is one obvious and popular of these techniques. Recently, a number of approaches have been developed dealing with specific aspects of Web usage mining for the purpose of automatically discovering user profiles. For example, Perkowitz and Etzioni [20] proposed the idea of optimizing the structure of Web sites based on co-occurrence patterns of pages within the site's usage data.

Another recommendation system, Letizia [14], learns the areas that are of interest to a user by recording the users' browsing behavior. It performs some tasks at idle times (when a user is not reading a document and is not browsing). These tasks include looking for more documents that are related to the user's interest or might be relevant to future requests.

Schechter et al [22] have developed techniques for using path profiles of users to predict future HTTP requests, which can be used for network and proxy caching. Spiliopoulou et al [24], Cooley et al [5], and Mobasher [16] have applied data mining techniques to extract usage patterns from Web logs, for the purpose of deriving marketing intelligence. Shahabi et al [23], and Nasraoui et al [19] have proposed clustering of user sessions to predict future user behavior. Yan et al [27] use Web server logs to discover clusters of users having similar access patterns. The system proposed in [27] consists of an offline module that will perform cluster analysis and an online module which is responsible for dynamic link generation of Web pages. There have been attempts to use association rules [18], sequential patterns [1], and Markov models [6, 21] in recommender systems.

2.4 Methodology for Automated Recommender Systems

The overall process of automated recommendation can be divided into three components [25]. Since the data source is Web server log data for Web usage mining, the first step is to clean the data and prepare for mining the usage patterns. The second step is to extract usage patterns, and the third step is to build a predictive model based on the extracted usage patterns. Fundamental methods of data cleaning and preparation have been well studied [25]. The prediction step is the real-time processing of the model, which considers the active user session and makes recommendations. Once the mining tasks are accomplished, the discovered patterns are used by the online component of the model to provide dynamic recommendations to users based on their current navigational activity. The Web server keeps track of the active user session as the user browser makes HTTP requests. This can be accomplished by a variety of methods such as URL rewriting, or by temporarily caching the Web server access logs. The produced recommendation set is then added to the last requested page as a set of links before the page is sent to the client browser.

3 Web Usage Based Recommendation Models

An important feature of the user's navigation path is the time that a user spends on different pages [23]. The time spent on a page is a good measure of the user's interest in that page, providing an implicit rating for it. If a user is interested in the content of a page, he or she will likely spend more time there compared to other pages in his or her session. In this section, we present two new models that we have developed and that use the time spent on page visits. The first model (User Interest Model) uses only the visiting time and visiting frequencies of pages without considering the access order of page requests in user sessions. The resulting model has lower run-time computation and memory requirements, while providing predictions that are at least as precise as previous proposals. The second model (Click-stream tree model) uses both the sequences of visiting pages and the time spent on those pages. As far as we know, existing tools for mining two different information types like the order of visited Web pages and the time spent on those pages, are hard to find. Therefore, we concentrate in this study on a model that well reflects the structural information of a user session and handles two-dimensional information.

For the first step of the models, we use cleaning and filtering methods in order to identify unique users and user sessions. The cleaning step is the same for both of the proposed models. Since the cleaning procedure is beyond the scope of this paper, the details of this procedure are not given here. In this research, we use server logs from the NASA Kennedy Space Center server collected over the months of July and August 1995 [15]. Approximately 30% of these cleaned transactions are randomly selected as the test set, and the remaining part as the training set.

