Web Usage Mining Using Self Organized Maps

Paola Britos, Damián Martinelli, Hernán Merlino, Ramón García-Martínez

PhD Computer Science Program, National University of La Plata. Software & Knowledge Engineering Center, Buenos Aires Institute of Technology. Intelligent Systems Laboratory, University of Buenos Aires. Buenos Aires. Argentina

Summary
This paper detail the capacity of use of Self Organized Maps, kind of artificial neural network, in the process of Web Usage Mining to detect user's patterns. The process detail the transformations necessaries to modify the data storage in the Web Servers Log files to an input of Self Organized Maps

Key words:
Web mining - Self Organized Maps - User's pattern - Web Servers Log files.

1. Introduction

Data mining is a set of techniques and tools used to the no trivial process of extracting and present implicit knowledge, no knowledge before, this information is useful and human reliable; this is processing from a great set of data; with the object of describing in automatic way models, no knowledge before; to detect tendencies and patterns [Felgaer, 2004; Piatetski-Shapiro *et al*., 1991; Chen *et al*., 1996; Mannila, 1997]

The Web Mining are the set of techniques of Data Mining applied to Web [Cernuzzi-Molas, 2004]. The Web Usage Mining is the process of applying techniques to detect patterns of usage to Web Page [Srivastava *et al*., 2000; Kosala-Blockeel, 2000]. The Web Usage Mining use the data storage in the Log files of Web server as first resource; in this file the Web server register the access at each resource in the server by the users [Batista-Silva, 2002; Borges-Levene, 2000; Chen y Chau, 2004].

With the Web usage mining we can obtain: (a) the understanding of the users pattern. (b) We can obtain the information by personalize to the site. (c) Build tune up the server. (d) Modify the site according to the preferences access by the users. (e) Build business rules. By this we can get: (a) new clients. (b) Marketing campaigns (c) build a more efficient site [Abraham y Ramos, 2003]. The user's habit detection has 3 steps: (a) pre-processing. (b) Detect commons patterns, and (c) analyze the patterns [Mobasher *et al*., 1999]

Pre-processing is the action to convert the data storage in logs files, using techniques of cleaning data, abstraction data, user detection and sessions detection. In the next step, detect commons patterns; we can use statistics techniques, rules association, gather, classification, etc. In the last step, analyze the patterns, Web Usage Mining try to understand the patterns detected in before step. The most common techniques is data visualization applying filters, zooms, etc [Keim, 2002; Ankerst, 2001]

The artificial neural networks (ANN), try to simulate the action doing by the human brain; RNA has the possibility of get abstraction of data and work with incomplete data or with errors, RNA has knowledge and can adapt it; and operate in real time [Grosser, 2004; Daza, 2003]. RNA is built by a common part called neurons. These units of processing are interconnected; each neuron has it this activation threshold. The learning in RNA is built by the adjustment of activation threshold in each neuron [Roy, 2000; Abidi, 1996; García Martínez *et al*., 2003]

2. Problem description

2.1 Logs processing

A critical step in the identification of user's habit in web sites is the cleaning and transformation of Web server Log files; and user's sessions identification [Mobasher, 1999].

2.2 Log files cleaning

Cleaning Web server Log files has a lot of steps [Huysmans *et al*., 2003; Mobasher, 1999; Pierrakos *et al*., 2001; Lalani, 2003; Kerkhofs, 2001; Eirinaki & Vazirgiannis, 2003]. When one user request a page, this request is added to the Log File, but, if this page has images, in they will be added in the Log file. This is the same for any resource in the page, for example JavaScript's, flash animations, videos, etc. In most of the cases these resources aren't necessary for the detection of user's habits; for this reason is good cat this records from the log file; to do this task we only need to search records by file extension. To give a little list we can consider cut extensions with jpg, jgeg, gif, js, css, swf, avi, mov, etc. In some case is proper to filter page inserted in others with frames; in other way is common to generate pages dynamically. Errors code in HTTP is used too filter records in the Logs files, the most common errors in HTTP are: error code 200, 4003 (recourse not found), 403 (access denage), and 500 (internal server error). For a complete list we can consult at RFC 2616 [RFC 2616].

