PAPER Special Issue on Information Processing Technology for web utilization

A new similarity measure to understand visitor behavior in a web site

Juan VELÁSQUEZ[†], *Nonmember*, Hiroshi YASUDA[†], Terumasa AOKI[†], *Regular Members*, *and* Richard WEBER[††], *Nonmember*

SUMMARY

The behavior of visitors browsing in a web site offers a lot of information about their requirements and the way they use the respective site.

Analyzing such behavior can provide the necessary information in order to improve the web site's structure. The literature contains already several suggestions on how to characterize web site usage and to identify the respective visitor requirements based on clustering of visitor sessions.

Here we propose to combine visitor behavior with the content of the respective web pages and the similarity between different page sequences in order to define a similarity measure between different visits. This similarity serves as input for clustering of visitor sessions.

The application of our approach to a bank's web site and its visitor sessions shows its potential for internet-based businesses.

key words: *web mining, browsing behavior, similarity measure, clustering.*

1. Introduction

Analyzing visitor browsing behavior in a web site can be the key for improving both, contents and structure of the site and this way assures the institutionals successful participation in Internet.

Web log files contain information about the visitors interaction with the respective web site. Depending on the traffic, these files can contain millions of registers with a lot of irrelevant information each, such that its analysis becomes a complex task [7].

Applying web usage mining techniques [12], allows to discover interesting pattern about the visitor behavior. Complemented with semantic web mining, results can be improved [4].

Here we propose a new similarity measure between different visitor sessions based on usage and content of the web site. In particular, this measure uses the following three variables, which we determine for all visitors: content of each visited web page, time spent on it, and the sequence of visited pages.

We proved the effectiveness of the proposed similarity measure applying self-organizing feature maps (SOFM) for session clustering. Any other unsupervised clustering method could be used as well for this purpose.

This way, similar visits are grouped together and typical visitor behavior can be identified, which gives way to improvements of web sites and better understanding of visitor behavior.

The special characteristic of the SOFM is its thoroidal topology, which has shown its advantages when it comes to maintain the continuity of clusters [16] or when the data correspond to a sequence of events, like e.g. voice patterns [17]. In the case of visitor behavior we have a similar situation.

Section 2 of this paper provides an overview on related work. In section 3 we describe the data preparation process, which is necessary for comparison of visitor sessions (section 4). In section 5 we show how a self-organizing feature map was used for session clustering using the previously introduced similarity measure. Section 6 describes the application of our work to the case of a Chilean bank. Section 7 concludes this work and points at future work.

2. Related Work

2.1 Overview on Web Mining

In order to understand the visitor behavior in the web, we will use web mining techniques. They aim at finding useful information from the World Wide Web (WWW). This task is not trivial, considering that the web is a huge collection of heterogeneous, unlabelled, distributed, time variant, semi-structured and high dimensional data [12]. Therefore, a data preparation process is necessary previous to any analysis.

The web mining techniques, can be categorized in three areas: Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM), see e.g. [4], [12] for a short description.

2.2 Analyzing the Web using cluster algorithms

The main idea of clustering is to identify classes of ob-

Manuscript received May 30, 2003.

Manuscript revised 0, 2003.

[†]E-mail:{jvelasqu,yasuda,aoki}@mpeg.rcast.u-tokyo.ac.jp, Research Center for Advanced Science and Technology, University of Tokyo, 3th Building, 4-6-1 Komaba, Meguro-Ku Tokyo, Japan P.C. 153-8904.

[††]E-mail: rweber@dii.uchile.cl, Department of Industrial Engineering, University of Chile, República 701,Santiago, Chile

jects (clusters) that are homogeneous within each class and heterogeneous between different classes. Therefore it is necessary to have a measure to determine similarity between objects.

Let Ω be a set of m vectors $\omega_i \in \Re^n$, i = 1, ..., m. The clustering goal is to partition Ω in K groups, where C_k denotes the k^{th} cluster. Then, $\omega_i \in C_l$ means that ω_i is more similar to the elements in this cluster than to objects from other clusters.

Clustering requires a similarity measure, $\zeta(\omega_p, \omega_q)$ in order to compare two vectors from Ω and the way to determine the number of clusters depends on the method used. In the case of supervised classification, the value K is set a priori. In unsupervised clustering, such as e.g. Self-organizing Feature Maps the cluster number is determined by the method itself.

