Using Self Organizing Feature Maps to acquire knowledge about visitor behavior in a web site

Juan D. Velásquez, Hiroshi Yasuda, Terumasa Aoki[1]
, Richard Weber[2], and Eduardo Vera[3]

[1] Research Center for Advanced Science and Technology, University of Tokyo
{jvelasqu,yasuda,aoki}@mpeg.rcast.u-tokyo.ac.jp
[2] Department of Industrial Engineering,
University of Chile, rweber@dii.uchile.cl
[3] AccessNova Program, Department of Computer Science,
University of Chile, esvera@accessnova.cl

Abstract. When a user visits a web site, important information concerning his/her preferences and behavior is stored implicitly in the associated log files. This information can be revealed by using data mining techniques and can be used in order to improve both, content and structure of the respective web site.
From the set of possible that define the visitor's behavior, two have been selected: the visited pages and the time spent in each one of them. With this information, a new distance was defined and used in a self organizing map which identifies clusters of similar sessions, allowing the analysis of visitors behavior.
The proposed methodology has been applied to the log files from a certain web site. The respective results gave very important insights regarding visitors behavior and preferences and prompted the reconfiguration of the web site.

1 Introduction

When a visitor enters a web site, the selected pages have direct relation with the desired information he/she is looking for. The ideal structure of a web site should support the visitors in finding such information.

However, reality is quite different. In many cases, the structure of a Web site does not help to find the desired information, although a page that contains it, does exist [3]. Studying visitors behavior is important in order to create more attractive contents, to predict her/his preferences and to prepare links with suggestions, among others [9]. These research initiatives aim at facilitating web site navigation, and in the case of commercial sites, at increasing market shares [1], transforming visitors into customers, increasing customers loyalty and predicting their preferences.

Each click of a web site visitor is stored in files, known as web logs [7]. The knowledge about visitors behavior contained in these files can be extracted using data mining techniques such as e.g. self-organizing feature maps (SOFM).

In this work, a new distance measure between web pages is introduced which is used as input for a specially developed self-organizing feature map that identifies clusters of different sessions. This way, behavior of a web sites visitors can be analyzed and employed for web site improvement.

The special characteristic of the SOFM is its thoroidal topology, which has shown already its advantages when it comes to maintain the continuity of clusters [10].

In section 2, a technique to compare user sessions in a web site is introduced. Section 3 shows how the user behavior vector and the distance between web page is used as input for self-organizing feature maps in order to cluster sessions. Section 4 presents the application of the suggested methodology for a particular web site. Finally, section 5 concludes the present work and points at extensions.

2 Comparing user sessions in a web site

2.1 User Behavior Vector based on Web Site Visits

We define two variables of interest: the set of pages visited by the user and the time spent on each one of them. This information can be obtained from the log files of the web site, which are preprocessed using the sessionization process [5, 6].

Definition 1. *User Behavior Vector.* $\mathbf{U} = \{u(1), \ldots, u(V)\}$
where $u(i) = (u_p(i), u_t(i))$, and $u_p(i)$ is the web page that the user visits in the event i of his session. $u_t(i)$ is the time the user spent visiting the web page. V is the number of pages visited in a certain session.

Figure 1 shows a common structure of a web site. If we have a user visiting the pages 1,3,6,11 and spending 3, 40, 5, 16 seconds respectively, the corresponding user vector is: U = ((1,3),(3,40),(6,5),(11,16))

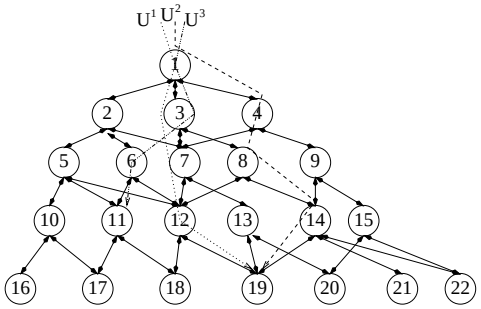

Fig. 1. A common structure of web site and the representation of user behavior vectors

After completing the sessionization step, we have the pages visited and the time spent in each one of them, except for the last visited page, of which we

only know when its visit began. An approximation for this value is to take the average of the time spent in the other pages visited in the same session.

