International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:2, No:9, 2008

Web Traffic Mining using Neural Networks

Farhad F. Yusifov

***Abstract*—**With the explosive growth of data available on the Internet, personalization of this information space become a necessity. At present time with the rapid increasing popularity of the WWW, Websites are playing a crucial role to convey knowledge and information to the end users. Discovering hidden and meaningful information about Web users usage patterns is critical to determine effective marketing strategies to optimize the Web server usage for accommodating future growth. The task of mining useful information becomes more challenging when the Web traffic volume is enormous and keeps on growing. In this paper, we propose a intelligent model to discover and analyze useful knowledge from the available Web log data.

***Keywords*—**Clustering, Self organizing map, Web log files, Web traffic.

I. Introduction

THE World Wide Web (WWW) is continuously growing with the information transaction volume from Web servers and the number of requests from Web users. Providing Web administrators with meaningful information about users' access behavior and usage patterns has become a necessity to improve the quality of Web information service performances. As such, the hidden knowledge obtained from mining Web server traffic and user access patterns could be applied directly for marketing and management of E-business, E-services, E-searching, E-education and so on [1].

The evolution of the Internet has lead to an enormous proliferation of the available information and the personalization of this information space has become a necessity. The knowledge obtained by learning web users' preferences can be used to improve the effectiveness of their web sites by adapting the web information structure to the users behavior. Automatic knowledge extraction from web log files can be useful for identifying such reading patterns and infer user profiles [2]- [4].

However, it is hard to find appropriate tools for analyzing raw web log data to retrieve significant and useful information. There are several commercially available web log analysis tools, but most of them are disliked by their users and considered too slow, inflexible, expensive, difficult to maintain or very limited in the results they can provide.

Recently, the advent of data mining techniques for discovering usage patterns from web data (web log mining or web usage mining) made it possible to mine typical user profiles from the vast amount of access logs. Web usage mining can be viewed as the extraction of usage patterns from access log data containing the behavior characteristics of users [4]-[6].

The statistical data available from the normal Web log files or even the information provided by most conventional Web server analysis tools including commercial Web trackers could only provide explicit information due to the natural limitation of statistic methodology used. Computational Web Intelligence (CWI), a recently coined paradigm, is aimed to improve the quality of intelligence in the Web technology and includes Web mining as one main stream [1]. Generally, the Web analysis relies on three general sets of information given: past usage patterns, degree of shared content and inter-memory associative link structures corresponding to the three subsets in Web mining namely: Web usage mining, Web content mining and Web structure mining. In Web usage mining, the pattern discovery consists of several steps including statistical analysis, clustering, classification and so on. Most of the current research are focusing on finding patterns but with little effort on the detailed pattern/trend analysis that varies with the Web environments and the intelligent paradigms considered [7]-[10].

In our research, we purposed a clustering algorithm to discover hidden relationships among the Web server data and access patterns. The unsupervised learning algorithm Self Organizing Map (SOM) was used for the clustering task to discover usage patterns from Web server logs. In order to make the analysis more intelligent we also used the clustered data to predict the daily and hourly traffic including request volume and page volume.

II. Mining Web Usage Data

Web usage mining is defined as the process of applying data mining techniques to the discovery of usage patterns from web logs data, to identify web users' behavior. In Web mining, data can be collected at the server-side, client-side and proxy servers. To summarize, Web server logs explicitly records browsing behavior of site visitors, Client-side data collection can be implemented by using a remote agent or by modifying the source code of an existing browser and Web proxies act as an intermediate level of caching between client browsers and Web servers [2]-[4].

The information provided by the data sources described above can be used to construct several data abstractions, namely users, page-views, click-streams, and server sessions. A user is defined as a single individual that is accessing file Web servers through a browser. In practice, it is very difficult to uniquely and repeatedly identify users. A user may access

Manuscript received May 30, 2006.