Recent studies [7], [10] [13], [16], underline the benefits of clustering for visitor behavior analysis.

2.3 Web content and usage mining proposition

It aims at combining the philosophies behind WCM and WUM through providing a complemented vision of both techniques [10].

Applying WUM [3] we can understand the visitor browsing behavior, but we cannot discover which content is interesting for the visitor. This analysis is possible using WCM [4].

The first step of our suggestion is the definition of a similarity measure that allows to compare behaviors of different visitors, through the analysis of visitor preferences. This measure is based on the content of the visited pages, the time a visitor spent on each page and the sequence of pages in his/her session.

As a second step we propose to use this measure in a cluster algorithm in order to find groups of similar visitor sessions and using this information, make prediction about the preferences of the future web site's visitors.

3. Data preparation process

We propose to use two kinds of data sources that are easily available: web log registers and web pages.

The web log registers contain information about the browsing behavior of the web site visitors, in particular the page navigation sequence and the time spent in each page visited. The data from web logs is based on the structure given by W3C[†].

The second data source is the web site itself. Each web page is defined by its content. For our similarity measure we need only the set of words from each page.

In order to study the visitor's behavior, it is necessary to prepare the data from both sources, i.e., web logs as well as web pages, using filters and identifying the real user sessions.

3.1 Web log data preparation process

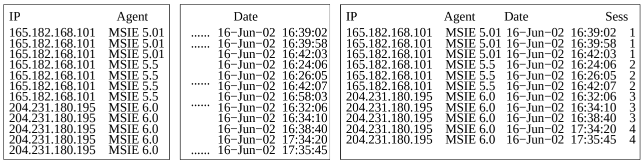
A web log file contains information on the access of all visitors to a particular web site in chronological order. In a common log file each access to one of its pages is stored together with the following information:

- IP address and agent.
- Time stamp.
- Embedded session Ids.
- Method.
- Status.
- Software Agents.
- Bytes transmitted.
- Objects required (page, pictures, etc).

Based on such log files we have to determine for each visitor, the sequence of web pages visited in his/her session. This process is known as **sessionization** [5]. It considers a maximum time duration given by a parameter, which is usually 30 minutes in the case of total session time. Based on this parameter we can identify the transactions that belong to a specific session using tables and program filters. Figure 1 shows the transformation sequence for web log registers.

This process can be realized using data streams and program code like **perl** in unix systems. The first stream contains the log registers grouped by IP and agents, which are separated by a space (see figure 1; left column). We only consider registers whose code is not error and whose URL parameter link to web page objects. Other objects, such as pictures, sound, links, etc. are not considered.

In the second column of figure 1, the registers are ordered by date, maintaining the groups. Finally, we select registers from a time window of 30 minutes that are grouped together in sessions in the third stream (third column in figure 1).



IP	Agent
165.182.168.101	MSIE 5.01
165.182.168.101	MSIE 5.01
165.182.168.101	MSIE 5.01
165.182.168.101	MSIE 5.5
165.182.168.101	MSIE 5.5
165.182.168.101	MSIE 5.5
165.182.168.101	MSIE 5.5
204.231.180.195	MSIE 6.0
204.231.180.195	MSIE 6.0
204.231.180.195	MSIE 6.0
204.231.180.195	MSIE 6.0
204.231.180.195	MSIE 6.0

	Date	
......	16−Jun−02	16:39:02
......	16−Jun−02	16:39:58
	16−Jun−02	16:42:03
	16−Jun−02	16:24:06
	16−Jun−02	16:26:05
......	16−Jun−02	16:42:07
	16−Jun−02	16:58:03
......	16−Jun−02	16:32:06
	16−Jun−02	16:34:10
	16−Jun−02	16:38:40
	16−Jun−02	17:34:20
......	16−Jun−02	17:35:45