Additionally, it is necessary to consider that the number of web pages visited by different users varies. Thus the numbers of components in the respective user behavior vectors differ. However, we can introduce a modification in order to create vectors with the same number of components.

Let L be the maximum number of components in a vector, and U a vector with S components so that $S \leq L$. Then the modified vector is:

$$U' = \begin{cases} (u_p(k), u_t(k)) & 1 \leq k \leq S \\ (0,0) & S < k \leq L \end{cases} \tag{1}$$

2.2 Preprocessing

A web page contains a variety of tags and words that do not have direct relation with the content of the page we want to study. Therefore we have to filter the text and to eliminate the following types of words: HTML Tags, Stopwords (i.e. pronouns, prepositions, conjunctions, etc.) and Word stem (suffix removal).

Next the web page is a document represented by a vector space model [2], in particular by vectors of words. Let $P = \{p_1, \ldots, p_Q\}$ be the set of Web pages in a web site. Its vectorial representation would be a matrix of QxR, where Q is the number of pages in the web site and R is the number of different words in P.

Then a matrix M that contains the vectors of words in its columns is:

$$M = (m_{ij}) \quad i = 1, \ldots, Q \quad and \quad j = 1, \ldots, R \tag{2}$$

where m_{ij} is the weight of word i in document j. In order to estimate these weights, we use the *tfxidf-weighting* [2], defined by equation 3.

$$m_{ij} = f_{ij} * \log(\frac{Q}{n_i}) \tag{3}$$

where f_{ij} is the number of occurrences of word i in document j and n_i is the total number of times that word i appears in the whole collection of documents.

2.3 Distance measure between two pages

With the above given definitions we can use vectorial linear algebra, in order to define a distance measure between two web pages.

Definition 2. *Word Pages Vectors.* $\mathbf{WP^i} = (wp_1^i, \ldots, wp_R^i) = (m_{i1}, \ldots, m_{iR})$

Thus the distance between page vectors [4] is:

$$dp(WP^i, WP^j) = \frac{\sum_{k=1}^{R} wp_k^i wp_k^j}{\sqrt{\sum_{k=1}^{R}(wp_k^i)^2}\sqrt{\sum_{k=1}^{R}(wp_k^j)^2}} \tag{4}$$

Definition 3. *Page Distance Vector.*
$\mathbf{D_{AB}} = (dp(a_1, b_1), \ldots, dp(a_m, b_m))$ where $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_m\}$ are sets of word page vectors with the same cardinality.

2.4 Comparison between two user behavior vectors

In order to compare two user sessions it is necessary to define a measure that determines the difference between two user behavior vectors based on both characteristics of the user behavior vector (time and page content), see equation 5.

$$dub(U^i, U^j) = \sum_{k=1}^{L} \min\{\frac{u_t^i(k)}{u_t^j(k)}, \frac{u_t^j(k)}{u_t^i(k)}\} * dp(u_p^i(k), u_p^j(k)) \quad (5)$$

with dp distance measure between the content of two pages.

We use dp (equation 4) because it is possible that two users visit different web pages in the web site, but the content is similar.

This is a variation of the approach proposed in [7], where only the user's path was considered but not the content of each page.

The second element of equation 5, $\min\{\frac{u_t^i(k)}{u_t^j(k)}, \frac{u_t^j(k)}{u_t^i(k)}\}$ is indicating the user's interest for the pages visited. The assumption is that the time spent on a page is proportional to the interest the user has in its content. In this way, if the times spent are closed, the value of the expression will be near 1. In the opposite case, it will be near 0.

The final expression of equation 5 combines the content of the visited pages with the time spent on each of the pages by a multiplication. This way we can distinguish between two users who had visited similar pages but spent different times on each of them. Similarly we can separate between users that spent the same time visiting pages with different content and position in the web.

3 Self Organizing Feature Map for Session Clustering

We used an artificial neural network of the Kohonen type (Self-organizing Feature Map; SOFM) [8]. Schematically, it is presented as a two-dimensional array in whose positions the neurons are located. Each neuron is constituted by an n-dimensional vector, whose components are the synaptic weights. By construction, all the neurons receive the same input at a given moment.