3.1 User Interest Model

The User Interest Model [7] clusters the transactions in the training set according to the similar amount of time spent on similar pages. We employ a model-based clustering algorithm to partition user sessions. The key idea behind this work is that user sessions can be clustered according to the similar amount of time that is spent on similar pages within a session. In particular, we model user sessions in log data as being generated in the following manner: (i) When a user arrives to the Web site, his or her current session is assigned to one of the clusters, (ii) the behavior of that user in this session, in terms of visiting time, is then generated from a Poisson model of visiting times of that cluster. The model parameters are learned with Expectation Maximization (EM) algorithm under the assumption that the data come from a mixture of Poisson distributions. To confirm our assumption that the data in each dimension (page of the Web site) have been generated by a Poisson distribution, the histogram of the occurrence of each of the time values at each dimension has been plotted. The histograms verify our assumption. For the last step, the transactions in the test set are assigned to one of the clusters that has the highest probability given the visiting time of current transaction's active page. The recommendation engine then predicts

three pages that have the highest recommendation scores in the active cluster. A hit is declared if any one of the three recommended pages is the next request of the user. The hit-ratio which is the number of hits divided by the total number of recommendations made by the system is 43% for the NASA data set.

3.2 Click-stream tree model

The click-stream tree model [8] considers page access sequences in addition to visiting times. The user sessions produced in the first step are clustered based on a similarity metric. Since user sessions are ordered URL requests, we can refer to them as sequences of Web pages. The similarity between sessions is then calculated using a dynamic programming approach such that only the identical match of page sequences and the time spent on those pages has a similarity value of 1. Using these pair-wise similarity values, a graph is constructed whose vertices are user sessions and edges are the calculated similarity values. In this study an efficient and fast graph partitioning algorithm called Cluto is used for graph partitioning [4]. We generate a click-stream-tree for each cluster. First a *root* node, which is labelled as "null", is generated for the click-stream-tree. When inserting a session in the tree, the first page of the session is stored as a child of the root node if the root node does not have a child with the same page and time information. If it has a child node with the same page and time information, the count of the corresponding node is incremented by one. The second page in the path is stored in a node that is a child of the first page's node. This may continue until the last page in the session. We start the algorithm with an empty tree of Web pages, which contains only the root node. The tree of a cluster is then constructed by inserting all the sessions in that cluster as mentioned above. The recommendation engine is the real time component of the model that selects the best path for predicting the next request of the active user session. When a request is received from an active user, a recommendation set consisting of three different pages that the user has not yet visited, is produced using the best matching user session[2]. For the first two requests of an active user session all clusters are explored to find the one that best matches the active user session. For the remaining requests, the best matching user session is found by exploring the top-N clusters that have the highest N similarity values computed using the first two requests of the active user session. The rest of the recommendations for the same active user session are made by using the top-N clusters.The resulting click-stream trees are then used for recommendation. The experimental results [8] indicate high precision in the recommendations made using this model.

4 Conclusion and Future Work

In this paper we provide a snapshot of recommendation methods by presenting the most representative techniques of collaborative filtering and data mining. We

[2] The user session that has the highest similarity to the active user session is defined as the best session.

have also considered the problem of modelling the behavior of a Web user during a single visit to the Web site. We proposed two models. The first model uses only the time information of the visiting pages whereas the second one uses both the time information and the access order of page requests. Our experimental results indicate that the techniques discussed here are promising, each with its own unique characteristics, and bear further investigation and development.

The field of recommender systems is still young and much work lays ahead. Given the huge amount of information available on the Internet and increasingly important role that the Web plays in today's society, data mining services on the Internet will become one of the most important and flourishing subfields in data mining. With the increasingly use of data mining tools on Internet, an important issue to face is privacy protection and information security. Evaluation of recommender systems is still a challenge. In the feature, there may be a development of new measures that go beyond those used for recommender systems to deal with more complex systems and that not only consider quantitative but sociological and economical factors as well.

Since recommender systems are becoming widely used by Web sites, a careful evaluation of their performance gets more important. However, for data mining part of recommender systems the question of how well found patterns match the user's concept of useful recommendation is often neglected.

