Hujun Yin
José A.F. Costa
Guilherme Barreto (Eds.)

LNCS 7435

Intelligent Data Engineering and Automated Learning – IDEAL 2012

13th International Conference
Natal, Brazil, August 2012
Proceedings

Springer

Using SOM to Clustering of Web Sessions Extracted by Techniques of Web Usage Mining

Fábio A. Procópio de Paiva[1] and José Alfredo F. Costa[2]

[1] IFRN, Zona Norte Campus, Natal, Brazil
fabio.procopio@ifrn.edu.br
[2] UFRN, Department of Electrical Engineering, Natal, Brazil
jafcosta@gmail.com

Abstract. Everyday a huge amount of pages are published on the Web, and, as a consequence, the users' difficulty to locate those that will meet their needs is increasingly bigger. The challenge for web designers and e-commerce companies is to identify groups of users that present similar interests in order to personalize navigation environments to meet those interests. In an attempt to offer that to the countless web users, in the last years, several researches have been done on clustering applied to Web Usage Mining. In this paper, a log file is preprocessed to map the sequence of visits for each user's session. A Session-Path Matrix is used as input to SOM Map and identifying patterns between each session. The results show the similarities between the sessions based on time spent on visited paths and volume transferred.

Keywords: Web Usage Mining, SOM, users' sessions, clustering, navigation path.

1 Introduction

The available content on WWW is stored in many formats (audio, images, text and others), covers different areas of knowledge and includes users with multiple interest profiles. However, the challenge is to provide rapid access to information and, above all, make them relevant to the users' interests. Thus, the WWW appears as an ideal environment to apply Data Mining techniques, known as Web Mining.

Web Mining consists in extracting interesting and potentially useful patterns and implicit information from artifacts or activity related to the WWW. Moreover, Web Mining is one of the most important areas of Computer Science and Information Science [9]. The Web Mining is classified in three categories [13]: Web Content Mining, Web Structure Mining and Web Usage Mining. The Web Content Mining is the process that extracts knowledge from the Web and analyzes the contents of its documents. Web Structure Mining tries to discover the model underlying the link structures of the Web. And finally, the purpose of Web Usage Mining is to apply statistical and data mining techniques to the preprocessed web log data, in order to discover useful patterns [1].

H. Yin et al. (Eds.): IDEAL 2012, LNCS 7435, pp. 484–491, 2012.

Nowadays Web Usage Mining is an area of interest for many researchers [13] because a) the record of accessed pages allows mapping the users' behavior; b) the frequent accesses can be used to improve link structure and; c) it allows suggesting changes in pages design.

In reference to [12], an important point to be observed in Web Usage Mining is the users' clustering according to their characteristics. From an analysis of clusters, a web designer can identify the users' interest and thus offer more personalized services to a group of them. Still, according to [14], the results generated by clustering techniques can be used to analyze the systems' performance and network communication. A method to cluster users is measuring the similarity between them based on their interests. There are several measures that can be used [12]: Usage Based Measure, Frequency Based Measure, Viewing-time Based Measure and Visiting-Order Based Measure.

In literature, there are several researches that propose methods to identify web users' interests, although this is a complex task [8]. According to [6], the authors considered the time as a good measure to evaluate the users' interest and used the naïve Bayes method to model and predict the users' navigational behavior. A model based on ant colonies has been proposed by [3] to identify users' browsing patterns. In this model, the authors used access frequency and the time spent as measures to identify the users' interest. In reference to [5], the authors proposed a method based on Self-Organizing Maps that uses Web Content Mining and Web Usage Mining clustering techniques to help visitors identify relevant information quickly.

In this paper, a log file is preprocessed to map the sequence of visits (i.e. path) for each user's session. Then a Session-Path Matrix is used as input data to Self-Organizing Map (SOM) and to identify patterns between each session. The results show the similarities between the sessions based on time spent on visited paths and volume transferred.

This paper is organized as it follows: section 2 presents the concepts of Web Usage Mining; Section 3 describes briefly the Self-Organizing Maps (SOM); Section 4 details the data preprocessing and sessions' clustering using the SOM. The last section presents conclusions and future research directions.