The idea in this learning process is to present an example to the network and by using a metric, to search the neuron in the network most similar to the example (center of excitation, winner neuron). Next we have to modify its weights and those of the center's neighbors.

The notion of neighborhood among the neurons provides diverse topologies. In this case the thoroidal topology was used [10], which means that the neurons closest to the ones of the superior edge, are located in the inferior and lateral edges (see figure 2)

The U vectors have two components (time and content) for each web page. Therefore it is necessary to modify both when the neural network changes the weights for the winner neuron and its neighbors.

The time component of the U vector is modified with a numerical adjustment, but the page component needs a different updating scheme [8]. In the

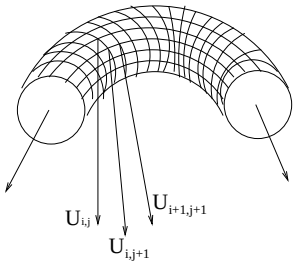

Fig. 2. Proximity of the User Behavior Vector in a network of thoroidal Kohonen

preprocessing step, we constructed a matrix with the pairwise distance among all pages in the web site. Using this information we can adjust the respective weights. Let N be a neuron in the network and E the user behavior vector example presented to the network, using definition 3, the page distance vector is:

$$D_{NE} = (d_p(N_p(1), E_p(1)), \ldots, d_p(N_p(M), E_p(M))) \tag{6}$$

Now the adjustment is over the D_{NE} vector, i.e., we have $D'_{NE} = D_{NE} * f_p$, with f_p an adjustment factor. Using D'_{NE}, it will be necessary to find a set of pages whose distances with N be near to D'_{NE}. Thus the final adjustment for the winner and its neighbor neurons is given by equation 7.

$$N^{n+1} = \left(N_t^n(i) * f_t, p \in P / D'_{NE}(i) \approx d_p(p, N_p^n(i))\right) \tag{7}$$

with $i = 1, \ldots, L$.

4 Application of the proposed methodology

4.1 Selecting the web site

In order to prove the effectiveness of the tools developed in this work, a web site was selected[4]. It contains information about programs of specialization for professionals and belongs to the Department of Industrial Engineering of the University of Chile.

This site is written in Spanish, has 142 static web pages and for this study approximately 24,000 web logs registers were considered, corresponding to the period August to October, 2002.

4.2 Preprocessing and Indexing

In the preprocessing step, the grammar particles (articles, prepositions, conjunctions, etc.), the characters with accent and html tags were eliminated. Additionally, the word stemming process was applied.

[4] http://www.dii.uchile.cl/~diplomas/

The links pages, i.e., pages that only contain links to other web pages, were not considered in the analysis. The total number of pages is 122.

Next we created the matrix with the distances among page vectors. The dimension of this matrix is 6234x122, i.e., R=6234 and Q=122.

4.3 Sessionization and User Behavior Vector

We used the time-based heuristic for the sessionization process and considered 30 minutes as the longest user session. Only 7% of the users have sessions with 7 or more pages visited and 11% visited at least 3 pages. Then we supposed three and six as minimum and maximum number of components in a user behavior vector, respectively. Using these filters, we identified 4113 user behavior vectors.

4.4 Training the Neural Network

We used a SOFM with 6 input neurons (corresponding to the six pages in a visit) and 256 output neurons. Using this structure, we could map the 4113 user behavior vectors to the 256 neurons of the feature map.

The thoroidal topology maintains the continuity in clusters, which allows to study the transition among the preferences of the users from one cluster to another .

The training of the neural network was carried out on a computer pentium IV, with 512 Mb in RAM and running Linux OS, distribution Redhat 7.1. The time necessary was 2,5 hours and the epoch parameter was 100.

4.5 Results

We identified six main clusters as shown in the following table. The second and third column of table 1, contain the center neurons of each of the clusters, representing the visited pages and the time spent in each one of them.

