Web Access Predictive Models

Jakub Snopek, Ivan Jelínek

Abstract: *Adaptation of web pages to needs of a specific user is today's trend of web technologies. The web adaptation and personalization problem, thus modification of web documents in sense of respecting needs and possibilities of a concrete user, could be solved by constructing and maintaining of a user predictive model, which provides predictions of the next user's step in his communication with the Web. In this paper we discuss current approaches used in this area of the research.*

Key words: *web access prediction, adaptive web, web personalization*

INTRODUCTION

Prediction of user's consecutive steps in his/her communication with the Web poses a big challenge for researchers in the web engineering area. If we are able to estimate next user's request with sufficient accuracy, based on this information we could modify behaviour of a web systems to accommodate user's needs and meet his expectations. Other motivation of researchers is growing concern of a web performance. Intelligent web caching can reduce response time, network bandwidth consumption and web latency. Further, response time reduction by loading the user's request before the actual access can be provided by web prefetching. Large area for predictive models usage is adaptive web. Adaptation of web-user interface to suit user needs, as well as web personalization can assist users to navigate and search web sites more effectively [1].

The Web can be seen as a structure containing information about hyperlinks, Web usage information, and Web contents in itself. Web site usage data, which contain records of how user has visited a Web site, can be used to identify collective user behavior in using the Web site, and use it as a base information for its predictive model.

Major sources of Web usage data on Web sites are web log files. A Web log file is a collection of records of user requests of documents on a Web server. Typical web log record contains following fields: IP address of the computer from which the request was made, User ID (Identification), date and time stamp of the request, URL of the requested document, status indicating whether the request was successful, document size, referring URL, and name and version of the browser and operating system being used for making the request. However, due to the influence of caching on the user side and proxy server, not all the requests are recorded in Web log files [3].

Web usage mining consists of web log preprocessing, request pattern discovery and pattern analysis. Data preprocessing as the first step of this process is time-consuming phase, requiring combining and cleaning logs, identificate and differentiating users and web robots/crawlers, group requests by sessions and more. There are several approaches to build predictive models from web usage data.

WEB MINING APPROACH

Web mining approach applies machine learning and data mining techniques to the Web for useful knowledge about the Web and its users. Web mining is typically concerned with characterizing the user, finding common attributes of classes of users, and predicting future actions without the concern for interactivity or immediate benefit.

Web mining approach uses traditional methods of machine learning, i.e. *k-nearest neighbor model* or naive *Bayes classifier model* [5] (Bayesian learning method, where attribute values are assumed to be conditionally independent) to predict next link. Other effective machine learning method are *artificial neural networks*, trained by incrementally adjusting the weights connecting the neurons to correctly categorize the training set. Widely the backpropagation algorithm is used.

Often used data mining method are *Decision trees.* Built on algorithms such as ID3, ASSISTANT or C4.5 [10] generate a set of *if-then-else* rules. These algorithms recursively construct a tree using a greedy algorithm by finding the next attribute that maximizes the separation of categories at each node.

Another class of machine learning algorithms are *Genetic algorithms.* There evolution-based algorihtms encode each rule set as a bit string and uses genetic search operators to explore this hypothesis space. It operates by iteratively updating a pool of rules, called population. On each generation, only some of rules are selected according to their fitness and are carried forward into the next generation population intact [3].

DYNAMICAL MODEL APPROACH

Dynamical prediction models are different from what data mining approaches do with Web logs. Predictive models based on Markov chains ([2], [4]) and n-grams are able to dynamically predict the next actions that the user will take. The n-gram models predict which URL will be requested next; the Markov models compute a probability of next request. The disadvantage of both n-gram and Markov method is that they cannot predict access to previously unvisited pages.

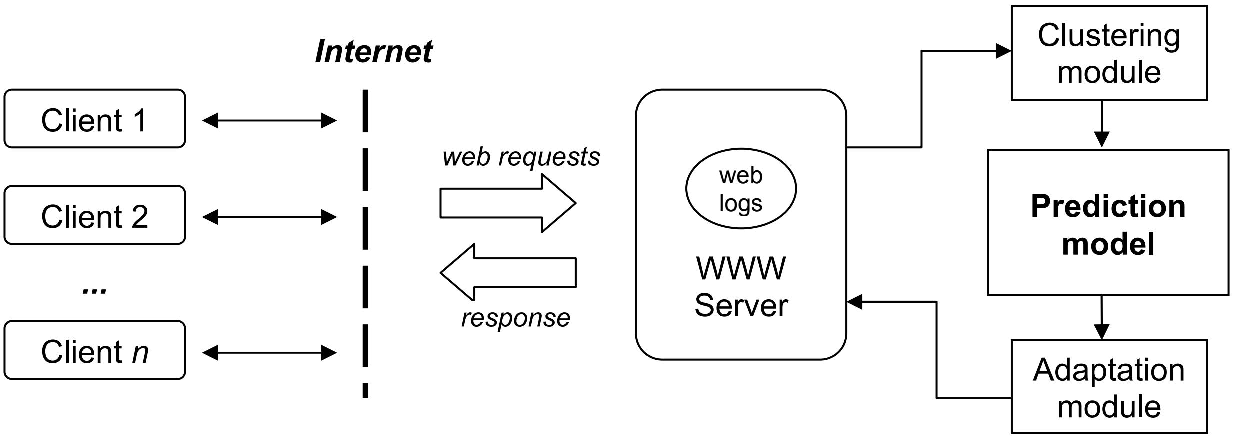

Figure 1. General architecture of web system with web access prediction

An *n-gram* is a sequence of *n* web requests, and the *n-gram* models learns how often each sequence of *n* requests was made in the training data. If an n-gram matches the suffix of the user's recent browsing trail, it is used to predict a future request. A disadvantage of the n-gram approach is that it requires large amount of training data and it just request next without estimating the probability of that next request [3].