Manuscript received June 5, 2007
Manuscript revised June 20, 2007

2.3 Users identifications

After the log files cleaning, we need to identify user's sessions. We have some methods to detect sessions each one pros and cons. One method is detecting the use of cookies [Eirinaki & Vazirgiannis, 2003; Huysmans et al., 2003; Kerkhofs, 2001]. W3C [WCA] define cookies as " data sent by the server to the client, data locally storage in cookies and is send to the server with each request". In other words the cookies are HTTP headers in string format. Cookies are used to identify users behind server's access, and what resources the user accesses. One problem with this method is; the users can lock the use of cookies, and the server after that can't storage information locally in the user machine; other problem is; the user can delete the cookies. Another method to identify users is using Identd [Eirinaki & Vazirgiannis, 2003]. Identd is a protocol defined in RFC 1413 [RFC 1413], this protocol permits detect to a user connected by the unique TCP connection. The problem with Identd is the terminal user needs to configure with the Identd support. Other method is detect the users in log files by the IP direction registered in each record. Another method is the explicit users registration each the user time accesses to the site. At last we can detect a users with the users name added in the log file in filed name authuser.

2.4 User session's identification

After identify the users, we need to identify the sessions. To do this we can divide the access of the same users in sessions. It's difficult to detect when one session is finish and start another. To detect sessions is common use of time between requests; if two requests are called in of time frame, we can suppose that these requests are in the same session; in other way below of time frame we can consider two different sessions. A good time frame is 25.5 minutes [Catledge, L., Pitkow, J.; 1995].

2.5 User's habit identification

After that all log processing, we can start to detect the user's habit.

3. Solution proposed

This paper proposes the use of SOM to identify the use's habits. This kind of artificial neural network will be try to gather the users by patterns of pages accesses. To obtain this result we need to process the Web Log files to identify users and session of users; after that with this session's user, we'll train the ANN. The selection of SOM is so because it isn't necessary to supervise to the training.

3.1 Steps of process data

This paper uses the Log Common Format (CLF) to analyze the Log Files. The programmer team has develop a parametric system to do this analysis. With this tool we can set filters, for example we can select which files extensions we can obtain from the Log. Other filter is the error codes, we can define what errors code will be consider; by default only the error code 200 is obtained. The third filter is to detect the dynamic pages. In other way we can identify a variable name to detect a dynamic page. To detect users the tool uses the IP direction; all request do it form the same IP direction are considered a unique user. We can define in the tool the minimal quantity of page in each session. At last we can define the minimal threshold of frequency for each page, less than the number inserted the page will be excluded of analyzes.

3.2 User's habits identify

After that identifying the session's users, we need to generate the correct format to insert the data in the artificial neural network. The normalization method is 0 (zero) for no presence of the page and 1 (one) for presence of the page; with each session's user we generate a vector with 0 or 1. These vectors are the input of the artificial neural network. The artificial neural network, in this case SOM, has an arbitrary number of input neurons, this number is defined by the user, to do this the users get the number of the most common pages in the site; by the way each site is probably has different artificial neural network architecture. The output of SOM is a map of NxN dimensions; the user configures N; this is the number of cluster that the users want to obtain; in the output map only one cluster will be activated. The same pattern of input will generate the activation of the same output cluster; similar inputs will be activated near output clusters.

4.Results

This paper compares two sites one of music and another of gastronomic; we can see the comparison between both methods, SOM and K-Means, in both sites. For each site it has been develop a complete process involved in Web Usages Mining.

4.1 Music site Log analysis

In this site we can buy song in MP3 format. In this site we can search songs and listening to a fragment of it before buying.

4.1.1 Session's in each cluster

In figures 1 and 2 we can see the quantity of clusters generated with both methods.