IP	Agent	Date		Sess
165.182.168.101	MSIE 5.01	16−Jun−02	16:39:02	1
165.182.168.101	MSIE 5.01	16−Jun−02	16:39:58	1
165.182.168.101	MSIE 5.01	16−Jun−02	16:42:03	1
165.182.168.101	MSIE 5.5	16−Jun−02	16:24:06	2
165.182.168.101	MSIE 5.5	16−Jun−02	16:26:05	2
165.182.168.101	MSIE 5.5	16−Jun−02	16:42:07	2
204.231.180.195	MSIE 6.0	16−Jun−02	16:32:06	3
204.231.180.195	MSIE 6.0	16−Jun−02	16:34:10	3
204.231.180.195	MSIE 6.0	16−Jun−02	16:38:40	3
204.231.180.195	MSIE 6.0	16−Jun−02	17:34:20	4
204.231.180.195	MSIE 6.0	16−Jun−02	17:35:45	4

Fig. 1 Sessionization process

3.2 Web page content preparation process

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 eliminating the following types of words:

[†] Web logs are delimited text files as specified by RFC 2616, "Hypertext Transfer Protocol – HTTP/1.1" http://www.rfc-editor.org/rfc/rfc2616.txt

- HTML Tags. Some tags show interesting information about the page content, for instance, the <title> tags mark the web page central theme. In this case, we used this information to identify special words inside the text.
- Stop words (e.g. pronouns, prepositions, conjunctions, etc.)
- Word stemming. Process of suffix removal, to generate word stems [2].

After filtering, we represent a document (web site) by a vector space model [1], in particular by vectors of words.

Let R be the number of different words in a web site and Q be the number of its pages. A vectorial representation of the web site would then be a matrix M of dimension RxQ with:

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

where m_{ij} is the weight of word i in page j.

We propose to estimate these weights using equation 2, which is based on the *tfxidf-weighting* [1].

$$m_{ij} = f_{ij}(1 + \frac{sw(i)}{TR}) * \log(\frac{R}{n_i}) \tag{2}$$

f_{ij} is the number of occurrences of word i in page j and n_i is the total number of times word i appears in the entire web site. Additionally, we propose to augment word importance if a user searches for a specific word. This is done by sw (special words), which is an array with dimension R. It contains in component i the number of times that a user searches word i during a given period (e.g.: one month). TR is the total number of times that a user searches words in the Web site during the same period. If TR is zero, $\frac{sw(i)}{TR}$ is defined as zero, i.e. if there has not been any word searching, the weight m_{ij} depends just on the number of occurrences of words.

3.2.1 Distance measure between two pages

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

Definition 1 (Word Page Vector):
$\mathbf{WP^k} = (wp^k_1, \ldots, wp^k_R) = (m_{1k}, \ldots, m_{Rk})$ k = 1,...,Q.

Based on this definition, we used the angle's cosine as similarity measure between two page vectors [2]:

$$dp(WP^i, WP^j) = \frac{\sum_{k=1}^R wp^i_k wp^j_k}{\sqrt{\sum_{k=1}^R (wp^i_k)^2}\sqrt{\sum_{k=1}^R (wp^j_k)^2}} \tag{3}$$

We define $dp_{ij} = dp(WP^i, WP^j)$ as the similarity between page i and page j of the web site.

3.3 Navigation sequence preparation process

The navigation sequence can be illustrated by a graph G as shown in figure 2, where each edge (web page) is represented by an identification number. Let $E(G)$ be the set of edges in graph G. Figure 2 shows the structure of a simple web site. If we suppose two visitors visiting the site, the respective sub-graphs can be $G_1 = \{1 \to 2, 2 \to 6, 2 \to 5, 5 \to 8\}$ and $G_2 = \{1 \to 3, 3 \to 6, 3 \to 7\}$. Then $E(G_1) = \{1, 2, 5, 6, 8\}$ and $E(G_2) = \{1, 3, 6, 7\}$ with $\parallel E(G_1) \parallel = 5$ and $\parallel E(G_2) \parallel = 4$, respectively. In this example, page 4 has not been visited.

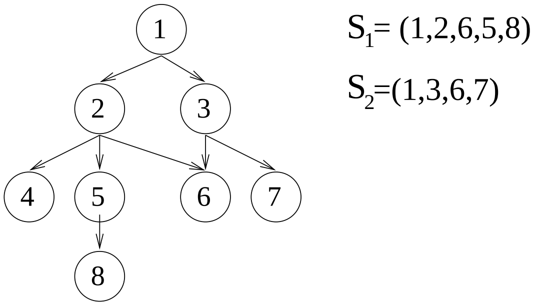


Fig. 2 A web site with two navigation sequences

From these graphs we can determine the visitors navigation sequence $S_1 = S(G_1) = (1, 2, 6, 5, 8)$ and $S_2 = S(G_2) = (1, 3, 6, 7)$, respectively. They represent the pages visited and the navigation sequence.