F. F. Yusifov is with the Institute of Information Technologies, Azerbaijan National Academy of Sciences, Baku, AZ1141 Azerbaijan (e-mail: y_ferhad@yahoo.com).

the Web through different machines, or use more than one browser at one time. A page-view consists of every file that contributes to the display on a user's browser at one time and is usually associated with a single user action such as a mouse-click. A click-stream is a sequential series of page-views requests. Note that any page view accessed through a client or proxy-level cache will not be recorded on the server side. A server session (or visit) is the click-stream for a single user for a particular Web site. The end of a server session is defined as the point when the user's browsing session at that site has ended [3]-[10].

The process of Web usage mining can be divided into three phases: preprocessing, pattern discovery, and pattern analysis [3]-[8].

Preprocessing consists of converting usage information contained in the various available data sources into the data abstractions necessary for pattern discovery. Another task is the treatment of outliers, errors, and incomplete data that can easily occur due reasons inherent to web browsing. The data recorded in server logs reflects the (possibly concurrent) access of a Web site by multiple users, and only the IP address, agent, and server side click-stream are available to identify users and server sessions. However, it is important to notice that the data collected by server logs may not be entirely reliable because some page views may be cached by the user's browser or by a proxy server. In a Web server log, all requests from a proxy server have the same identifier, even though the requests potentially represent more than one user. Also, due to proxy server level caching, multiple users throughout an extended period of time could actually view a single request from the server. The Web server can also store other kinds of usage information such as cookies, which are markers generated by the Web server for individual client browsers to automatically track the site visitors [3], [4].

After each user has been identified (through cookies, logins, or IP/agent analysis), the click-stream for each user must be divided into sessions. As we cannot know when the user has left the Web site, a timeout is often used as the default method of breaking a user's click-stream into sessions [2].

The next phase is the pattern discovery phase. Methods and algorithms used in this phase have been developed from several fields such as statistics, machine learning, and databases. This phase of Web usage mining has three main operations of interest: association (i.e. which pages tend to be accessed together), clustering (i.e. finding groups of users, transactions, pages, etc.), and sequential analysis (the order in which web pages tend to be accessed) [3]-[5]. The first two are the focus of our ongoing work. Pattern analysis is the last phase in the overall process of Web usage mining. In this phase the motivation is to filter out uninteresting rules or patterns found in the previous phase. Visualization techniques are useful to help application domains expert analyze the discovered patterns.

III. Experimental Analysis using SOM

In Web usage mining research, the method of clustering is broadly used in different projects by researchers for finding the usage patterns or user profiles. Among all the popular clustering algorithms, SOM has been successfully used in Web mining projects [1], [7]. A self organizing map was used to cluster the user access records. The SOM is an algorithm used to visualize and interpret large high-dimensional data sets. The map consists of a regular grid of processing units, neurons. A model of some multidimensional observation, eventually a vector consisting of features, is associated with each unit. The map attempts to represent all the available observations with optimal accuracy using a restricted set of models. At the same time the models become ordered on the grid so that similar models are close to each other and dissimilar models far from each other (as in Fig. 1) [1], [11], [12]. Fitting of the model vectors is usually carried out by a sequential regression process. In our approach, with the Web usage data from high dimensional input data, finally a 2D map of Web usage patterns with different clusters could be formed from the SOM training process. From all the experiments with different parameter settings, the best solution was selected when minimum errors were obtained.

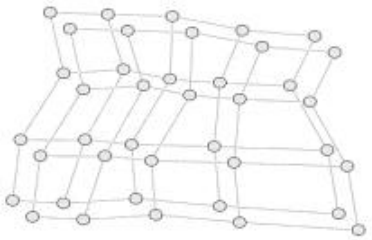

Fig. 1 Self Organizing Map

We used the MATLAB environment to simulate the various experiments. For example, in experiments, we attempted to cluster the data depending on the total activity for each day of the month using 'number of request', 'page volume' and 'time index' as input features. In experiments used web log data (server usage statistics) of main Web server located at http://www.science.az. The training process using SOM produced three clusters and the developed 2D cluster map is shown in Fig. 2. As shown in Fig. 2. More requests consist of Cluster 3, which occurred with heavy traffic volume during normal working weekdays.