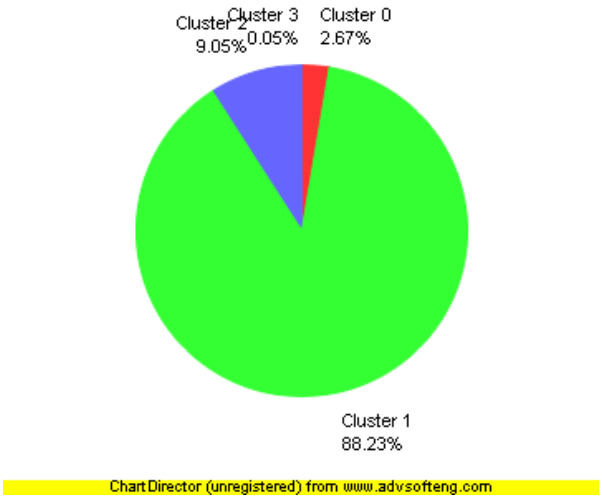

Fig. 1. Percentage of sessions in each cluster with SOM

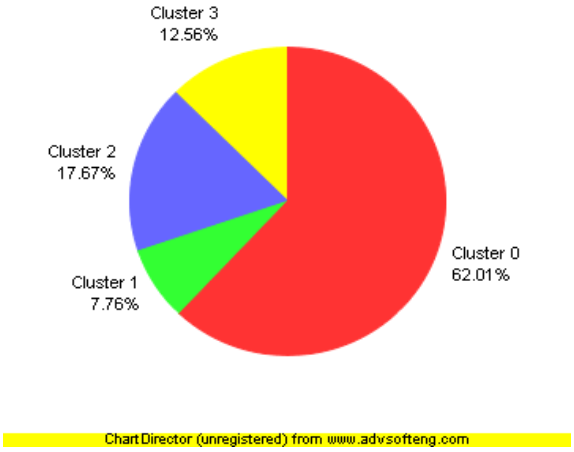

Fig. 2. Percentage of sessions in each cluster with K-Means

4.1.2 Detail of clusters by pages accesses

In table 1 we can see the detail of pages accesses by users in each cluster.

4.1.3 Conclusion of Music site

The result of the comparing two methods, SOM and K-Means, is that SOM is better than K-Means. SOM has a better group of cluster by pages accesses by users, with K-Means; We can obtain only a few pages accesses.

Pages in each Cluster (Porcentage Min = 10%; Quantity Max by Cluster = 100)			
SOM		SOM	
Page	Page	Página	Porcentaje
Cluster 0			
/top.php	98.8312	/login.php	92.6187
/thumbnail.php	96.7532		
/flashes_home.php	95.7143		
Cluster 1			
/login.php	73.8426	/detalle_album.php	100.0
/detalle_album.php	28.1341	/login.php	100.0
Cluster 2			
/top.php	96.7062	/detalle_album.php	100.0
/thumbnail.php	96.1317		
/detalle_album.php	76.7905		
/preview.php	37.8782		
/detalle_grupo.php	29.1076		

Table 1. Detail of pages accesses by users in each cluster

4.2 El cuerpo de Cristo Log analysis

This is a site of food. The site () put on the Web articles about foods and cooking; each registered users can add his articles, when one user adds an article, this is added in the workflow to be approve. In other way this site use a Wiki of cooking. The site has a gallery of images and an engine of search to obtain by the users cooking and culling

4.2.1 Session's in each cluster

In figures 3 and 4 we can see the quantity of clusters obtained with both methods.