3.3.1 Comparing navigation sequences

For the similarity of visits we propose in this paper we need the similarity of the respective navigation sequences. Therefore it is necessary to define a measure that considers how much two sub-sequences have in common. Equation 4 introduces a simple way to compare two navigation sequences [14].

$$dG(G_1, G_2) = 2\frac{\parallel E(G_1) \cap E(G_2) \parallel}{\parallel E(G_1) \parallel + \parallel E(G_2) \parallel} \tag{4}$$

Notice that $dG \in [0, 1]$ and if $G_1 = G_2$, we have $dG(G_1, G_2) = 1$. In the case of disjoint graphs, $dG(G_1, G_2) = 0$ because $\parallel E(G_1) \cap E(G_2) \parallel = 0$.

The problem of equation 4 is that its nominator does not take into consideration the sequence of visited pages in each sub-graph; it just looks at the set of visited pages.

If we want to consider how similar two sequences are, we have to compare also the respective sequences. i.e., instead of calculating $\parallel E(G_1) \cap E(G_2) \parallel$ we have to determine how similar two sequences are, considering the order in which the nodes are visited.

For example, both sequences in figure 2 can be represented by a string of tokens [14] such as $S_1 =$ "12658" and $S_2 =$ "1367". We need to know how similar or different are both sequences in its string representation.

For this purpose we use the *Levenshtein distance* [9], also known as *edit distance*. It determines the number of transformations necessary to convert S_1 in S_2. This number can be used as dissimilarity measure for the two sequences.

For two sequences, $\hat{x_p} = (x_1, \ldots, x_p)$ and $\hat{y_q} = (y_1, \ldots, y_q)$, the Levenshtein distance is defined as:

$$L(\hat{x_p}, \hat{y_p}) = \begin{cases} p & q = 0 \\ q & p = 0 \\ \min\{L(x_{\hat{p-1}}, \hat{y_q}) + 1, & else \\ L(\hat{x_p}, y_{\hat{q-1}}) + 1, & \\ L(x_{\hat{p-1}}, y_{\hat{q-1}}) + z(x_p, y_q)\} & \end{cases} \qquad (5)$$

where

$$z(i,j) = \begin{cases} 0 & if \quad i = j \\ 1 & otherwise \end{cases} \qquad (6)$$

Analyzing the definition of the Levenshtein distance reveals the importance of the order of the token to be compared. For instance, if we compare S_1 and S_2, three transformations will be necessary. If we have e.g. $S_3 =$ "12856" and $S_4 =$ "1367" four transformations will be necessary. This characteristic makes the Levenshtein distance very suitable to compare two navigation sequences.

$$dG(G_1, G_2) = 1 - 2\frac{\text{Ł}(S_1, S_2)}{\parallel E(G_1) \parallel + \parallel E(G_2) \parallel} \qquad (7)$$

Using the example in figure 2 and equation 7 leads to $dG(G_1, G_2) = 0.\bar{6}$.

4. Comparing visitor sessions in a web site

We propose a model to represent visitor behavior using three variables: the sequence of visited pages, their content and the time spent in each one of them. The model is based on a visitor behavior vector with dimension n and two parameters in each component.

Definition 2 (Visitor Behavior Vector):
$\upsilon = [(p_1, t_1) \ldots (p_n, t_n)]$, being (p_i, t_i) parameters that represent the i^{th} page from a visit and the time spent on it, respectively. In this expression, p_i is the page identifier.

For instance, using the browsing navigation shown in figure 2, we have $\upsilon_1 = [(1,3), (2,40), (6,5), (5,16), (8,15)]$.

Let α, β be two visitor behavior vectors with cardinality C^α and C^β respectively and $\Gamma(\cdot)$ a function that applied over α or β returns the respective navigation sequence.