2 Web Usage Mining

Web Usage Mining is a Data Mining process used to discover usage patterns of the information on the Web and aims to provide an understanding of the interests and behavior of web users[10]. Thus, analyzing the user's access logs on websites, it is possible to understand their actions and thus enable customization of a navigation environment.

Often, Web Usage Mining includes the following steps [2]:

(a) Preprocessing – removes inconsistencies and noise of the data sources in order to leave only those that are really significant. The stage also includes tasks such as users' identification, sessions' identification and definition of the full path of navigation [9].

(b) Pattern discovery – discovers the user's interests (or a group of them) and build a model according to these preferences.
(c) Pattern Analysis – the main objective is to filter the information that are apparently irrelevant to viewing and interpreting the user interest patterns.

When a user navigates on a website his/her interactions are recorded in files called web server log file. This record has the form of a single transaction and is appended in ASCII text file. Nowadays there are three ways to obtain user's access logs: a) client log file, b) proxy log file and c) server log file. The delimiter of this file type can be a comma, a blank or a tab.

There are three types of server log file available to capture the activities of a user on websites [4]: Common Log Format, Log Format and Extended Log Format IIS. This work is based on a server log file using Common Log Format.

The most common and simple way to analyze a log file is using a statistical method. However, there are more sophisticated methods, such as Association Rules Mining, Sequential Pattern Discovery, Clustering and Classification [1].

3 Self-Organizing Maps (SOM)

The Self-Organizing Map (SOM) is one of the most popular artificial neural network algorithms and it is based on unsupervised competitive learning, which means that the training is entirely data-driven. The training of the neurons present competitive and cooperative processes [15]. The SOM defines a mapping from the high dimensional input data space onto a regular, usually, two-dimensional array of nodes. Each neuron i of the SOM is represented by an p-dimensional weight vector $m_i = [m_{i1}, m_{i2}, ..., m_{ip}]^T$, where p is equal to the dimension of the input vectors[16].

3.1 SOM Properties

Assume that $\Re^p$ is an input space with a topology defined by the metric relation between the vectors $\boldsymbol{x} \in \Re^p$. Consider K a discrete output space with a topology that is defined by arranging a set of neurons like nodes on a grid. According to Haykin [21], the SOM algorithm can be defined as a non-linear transformation, $\boldsymbol{\Phi}$, called a *feature map*, which maps the input space $\Re^p$ to the output space K:

$$\boldsymbol{\Phi}: \Re^p \rightarrow K$$

Given an input vector $\boldsymbol{x}$, the SOM algorithm identifies the winner neuron c in the output space K according to a features map $\boldsymbol{\Phi}$. The main properties of feature mapping computed by the SOM include [22]:

1. Approximation of the input space – the features map $\boldsymbol{\Phi}$, represented by the set of code vectors $\boldsymbol{w}_i$ in the output space K, provides a good approximation of the input space $\Re^p$. This strategy is based on the vector quantization theory, the motivation for which is data compression [21];
2. Topological ordering – the SOM algorithm attempts to preserve as well as possible the topology of the original space, i.e., it tries to make the neighboring

neurons in the rectangular grid (output space *K*) present weight vectors that represent neighboring patterns in the input space $\Re^p$.

3. Density matching – when properly trained, the features map Φ approximates the probability distribution of data in the input space $\Re^p$.

4 Data Preprocessing

The log files cannot be used before a treatment [7] because they have some records which do not add any value to the Data Mining. However, it makes it difficult to analyze the users' behavior. These irrelevant data are called noise and usually are generated by web robots or when images, videos, audios, CSS, javascripts and flash animations are loaded within the pages of a website.

The access logs used in this paper did not show which user made a particular access, thus users were identified from the recorded IP address into a log file. However, this does not prevent that the same IP can be used by different users or even the same user to different IPs. Thus, the analysis was performed on the basis of access sessions. In this context, a session consists of an access to a page (or set of them) recorded to the same IP and the time difference between the instants t_i and t_{i-1} (where t_i is the time that the page p_j was accessed and t_{i-1} the access time to the page p_{j-1}) is less than or equal to 30 minutes [11][18]. The method which is often used to distinguish two sessions is setting the time of timeout. Many web usage analysts and commercial applications set the timeout threshold at 30 minutes [7][19].