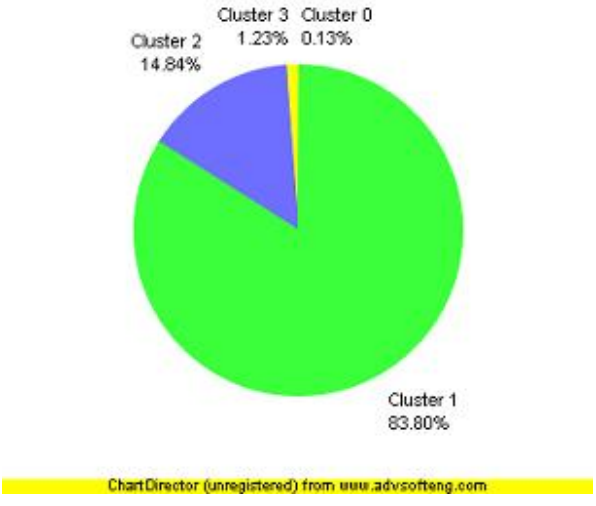

Fig. 3. Percentage of sessions in each cluster with SOM

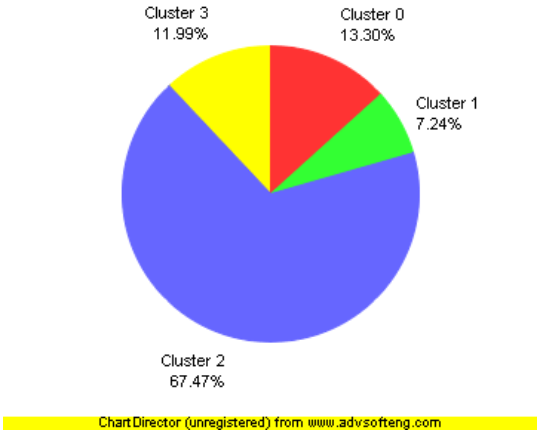

Fig. 4. Percentage of sessions in each cluster with K-Means

4.2.2 Detail of clusters by pages accesses

In tables 2a and 2b we can see the detail of pages accesses by users in each cluster.

Table 2 (a). Detail of pages accesses by users in each cluster

Pages in each Cluster (Porcentage Min = 10%; Quantity Max by Cluster = 100)			
SOM		K-Means	
Page	Porcentage	Page	
Cluster 0			
/tiki-edit_submission.php	100.0	/tiki-editpage.php	100.0
/tiki-view_forum_thread.php	85.7143	/tiki-random_num_img.php	62.8453
/tiki-list_submissions.php	71.4286		
/tiki-browse_image.php	57.1429		
/show_image.php	42.8571		
/tiki-edit_article.php	42.8571		
/tiki-galleries.php	42.8571		
/tiki-pagehistory.php	42.8571		
/topic_image.php	42.8571		
/tiki-directory_browse.php	28.5714		
/tiki-editpage.php	28.5714		
/tiki-index.php	28.5714		
/tiki-list_articles.php	28.5714		
/tiki-meta.php	28.5714		
/tiki-upload_image.php	28.5714		
/tiki-wiki_rss.php	28.5714		
/tiki-searchindex.php	14.2857		
/tiki-view_articles.php	14.2857		
Cluster 1			
/recetodromoII/index.php	19.2417	/tiki-index.php	99.7462
/tiki-editpage.php	18.0145	/tiki-editpage.php	30.203
/tiki-articles_rss.php	14.3765	/tiki-listpages.php	23.0964
/tiki-wiki_rss.php	11.5494	/tiki-list_articles.php	22.5888
/tiki-random_num_img.php	10.103	/tiki-pagehistory.php	21.3198
		/tiki-edit_translation.php	18.5279
		/tiki-wiki_rss.php	18.2741
		/tiki-edit_submission.php	17.7665
		/tiki-register.php	17.7665
		/tiki-articles_rss.php	17.5127
		/tiki-user_information.php	17.2589
		/tiki-view_forum_thread.php	17.2589
		/tiki-print.php	16.7513
		/tiki-searchindex.php	10.9137
		/show_image.php	10.4061
		/tiki-browse_categories.php	10.1523

Table 3 (b). Detail of pages accesses by users in each cluster

Cluster 2			
/show_image.php	78.7129	/recetodromoII/index.php	21.8018
/tiki-browse_image.php	26.1139	/tiki-articles_rss.php	17.583
		/show_image.php	15.1606
		/tiki-wiki_rss.php	14.1535