References

1. R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. 11th Int. Conf. on Data Engineering (ICDE)*, pages 3–14, Taipei, Taiwan, March 1995.
2. M. Balabonovic and Y. Shoham. Learning information retrieval agents: Experiments with automated web browsing. In *Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogenous, Distributed Resources*, pages 13–18, 1995.
3. P. Buono, M. F. Costabile, S. Guida, A. Piccinno, and G. Tesoro. Integrating user data and collaborative filtering in a web recommendation system. In *Proc. 8th Int. Conf. on User Modeling*, Sonthofen, Germany, 2001.
4. Cluto. http://www-users.cs.umn.edu/ karypis/cluto/index.html.
5. 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.
6. M. Deshpande and G. Karypis. Selective markov models for predicting web-page accesses, 2001.
7. Ş. Gündüz and M. T. Özsu. A user interest model for web page navigation. In *Proc. Int. Workshop on Data Mining for Actionable Knowledge*, Seoul, Korea, April 2003. to appear.
8. Ş. Gündüz and M. T. Özsu. A web page prediction model based on click-stream tree. In *submitted for publication*, 2003.
9. O. Etzioni. The world wide web: Quagmire or gold mine. *Communications of the ACM*, 39(11):65–68, 1996.
10. D. Goldberg, D. Nichols, B. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

11. T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In *Proc. 15th Int. Conf. on Artificial Intelligence*, pages 770–777, Nagoya, Japan, 1997.
12. J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, 1997.
13. R. Kosala and H. Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations*, 2(1):1–15, 2000.
14. H. Lieberman. Letizia: An agent that assists web browsing. In Chris S. Mellish, editor, *Proc. of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 924–929, Montreal, Quebec, Canada, 1995. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.
15. NASA Kennedy Space Center Log. http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html.
16. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery of aggregate usage profiles for web personalization. In *Proc. International WEBKDD Workshop – Web Mining for E-Commerce: Challenges and Opportunities*, Boston, USA, 2000.
17. J. Mostafa, S. Mukhopadhyay, W. Lam, and M.Palakal. A multilevel approach to intelligent information filtering: Model, system and evaluation. *ACM Transactions on Information Systems*, 15(4):368–399, 1997.
18. A. Nanopoulos, D. Katsaros, and Y. Manolopoulos. Effective prediction of web-user accesses: a data mining approach. In *Proc. International WEBKDD Workshop – Mining Log Data Across All Customer TouchPoints*, 2001. San Francisco, CA, USA.
19. O. Nasraoui, H. Frigui, A. Joshi, and R. Krishnapuram. Mining web access logs using a fuzzy relational clustering algorithm based on a robust estimator. In *Proceedings of the Eighth International Fuzzy Systems Association Congress*, 1999. Hsinchu, Taiwan.
20. M. Perkowitz and O Etzioni. Adaptive web sites. *Communications of the ACM*, 43(8):152–158, 2000.
21. R. R. Sarukkai. Link prediction and path analysis using markov chains. In *Proc. 9th Int. World Wide Web Conference*, pages 377–386, 2000. Amsterdam.
22. S. Schechter, M. Krishnan, and M. D. Smith. Using path profiles to predict http requests. In *Proc. 7th Int. World Wide Web Conference*, Brisbane, Australia, November 1998.
23. C. Shahabi, A. Zarkesh, J. Adibi, and V. Shah. Knowledge discovery from users web-page navigation. In *Proc. 7th Int. Workshop on Research Issues in Data Engineering*, pages 20–29, April 1997. Birmingham, England.
24. M. Spiliopoulou and L. C. Faulstich. Wum: A tool for web utilization analysis. In *extended version of Proc. EDBT Workshop WebDB'98*, pages 184–203, Valencia, Spain, 1998. Springer-Verlag.
25. J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan. Web usage mining: Discovery and application of usage patterns from web data. *ACM SIGKDD Explorations*, 1(2):12–23, 2000.
26. T.W Yan and H.Garcia Molina. The SIFT information dissemnination system. *ACM Transactions on Database Systems*, 24(4):529–565, 1999.
27. Y. Yan, M. Jacobsen, Garcïa-Molina H, and U. Dayal. From user access patterns to dynamic hypertext linking. In *Proc. 5th Int. World Wide Web Conference*, pages 1007–1014, Paris, France, 1996.