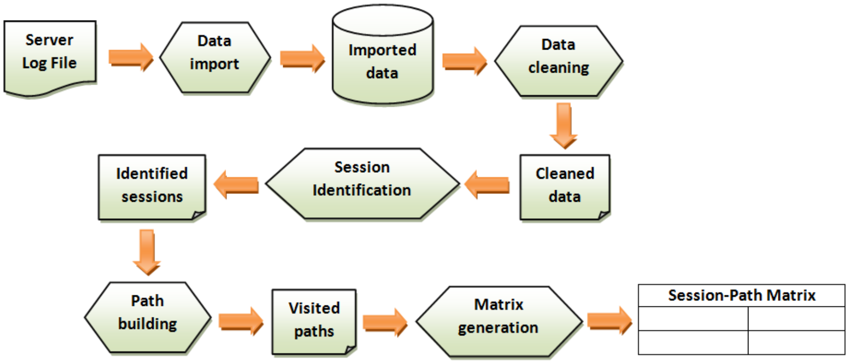

Fig. 1. Data preprocessing to Session-Path Matrix generation

As it was mentioned before, data from a log file cannot be used before being preprocessed. Fig. 1 shows the procedures used for this. Next, each step is described:

1. Data import – imports logs from a text file to database. To facilitate the data manipulation, the logs were imported into SQL Server 2008 R2 Express.
2. Data cleaning – performs noise removal from original data. After importing data, unsuccessful requests logs have been removed. In addition, records of images, videos, audios, javascript and flash animations.

3. Session identification – aims to identify the user's sessions. The data were sorted by timestamp field. When the user (identified by IP) of the r_i record is different from r_{i-1} record or when the difference between the timestamps of r_i and r_{i-1} is greater than 30 minutes a new session is created.
4. Path building – consists of mapping the full path (sequence of visits to pages) of each session, calculating the time spent and the transferred data volume.
5. Session-Path Matrix – creates a matrix containing session identifier, path identifier, time spent on each page and transferred volume.

The result of preprocessing produces the Session-Path Matrix, Fig. 1, in which rows represent sessions and columns are visited path information by each session.

5 Sessions Clustering

After preprocessing we use the data as input to SOM Toolbox, a SOM library developed to Matlab. Based on Path-Session Matrix, a file sample is generated by preprocessing script.

In reference to [12] the authors highlighted Visiting-Order as a measure to evaluate degree of interest of web users and this contributed to what we assume that sequence of pages accessed by users in a user's sessions (i.e. path) is a important variable for measuring the similarity in the degree of web sessions.

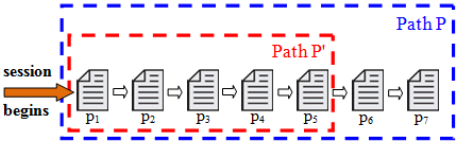

Fig. 2. Path P' is a segment of Path P

We have used access-logs of the University of Saskatchewan (available at http://ita.ee.lbl.gov/html/traces.html), located in Saskatoon, Saskatchewan, Canada. After preprocessing, we have extracted 125 records from Session-Path Matrix and such records are divided into 5 classes, each with 25 instances of visited paths. The classes indicate the paths visited by sessions: Path6, Path9, Path13, Path29 and Path34. Here, P is the full path visited by a session S. Eventually P can be segmented into smaller path called P'. In Fig. 2, e.g., P is composed of 7 pages (p_1, p_2, p_3, p_4, p_5, p_6 and p_7) and S begins when page p_1 is accessed. From path P, path P' has been created which is a shortest length path containing 5 pages (p_1, p_2, p_3, p_4 and p_5). In our experiments, we evaluate paths with length greater than or equal to 5 pages because they presented a higher representation of data. When paths P presented length greater than 5 pages, we generated a path P' of length 5 in order to compare paths only of the same length. The attributes defined for each instance are:

- timeToPg2: elapsed time between accesses page p_1 and page p_2;
- timeToPg3: elapsed time between accesses page p_1 and page p_3;

- timeToPg4: elapsed time between accesses page p_1 and page p_4;
- timeToPg5: elapsed time between accesses page p_1 and page p_5 and;
- volume: volume transferred (bytes) in P'.

In order to visualize the results produced by SOM Toolbox, we set parameters for initialization, for training and for visualization of the map. The evaluation was performed in several experiments using the combination of parameters.