Cluster 3			
/tiki-editpage.php	100.0	/tiki-pagehistory.php	69.0658
/tiki-edit_submission.php	100.0	/tiki-print.php	36.9066
/tiki-listpages.php	100.0	/tiki-listpages.php	24.8086
/tiki-list_articles.php	100.0	/tiki-editpage.php	18.0704
/tiki-articles_rss.php	98.5075	/tiki-browse_categories.php	15.6202
/tiki-index.php	98.5075	/tiki-edit_translation.php	14.5482
/tiki-register.php	98.5075	/tiki-backlinks.php	14.242
/tiki-wiki_rss.php	98.5075		
/tiki-user_information.php	95.5224		
/tiki-pagehistory.php	92.5373		
/tiki-edit_translation.php	91.0448		
/tiki-print.php	88.0597		
/tiki-view_forum_thread.php	44.7761		
/tiki-edit_templates.php	40.2985		
/tiki-view_forum.php	25.3731		
/tiki-print_article.php	23.8806		
/tiki-backlinks.php	22.3881		
/tiki-list_submissions.php	22.3881		
/tiki-browse_categories.php	20.8955		
/tiki-forums.php	20.8955		
/tiki-browse_image.php	19.403		
/tiki-galleries.php	19.403		
/tiki-list_gallery.php	16.4179		

4.2.3 Conclusion of "El Cuerpo de Cristo" site.

We can see that the SOM method is better than K-Means, because SOM has a better gathering of users.
We obtain with SOM a great number of page accesses by the users in each cluster; in other way, with K-Means, We can obtain only a few page accesses, only one or two with a great number of access. In comparison SOM has more information for the users than K-Means.

5. Conclusions

We can conclude that; to identify common patterns in Web, the self-organized map (SOM) is better than K-Means. SOM has a better group of users. With K-Means we get a few information about user's habits. In other way SOM build some gathering with a great quantity of user's sessions for the same users; but K-Means has a better distribution of user's sessions in each group. At last, the time to identified user's habits is similar for the two tools; but when the quantity of session is increased K-Means is better than SOM.

Acknowledgments

We wish to thank to Miss Carolina Procopio who was deeply concern in helping us in the translation of this paper from spanish to english.

References

Abidi, S.S.R. (1996). Neural networks: their efficacy towards the Malaysian IT environment. School of Computer Sciences. Universiti Sains Malaysia. Penang. Malaysia.

Abraham, A., Ramos, V. (2003). Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming.

Ankerst, M. (2001). Visual Data Mining with Pixel-oriented Visualization Techniques. The Boing Company. Seattle, WA.

Batista, P., Silva, M.J. (2002). Mining Web Access Logs of an On-line Newspaper. Departamento de Informática, Faculdade de Ciencias – Universidade de Lisboa. Lisboa. Portugal.

Borges, J., Levene, M. (2000). A Fine Grained Heuristic to Capture Web Navigation Patterns. ACM SIGKDD, July 2000.

Catledge, L., Pitkow, J., Characterizing browsing strategies in the world wide web. Computer Networks and ISDN Systems, 27(6): 1065-1073, Abril 1995

Cernuzzi, L., Molas, M.L. (2004). Integrando diferentes técnicas de Data Mining en procesos de Web Usage Mining. Universidad Católica "Nuestra Señora de la Asunción". Asunción. Paraguay.

Chen, H., Chau, M. (2004). Web Mining: Machine Learning for Web Applications. Annual Review of Information Science and Technology, 38, 289-329.

Chen, M., Han, J., Yu, P. (1996). Data mining: An overview from database perspective. IEEE Transactions on Knowledge and Data Eng.

Cooley, R., Tan, P.N., Srivastava, J. (1999). Discovery of interesting usage patterns from web data. University of Minnesota.