The navigation process of a user on a Web site can be modeled as a *Markov chain*, i.e., the pages that the user is likely to request in the future are determined by the pages already requested by the user. Web pages can be treated as states, and hyperlinks between Web pages as transitions between the states in a Markov chain model [2]. Information about Web usage contained in a Web link structure can be used to infer the transition probabilities between these states. The Markov chain model can be used to predict the Web pages that a user is likely to visit given a sequence of Web pages already visited by the user. To predict the *n*-th step in the future, we need to compute the *n*-th power of the transition matrix [1], which can be computationally expensive given the large number of pages on a Web site [4]. This problem is solved by several algorithms:

- *Transition matrix compression* algorithm reduces the size of the state space of the Markov chain model by aggregating most similar transition behaviors into one state, while retaining the accuracy of prediction. The compressed transition matrix does not result in a significant increase of errors when being raised to a higher power [7].
- *Hybrid-order tree-like Markov* model is intelligently merging tree-like structure that aggregates the access sequences by pattern matching and a hybrid-order method that combines varying-order Markov model, can predict Web access precisely, providing high coverage and good scalability [11].
- Another possible way to improve predictive model performance is web page clustering. Algorithm *PageGather* creates a co-occurrence matrix of all pairs of pages visited and finds clusters of pages that is frequently viewed in the same session. Top *n* pages are recommended, that are most likely to co-occur with the visitor's current session [6]. Clustering by Web page contents used a vector to represent the contents of a Web page as a set of weighted terms [8]. The content similarity of two Web pages is defined as a distance-based measure between the two vectors representing the two pages respectively. A Web page is initially treated as a singleton cluster and used as an input to an *agglomerative hierarchical clustering* algorithm, in which clusters are combined sequentially based on the similarity measure between them.
- The agglomerative hierarchical clustering and k-means algorithms have been used in the Scatter/Gather system [9] to cluster documents on the basis of their contents for browsing large information spaces. Scatter/Gather presents summaries of clusters to users, who can then select some of these clusters for re-clustering the documents in them. New clusters become smaller and their contents are revealed in more detail. Important role in clustering web pages have web metrics, used to „measure" web pages content similarity and their properties in web hypertext structure (graph-based metrics) [12].

CONCLUSIONS AND FUTURE WORK

Promising approach among techniques for web access prediction are Markov chains, widely used to model user navigation on the Web. Difficulties concerning their usage arise in web structures with large number of heterogeneous pages. Current research goal is to find methods how to increase their prediction accuracy and usability for large web sites. Research of technologies stated above should result in model design, able to predict and offer the most relevant web documents from desired information area to web users. We address especially Markov models and their design enhanced by semantic web page clustering, focusing on web access prediction in large-structure web sites.

REFERENCES

[1] Zhu, J: Mining Web Site Link Structures for Adaptive Web Site Navigation and Search. Dissertation. Faculty of Informatics, University of Ulster at Jordanstown, October 2003.

[2] Zhu, J., Hong, J., Hughes, J. G.: Using Markov Chains for Link Prediction in Adaptive Web Sites. Springer-Verlag Berlin, Heidelberg 2002.

[3] Brian D. Davison. Learning Web request patterns. In A. Poulovassilis and M. Levene (eds), *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*, pp. 435-460, Springer 2004.

[4] Sarukkai, R. R.: Link prediction and path analysis using Markov chains. WWW, 2000.

[5] Mladenic, D. Machine leaning for better web browsing. In *AAAI Spring, Symposium on Adaptive User Interfaces*, 2000.

[6] Perkowitz, M., and Etzioni, O. Towards adaptive Web sites: Conceptual framework and case study. In *Artificial Intelligence* 118, 245-275, 2000.

[7] Spears, W. M. (1998) A compression algorithm for probability transition matrices. *SIAM Matrix Analysis and Applications*, Vol. 20, No. 1, pp. 60-77.

[8] Crouch, D. B., Crouch, C. J., and Andreas, G. (1989) The use of cluster hierarchies in hypertext information retrieval. In *Proc. of Hypertext'89,* pp. 225-237.

[9] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992) Scatter/Gather: A cluster based approach to browsing large document collections. In *Proc. of ACM SIGIR'92*, pp. 318-329.

[10] Smith, A. S. G.: Application of Machine Learning Algorithms in Adaptive Web-based Information Systems. Thesis. School of Computing Science Middlesex University. United Kingdom, 1999.

[11] Dongshan, X., Junyi, S.: A New Markov Model for Web Access Prediction. In *Computing in Science & Engineering*, Vol. 8, pp. 34-39, 2002.

[12] Dhyani, D., NG Keong, W., Bhowmick, S. S.: A survey of Web Metrics. In ACM Computing Surveys, Vol. 34, No. 4, pp. 469-503, 2002.

ABOUT THE AUTHORS

Jakub Snopek, Ivan Jelínek, Department of Computer Science and Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, E-mail: snopekj@fel.cvut.cz, jelinek@fel.cvut.cz.