The proposed similarity measure is introduced in equation 8 as:

$$sm(\alpha, \beta) = dG(\Gamma(\alpha), \Gamma(\beta))\frac{1}{\eta}\sum_{k=1}^{\eta} \tau_k * dp(p^h_{\alpha,k}, p^h_{\beta,k}) \qquad (8)$$

where $\eta = \min\{C^\alpha, C^\beta\}$, and $dp(p^h_{\alpha,k}, p^h_{\beta,k})$ is the similarity between the k^{th}page of vector α and the k^{th}page of vector β.

The first element of equation 8 is the sequence similarity introduced in equation 7.

As second element, $\tau_k = \min\{\frac{t^\alpha_k}{t^\beta_k}, \frac{t^\beta_k}{t^\alpha_k}\}$ is indicating the visitor's interest in the pages visited. We assume that the time spent on a page is proportional to the interest the visitor has in its content. If the times by visitors α and β on the k^{th} page visited (t^α_k, t^β_k, respectively) are close to each other, the value of the expression will be close to **1**. In the opposite case, it will be close to **0**.

The third element, dp, measures the similarity of the pages visited. It is possible that two users visit different web pages in the web site, but the content is similar, e.g., one page contains information about classic rock and another one about progressive rock. In both cases the users have interest in music, specifically in rock. This is a variation compared to the approach proposed in [7], where only the user's path was considered but not the content of each page.

Finally, we combine in equation 8 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 visitors that spent the same time visiting pages with different content in the web.

5. Self Organizing Map for Session Clustering

We used an artificial neural network of the Kohonen type (Self-organizing Feature Map; SOFM), in order to mine the visitor behavior vectors and discover knowledge [11] about the visitor preferences. 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 of this learning process is to present an example to the network and, by using a metric, to determine the neuron in the network most similar to the presented example (center of excitation, winner neuron). Next, we have to modify its weights and those of the center's neighbors.

This type of learning is **unsupervised**, because the neurons "move" towards the centers of the groups of examples that they try to represent. The described process was proposed by Kohonen, in his *Algorithm of training of Self Organizing networks* [8].

5.1 Operation of the Self-Organizing Feature Map

The notion of neighborhood among the neurons provides diverse topologies. In this case the thoroidal topology was used [16], [17], which means that the neurons closest to the ones of the superior edge, are located in the inferior and lateral edges (see figure 3)

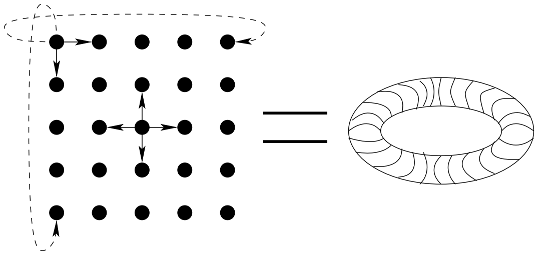


Fig. 3 Neighborhood of neurons in a thoroidal Kohonen network

The SOFM's operation needs vectors with the same cardinality. It is a configurable parameter in the creation of the visitor behavior vector. Let H be the number of elements in each vector. If a visitor session does not have at least H elements, it is not considered in our analysis. On the other hand, we only consider up to the H^{th} component of a session for the visitor behavior vector.

Since the visitor behavior vectors have two components for the visited pages (page identifier and time spent on each page), it is necessary to modify both when the neural network changes the weights for the winner neuron and its neighbors.

Let N be a neuron in the network and E the visitor behavior vector example presented to the network. The vector's time component is modified with a numerical adjustment, i.e., $t_{N,i+1} = t_{N,i} * f_t$ with $i = 1, \ldots, H$.

The page component needs another updating scheme. Using the page distance, the difference between page content components is shown in equation 9.

$$D_{NE} = [dp(p^h_{N,1}, p^h_{E,1}), \ldots, dp(p^h_{N,H}, p^h_{E,H})] \qquad (9)$$

Equation 9 represents a vector with distance between pages, i.e., its components are numeric values. Then the adjustment is over the D_{NE} expression, i.e., we have $D'_{NE} = D_{NE} * f_\rho$, with f_ρ adjustment factor. Using D'_{NE}, it will be necessary to find a set of pages whose distances with N are close to D'_{NE}. Thus the final adjustment for page component of the winner neuron and its neighbor neurons is given by equation 10.

$$p_{N,i+1} = \pi \in \Pi / D'_{NE,i} \approx dp(\pi, p^h_{N,i}) \qquad (10)$$

with $\Pi = \{\pi_1, \ldots, \pi_Q\}$ the set of all pages in the web site, $D'_{NE,i}$ the i^{th} component of D'_{NE}. Then given $D'_{NE,i}$ we have to find the page π in Π whose $dp(\pi, p^h_{N,i})$ is closest to $D'_{NE,i}$.