Daza P., S.P. (2003). Redes neuronales artificiales: Fundamentos, modelos y aplicaciones. Universidad Militar Nueva Granada. Facultad de Ingeniería Mecatronica. Bogota. Colombia.

Eirinaki, M., Vazirgiannis, M., Web Mining for Web Personalization. Athens University of Economics and Business, 2003.

Felgaer, Pablo E. (2004). Redes bayesianas aplicadas a minería de datos inteligente. Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. UBA. Argentina.

García Martínez, R., Servente, M., Pasquini, D. (2003). Sistemas Inteligentes. Nueva Librería. 67-148.

Grosser, H. (2004). Detección de fraude en telefonía celular usando redes neuronales. Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. UBA. Argentina.

Huysmans, J.,Baesens B., Vanthienen J., Web Usage Mining: A Practical Study.Katholieke Universiteit Leuven, Dept. of Applied Economic Sciences,2003.

Keim, D.A., (2002). Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics, VOL. 7, NO. 1, January-March 2002.

Kerkhofs, J., Vanhoof, K., Pannemans, D., Web Usage Mining on Proxy Servers: A Case Study. Limburg University Centre, 2001.

Kosala, R., Blockeel, H. (2000). Web Mining Research: A Survey. ACM SIGKDD, July 2000.

Lalani, A.S., Data mining of web access logs. School of Computer Science and Information Technology. Royal Melbourne Institute of Technology. Melbourne, Victoria, Australia, 2003

Mannila, H. (1997). Methods and problems in data mining. In Proc. of International Conference on Database Theory, Delphi, Greece.

Mobasher, B., Cooley R., Srivastava, J. (1999). Creating Adaptive Web Sites Through Usage-Based Clustering of URLs.

Piatetski-Shapiro, G., Frawley, W.J., Matheus, C.J. (1991). Knowledge discovery in databases: an overview. AAAI-MIT Press, Menlo Park, California.

Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropulos, C.D., KOINOTITES: A Web Usage Mining Tool for Personalization, 2001

RFC 1413. Identification Protocol. http://www.rfc-editor.org/rfc/rfc1413.txt. Vigente al 19/11/2005.

RFC 2616 - Hypertext Transfer Protocol - HTTP/1.1. http://www.faqs.org/rfcs/rfc2616.html. Vigente al 19/11/2005.

Roy, A. (2000). Artificial Neural Networks - A Science in Trouble. ACM SIGKDD, Jan 2000.

Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDD, Jan 2000.

WCA. Web characterization terminology & definitions. http://www.w3.org/1999/05/WCA-terms/. Vigente al 19/11/2005

Paola Britos received the B.S. in Information Systems from National University of Luján in 1999 and M.S. degree in Knowledge Engineering from Polytechnic University of Madrid in 2001. She is Senior Researcher at the Intelligent Systems Laboratory of the School of Engineering of the University of Buenos Aires and Associate Professor of the Software and Knowledge Engineering Center of the Graduate School of the Buenos Aires Institute of Technology.

Damian Martinelli received the B.Eng. in Information Systems from University of Buenos Aires in 2007. He is Postgraduate Student at the Intelligent Systems Laboratory of the School of Engineering of the University of Buenos Aires.

.

Hernán Merlino received the B.S. in Information Systems from University of Belgrano in 1999 and M.S. degree in Software Engineering from Buenos Aires Institute of Technology in 2006. He is Assistant Researcher at the Intelligent Systems Laboratory of the School of Engineering of the University of Buenos Aires and Assistant Professor of the Software and Knowledge Engineering Center of the Graduate School of the Buenos Aires Institute of Technology.

Ramón García-Martínez received the B.S. in Computer Science from National University of La Plata in 1988 and M.S. and Ph.D. degrees in Computer Science from Polytechnic University of Madrid in 1992 and 1997, respectively. He is Founder Director of the Intelligent Systems Laboratory of the School of Engineering of the University of Buenos Aires and since 2000 he is Director of the Software and Knowledge Engineering Center of the Graduate School of the Buenos Aires Institute of Technology.