Table 1. Configuration parameters of the SOM

Parameter	Value
Data normalization	Variance
Inicialization	Linear
Topology	Hexa
Neighborhood function	Gaussian
Training	Batch
Epochs	3000
Dimension X / Dimension Y	15 / 15

To measure the map generation quality after training, we use two metrics: Quantization Error (Qe) and Topographical Error (Te). Quantization Error measures the average distance between each data vector and its best matching unit [20] and Topographical Error indicates how much the SOM preserving topology of the input data. Based on the parameters defined, in Table 1, the results obtained for the two measurements were Qe = 0.2533 and Te = 0.0400.

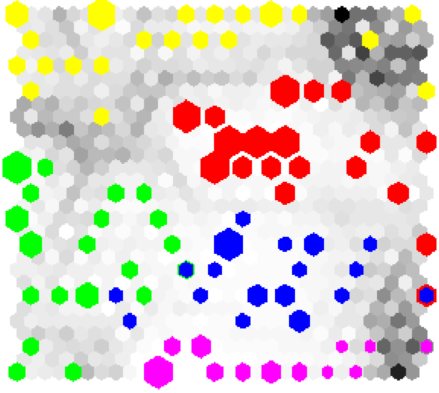

Fig. 3. U-Matrix

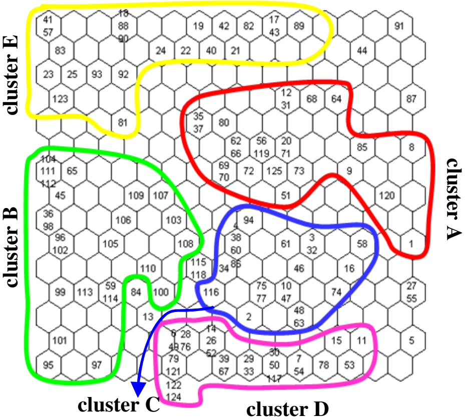

Fig. 4. Clustering of web sessions

After training, the map is displayed by U-Matrix in which different colors are used to represent the distances between neurons. The U-Matrix is commonly used visualization method for the cluster analysis using SOM and based on the distance in input space between a weight vector and its neighbors on map [17]. In U-Matrix a light shade means that they are close and a darker shade can be interpreted as a clusters separator.

We also use the histogram hits that allow the identification of the parts of the map that best represent the data. In Fig. 3, the U-Matrix is colored in 5 different classes, each representing a sessions' cluster. In Fig. 4, the labels, such as 41 and 57 (cluster E), represent the sessions' identifiers. We draw the clusters' line manually and we noted that:

- Cluster A – represents the users' sessions that accessed Path 6;
- Cluster B – represents the set of users' sessions that browsed on Path 9;
- Cluster C – represents the users' sessions that accessed Path 13;
- Cluster D – represents the set of users' sessions that browsed on Path 29;
- Cluster E – represents the users' sessions that accessed Path 34.

6 Conclusion

The Internet popularization has had an impact on the number of pages published on the network. The amount of available information configures a huge source of data and the difficulty encountered by web users to find resources that meet their interests is a problem that some areas such as Information Retrieval and Web Mining have been trying to solve.

In this paper, five steps were performed to preprocess data extracted from a log file: Data import, Data cleaning, Session identification, Path building and Session-Path Matrix building. In order to visualize the similarities between users' sessions, the Session-Path Matrix has been used as an input data to Kohonen's Map. The figures 3 and 4 show the existence of similar patterns between sessions that were browsed by the same path.

Future works will focus ontologies to assign semantics to the pages visited and therefore the browsed paths. Furthermore, such works will create a mechanism to enable the similar paths recommendation to analyzed sessions.