6. Practical application

In order to prove the effectiveness of the tools developed in this work, a web site was selected, considering the following characteristics [13]:

- It includes many pages with different information.
- Each visitor has interest in some pages, but is not interested in others.
- The web site is maintained by a web master observing pages of interest, i.e., if a page is not visited it will be removed from the site.

The web site selected[†], is about the first Chilean virtual bank, i.e., it does not have physical branches and all the transactions are made using electronic means, like e-mails, portals, etc.

We have the following information about the web site:

- Written in Spanish.
- 217 static web pages.
- Approximately eight million web log registers from the period January to March 2003.

6.1 Sessionization process

This task was implemented by a **perl** code and considers 30 minutes as the longest user session. In order to clean very short sessions, it is necessary to apply a previous heuristic to the data.

Only 16% of the visitors visit 10 or more pages and 18% less than 4. The average of visited pages is 6. Based on this information, we fix 6 as maximum number of components in a visitor behavior vector, i.e. the parameter $H = 6$.

Vector with more than 6 components are considered only up to the sixth component. Vector with less than 4 components are not considered in our analysis.

Finally, applying the above described filters, approximately 400,000 visitor behavior vectors were identified.

6.2 Web page content processing

Applying web page text filters, we found 710 different words for our analysis in the complete web site (i.e. R= 710).

Regarding word weights, especially the special words (see equation 2), we applied the following procedure.

[†] http://www.tbanc.cl/

The web site offers visitors the option to send e-mails to the call center platform. Then the text sent is a source to identified the most interesting words. In this case, only 20 special words were used for page vector calculation.

Using the above described data and applying equation 3, a 3-dimensional matrix with the similarity between pairs of pages is created and shown graphically in figure 4.

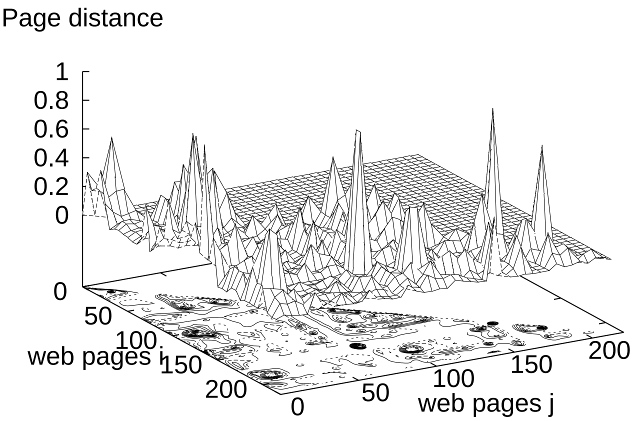


Fig. 4 Similarity measure among web pages

Since the matrix is symmetric, figure 4 shows only its superior side.

6.3 Applying Self-Organizing Feature Maps

The SOFM used has 6 input neurons and 32*32 output neurons in the feature map. The thoroidal topology maintains the continuity in clusters [17], which allows to study the transition among the preferences of the visitor from one cluster to another.

We update a matrix that contains the number of times each neuron wins during the training of the SOFM. Using this information, the clusters found by the SOFM are shown in figure 5. The x and y axes are the neuron's position and the z axis is the neuron's winner frequency, with the scale adapted.

From figure 5, we can identify four main clusters. They are checked using the information contained in the matrix. This allows to find the centroid neurons (winner neuron) of each of the clusters.

6.4 Results

Table 1 shows the clusters found by the SOFM. The second and third column contain the cluster's center, represented by the visited pages and the time spent in each one of them.