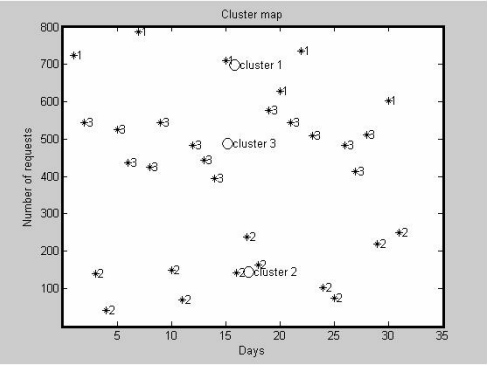

Fig. 2 Cluster requests number analysis

From the SOM clusters, provided useful information related to the user access patterns, which could not be possible by using conventional statistical approaches. Also, these techniques might be useful for the Website tracker software vendors to provide more useful information to the Web administrators.

IV. Conclusion

The discovery of useful knowledge, user information and server access patterns allows Web based organizations to mining user access patterns and helps in future developments, maintenance planning and also to target more rigorous advertising campaigns aimed at groups of users. Previous studies have indicated that the size of the Website and its traffic often imposes a serious constraint on the scalability of the methods.

As popularity of the web continues to increase, there is a growing need to develop tools and techniques that will help improve its overall usefulness.

References

[1] X. Wanga, A. Abraham, K. A. Smitha. Intelligent web traffic mining and analysis. Journal of Network and Computer Applications, vol. 28, 2004, pp. 147–165.

[2] P. Batista, M. J. Silva, "Mining web access logs of an on-line newspaper," (2002), http://www.ectrl.itc.it/rpec/RPEC-Papers/11-batista.pdf.

[3] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: Information and pattern discovery on the World Wide Web," Proc. 9th IEEE Int. Conf. Tools with Artificial Intelligence, Nov. 1997, pp. 558–567.

[4] R. Kosala, H. Blockeel, Web Mining Research: A Survey, SIGKKD Explorations, vol. 2(1), July 2000.

[5] M. Baglioni, U. Ferrara, A. Romei1, S. Ruggieri, and F. Turini, Preprocessing and Mining Web Log Data for Web Personalization. (2003), http://www.di.unipi.it/~ruggieri/Papers/aiia2003.pdf.

[6] R. Iváncsy, I. Vajk, Different Aspects of Web Log Mining. 6th International Symposium of Hungarian Researchers on Computational Intelligence. Budapest, Nov., 2005.

[7] J. Everts and M. Bulacu, Assignment: Clustering of Web Users. Groningen University, Netherlands, Nov. 22, (2005) http://www.ai.rug.nl/ki2/assignments/ki2-assig03.pdf.

[8] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKKD Explorations, vol.1, Jan 2000.

[9] B. Mobasher, H. Dai, T. Luo, N. Nakagawa, Y. Sun, J. Wiltshire, Discovery of Aggregate Usage Profiles for Web Personalization, Proc. of the Web Mining for E-Commerce Workshop (WebKDD'2000), August 2000.

[10] B. Moshaber, R. Cooley, J. Srivastava, Automatic Personalization Based on Web Usage Mining, Communications of the ACM, vol.43(8), 2000.

[11] A. Abraham. Business Intelligence from Web Usage Mining. Journal of Information & Knowledge Management, Vol. 2, 2003, pp. 375-390

[12] S. K. Pal, V Talwar, P Mitra. Web Mining in Soft Computing Framework: Relevance. State of the Art and Future Directions. *IEEE Trans. on Neural Networks*, vol.13 (5), 2002, pp. 1163–77.

Farhad F. Yusifov is postgraduate student and scientific worker at the Institute of Information of Technologies of Azerbaijan National Academy of Sciences. He received his Master's degree in Data processing and automation control systems from State Oil Academy in Baku, Azerbaijan. His primary research interests include various areas in artificial intelligence, particularly in the area of Web usage mining.