Cluster	Pages Visited	Time spent in seconds
1	(2,15,60,42,70,62)	(3,5,113,67,87,43)
2	(5,43,65,75,112,1)	(4,53,40,63,107,10)
3	(6,47,67,7,48,112)	(4,61,35,5,65,97)
4	(10,51,118,87,105,1)	(5,80,121,108,30,5)
5	(11,55,37,87,114,12)	(3,75,31,43,76,8)
6	(13,57,41,98,120,107)	(4,105,84,63,107,30)

Table 1. User behavior clusters

The pages in the web site were labelled with a number to facilitate its analysis. Table 2 shows the main content of each page.

Pages	Contain
1	Home page
$2, \ldots, 14$	Main page about a course
$15, \ldots, 28$	Presentation of the program
$29, \ldots, 41$	Objectives
$42, \ldots, 58$	Program: Course's modules
$59, \ldots, 61$	Student profile
$62, \ldots, 68$	Schedule and dedication
$69, \ldots, 91$	Curriculums of the staff of instructors
$92, \ldots, 108$	Menu to solicited information
$108, \ldots, 121$	Information:cost, schedule, postulation, etc.
122	News page

Table 2. Pages and their content

The clusters analysis show:

- Cluster 1. The users are interested in the profile of the students, the program and the faculty staff's curriculums.
- Cluster 2 and 3. The users show preferences for courses about environmental topics and visit the page where they can ask for information. In both clusters, program and schedule are very important for the user.
- Cluster 4. This cluster contains sessions of users interested in new courses.
- Cluster 5. The users are interested in the students profile and the course objectives.
- Cluster 6. In this case sessions are similar to the sessions in cluster 5. This kind of course is seminar.

Reviewing the clusters found, it can be inferred that the users show interest in the profile of the students, the schedules and contents of the courses and the professors who are in charge of each subject. Based on our analysis we propose to change the structure of the Web site, privileging the described information. The following step is to do an analysis of the content of the pages using the distance defined in the equation 5.

5 Conclusions

In this work we introduced a methodology to study the user behavior in a web site. In the first part we propose a way to study the user behavior in the web, using a new distance measure based on two characteristics derived from the user sessions: pages visited and time spent in each one of them. Using this distance in a self organizing map, we found clusters from users sessions, which allow us to study the user behavior in the particular web site.

The experiments made with data from a certain Web site, showed that the methodology used allows to create clusters of user sessions, and - using this information - to study the user behavior in the web site.

Since the distance considers the content and position of the page in the web site, its structure and the words used in the pages are variables that influence directly in the capacity of the SOFM to create clusters.

The distance introduced, is very useful to increase the knowledge about the user behavior in a web site. As future work, it is proposed to improve the presented methodology introducing new variables derived from user sessions. It will also be necessary to continue applying our methodology to other web sites in order to get new hints on future developments.

References

1. S. Araya , M. Silva and R. Weber, Identifying web usage behavior of bank customers. *Proceedings of SPIE, Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV*, Vol. 4730, pages 245-251 April, 1-5,Orlando, USA 2002
2. R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, chapter 2. *Addison-Wesley* 1999
3. N.J. Belkin, Helping people find what they dont know *Communications of the ACM*, Vol. 43(8), pages 58-61, 2000
4. M. W. Berry, S. T. Dumais and G.W. O'Brien, Using linear algebra for intelligent information retrieval, *SIAM Review*, Vol. 37, pages 573-595, December 1995
5. R. Cooley, B. Mobasher, J. Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowlegde and Information Systems* Vol. 1, pages 5-32, 1999
6. R. Cooley, B. Mobasher and J. Srivastava, Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns, *In Knowledge and Data Engineering Workshop*, pages 2-9, Newport Beach, CA, 1997
7. A. Joshi and R. Krishnapuram, On Mining Web Access Logs. *In Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 63-69, 2000.
8. T. Kohonen, Self-Organization and Associative Memory, *Springer-Verlag* 1987, 2nd edition.
9. B. Mobasher, R. Cooley and J. Srivastava, Creating Adaptive Web Sites Through Usage-Based Clustering of URLs, *Proceedings of IEEE Knowledge and Data Engineering Exchange*, November, 1999.
10. J. Velásquez, H. Yasuda, T. Aoki and R. Weber, Voice Codification using Self Organizing Maps as Data Mining Tool. *Proceedings of Second International Conference on Hybrid Intelligent Systems* , pages 480-489, Santiago, Chile, December, 2002