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

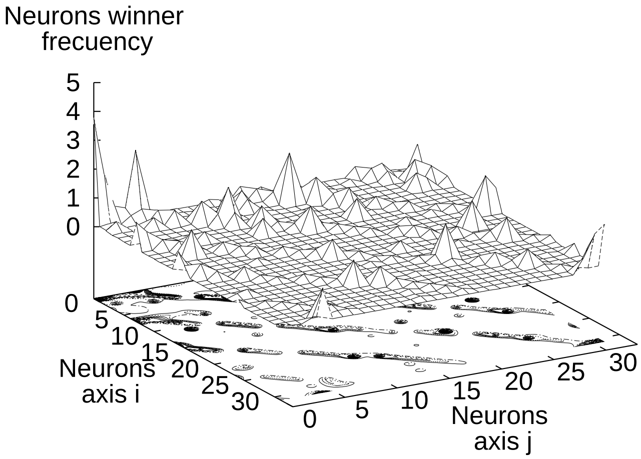


Fig. 5 Clusters among visitor behavior vectors

Table 1 Visitor behavior clusters

Cluster	Pages Visited	Time spent in seconds
1	(1,3,8,9,147,190)	(40,67,175,113,184,43)
2	(100,101,126,128,30,58)	(20,69,40,63,107,10)
3	(70,86,150,186,137,97)	(4,61,35,5,65,97)
4	(157,169,180,101,105,1)	(5,80,121,108,30,5)

Table 2 Pages and their content

Pages	Content
1	Home page
2, . . . , 65	Products and Services
66, . . . , 98	Agreements with other institutions
99, . . . , 115	Remote services
116, . . . , 130	Credit cards
131, . . . , 155	Promotions
156, . . . , 184	Investments
185, . . . , 217	Different kinds of credits

A detailed cluster analysis together with the banks experts and an interpretation of the web pages visited revealed the following results:

- Cluster 1. The visitors are interested in general products and services offered by the bank.
- Cluster 2. The visitors search information about credit cards.
- Cluster 3. Visitors are interested in agreements between the bank and other institutions.
- Cluster 4. Visitors are interested in investments and remote services offered by the bank

Based on our analysis we have proposed changes of the structure of the web site, privileging the described information.

This new information about the visitor behavior, can be used e.g. for navigation suggestions. For instance, we can classify a person visiting the web site based on his/her navigation behavior.

Based on the cluster he or she belongs to we can suggest online other probably interesting pages. This way, the web site effectiveness is improved.

6.5 Using an alternative clustering algorithm

In order to analyze the segmentation found we applied another unsupervised cluster algorithms, the well-known k-means [6]. Its main idea is to assign each vector to a set of given cluster centroids and then update this centroids given the previously established assignment. This procedure is repeated iteratively until a certain stopping criterion is fulfilled.

The number of clusters to be found (k) is a required input value for k-means. Since we found four clusters with the SOFM we used $k = 4$ as input value for k-means. We took a random selection from the original set of training vectors as initial centroids.

Given $c_1^j, \ldots, c_k^j$ as cluster centroids in iteration j, we compute $c_1^{j+1}, \ldots, c_k^{j+1}$ as cluster centroids in iteration $j + 1$, according to the following steps:

1. Cluster assignment. For each vector in the training set, we determine the cluster to which it belongs, using the similarity introduced in this work.
2. Cluster centroid update. Let $V_l^j = \{v_1, \ldots, v_{q_l^j}\}$ be the set of q_l^j vectors associated to centroid c_l^j, with $l = 1, \ldots, k$. We determine the c_l^{j+1}, as a mean of V_l^j,i.e., $v_i \in V_l^j \ / \ \max\{\sum_{j=1}^{q_l^j} sm(v_i, v_j)\}\ , i \neq j$
3. Stop when $c_l^{j+1} \approx c_l^j$.

We apply the k-means over the same set of visitor behavior vectors used in the training of the SOFM. The algorithm converged to the result shown in table 3.

Table 3 k-means visitor behavior clusters

Cluster	Pages Visited	Time spent in seconds
1	(2,29,45,112,120,154)	(20,69,35,126,134,90)
2	(200,135,10,50,132,128)	(3,101,108,130,20,13)
3	(1,100,114,128,141,148)	(4,76,35,8,89,107)
4	(131,135,156,182,118,7)	(25,62,134,103,154,43)

While cluster centroids in table 3 are different from those in table 1 a more detailed analysis of the assigned visitor behavior vectors showed similarities between the clusters found. It cannot be concluded, which method gives a better solution; both results, however, have been confirmed and interpreted by the business's expert user.