References

1. Eirinaki, M., Vazirgiannis, M.: Web Mining for Web Personalization. ACM Transactions on Internet Technology 3, 1–27 (2003)
2. Etminani, K., Delui, A.R., Yanehsari, N.R., Rouhani, M.: Web usage mining: Discovery of the users' navigational patterns using SOM. In: First International Conference on Networked Digital Technologies, pp. 224–249. IEEE Press, New York (2009)
3. Ling, H., Liu, Y., Yang, S.: An Ant Colony Model for Dynamic Mining of Users Interest Navigation Patterns. In: IEEE International Conference on Control and Automation, pp. 281–283. IEEE Press, New York (2007)
4. Hussain, T., Asghar, S., Masood, N.: Web usage mining: A survey on preprocessing of web log file. In: International Conference on Information and Emerging Technologies, pp. 1–6. IEEE Press, New York (2010)

5. Petrilis, D., Halatsis, C.: Combining SOMs and Ontologies for Effective Web Site Mining. In: Self Organizing Maps - Applications and Novel Algorithm Design, pp. 109–124. InTech, Rijeka (2011)
6. Khosravi, M., Tarokh, M.J.: Dynamic mining of users interest navigation patterns using naive Bayesian method. In: IEEE International Conference on Intelligent Computer Communication and Processing, pp. 119–122. IEEE Press, New York (2010)
7. Markov, Z., Larose, D.T.: Data Mining the Web – uncovering patterns in Web Context, Structure, and Usage. Wiley, New Jersey (2007)
8. Nadi, S., Saraee, M., Davarpanah-Jazi, M.: A fuzzy recommender system for dynamic prediction of user's behavior. In: International Conference on Internet Technology and Secured Transactions, pp. 1–5. IEEE Press, New York (2010)
9. Pamnani, R., Chawan, P.: Web Usage Mining: A Research Area in Web Mining. In: Proceedings of ISCET 2010, pp. 73–77 (2010)
10. Vellingiri, J., Pandian, S.C.: A Survey on Web Usage Mining. Global Journal of Computer Science and Technology, 66–72 (2011)
11. Xiao, J., Zhang, Y.: Clustering of web users using session-based similarity measures. In: International Conference on Computer Networks and Mobile Computing, pp. 223–228. IEEE Press, New York (2001)
12. Xiao, J., Zhang, Y., Jia, X., Li, T.: Measuring similarity of interests for clustering Web-users. In: 12th Australasian on Database Conference, pp. 107–114. IEEE Press, New York (2001)
13. Pani, S.K., Panigrahy, L., Sankar, V.H., Ratha, B.K., Mandal, A.K., Padhi, S.K.K.: Web Usage Mining: A Survey on Pattern Extraction from Web Logs. International Journal of Instrumentation, Control & Automation 1(1) (2011)
14. Wang, S., Xu, C., Wu, R.: Clustering Method Based on Fuzzy Multisets for Web Pages and Customer Segments. In: International Seminar on Business and Information Management, vol. 2, pp. 125–128. IEEE Press, New York (2008)
15. Kohonen, T.: Self-Organizing Maps, 2nd edn. Springer, Berlin (1997)
16. Costa, J.A.F., Netto, M.L.A.: Clustering of complex shaped data sets via Kohonen maps and mathematical morphology. In: Dasarathy, B. (ed.) Proceedings of the SPIE, Data Mining and Knowledge Discovery, vol. 4384, pp. 16–27 (2001)
17. Yamaguchi, T., Ichimura, T.: Visualization using multi-layered U-Matrix in growing Tree-Structured self-organizing feature map. In: IEEE International Conference Systems, Man and Cybernetics (SMC), pp. 3580–3585 (2011)
18. Cheng, X., Liu, H.: Personalized Services Research Based on Web Data Mining Technology. In: Second International Symposium on Computational Intelligence and Design, vol. 2, pp. 177–180 (2009)
19. Dong, Y., Zhang, H., Jiao, L.: Research on Application of User Navigation Pattern Mining Recommendation. In: Proceedings of the 6th World Congress on Intelligent Control and Automation, vol. 2, pp. 6106–6110 (2006)
20. Uriarte, E.A., Martín, F.D.: Topology Preservation in SOM. International Journal of Mathematical and Computer Sciences, 19–22 (2005)
21. Haykin, S.: Neural networks: A comprehensive foundation, 2nd edn. Macmillan College Publishing Company, N. York (1999)
22. Gonçalves, M., Netto, M., Zullo Jr., J., Costa, J.A.F.: A new method for unsupervised classification of remotely sensed images using Kohonen self-organizing maps and agglomerative hierarchical clustering methods. Intl. Journal of Remote Sensing 29(11), 3171–3207 (2008)