The performance of k-means depends directly on the similarity measure used. In the presented case, it is complex and expensive in calculation, compared with the traditional Euclidean distance typically used in k-means. Comparing computation time of k-means and SOFM showed that k-means requires approximately 2.5 times more resources than the neural network.

7. Conclusions

We introduced a methodology to understand the visitor behavior in a web site, using a new similarity measure based on three characteristics derived from the visitor sessions: the sequence of visited pages, their content and the time spent in each one of them. Using this similarity in a self organizing map, we can identify clusters of visitor sessions, which allow us to study the user behavior in the respective web site.

The experiments made with data from a particular web site showed that the methodology used allows to identify meaningful clusters of user sessions, and - using this information - to understand the visitor behavior in the web.

The proposed similarity measure assumes a relatively static web site, i.e., the set of web pages and the text of each page does not change very much in the considered time horizon. If, however, page variation is high, for instance in a newspaper web site, the proposed similarity can be negatively affected. A solution is to label automatically the sites pages in order to compare pages with similar content.

The similarity introduced can be very useful to increase the knowledge about the visitor behavior in a web site. As future work it is proposed to improve the presented methodology introducing advanced variables derived from visitor sessions as well as to develop alternative clustering algorithms using the proposed similarity measure. 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] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, chapter 2. *Addison-Wesley* 1999

[2] M. W. Berry, S. T. Dumais and G.W. O'Brien, Using linear algebra for intelligent information retrival, *SIAM Review*, Vol. 37, pages 573-595, December 1995

[3] B. Berendt and M. Spiliopoulou, Analysis of navigation behavior in web sites integrating multiple information systems, *The VLDB Journal*, Vol. 9, pages 56-75, 2001

[4] B. Berent, A. Hotho and G. Stumme, Towards Semantic Web Mining, *Proceedings of the First International Semantic Web Conference*, pages 264-278, Sardinia, Italy, June 9-12, 2002.

[5] R. Cooley, B. Mobasher, J. Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems* Vol. 1, pages 5-32, 1999.

[6] J.A. Hartigan and M.A. Wong, A K-means clustering algorithm, *Journal of the Applied Statistics*, Vol. 28, pages 100108, 1979

[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*, 2nd edition,1987.

[9] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, *Sov. Phys. Dokl.*, pages 705-710, 1966

[10] B. Mobasher, T. Luo, Y. Sung, and J. Zhu, Integrating Web Usage and Content Mining for More Effective Personalization, *In Proceedings of the International Conference on E-Commerce and Web Technologies*, September, Greenwich, UK, 2000

[11] J. Shavlik and G. G. Towell, Knowledge-based artificial neural networks, *Artificial Intelligence*, vol. 70, no. 1/2, pages 119-165,1994

[12] S. K. Pal, V. Talwar and P. Mitra, Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, *IEEE Transactions on Neural Networks*, Vol. 13, No. 5 pages 1163-1177, September, 2002

[13] M. Perkowitz, Adaptive Web Site: Cluster Mining and Conceptual Clustering for Index Page Synthesis, *Dissertation for degree of Doctor of Philosophy, University of Washington*, 2001

[14] T.A. Runkler and J.C. Bezdek, Web mining with Relational Clustering, *International Journal of Approximate Reasoning*, Vol. 32, no. 2-3, pages 217-236, February, 2003.

[15] J. Velásquez, H. Yasuda, T. Aoki and Richard Weber, Acquiring Knowledge About Users's Preferences in a Web Site, *Proc. First IEEE Int. Conf. on Information Technology: Research and Education*, pages 375-379, Newark, New Jersey, USA,August 2003

[16] J. Velásquez, H. Yasuda, T. Aoki and Richard Weber, Using Self Organizing Feature Maps to acquire knowledge about visitor behavior in a web site, *In Proc. of the Knowledge-Based Intelligent Information & Engineering Systems*, pages 951-958, Oxford, UK, September, 2003

[17] J. Velásquez, H. Yasuda, T. Aoki and R. Weber, Voice Codification using Self Organizing Maps as Data Mining Tool. *Proc. of Second Int. Conf. on Hybrid Intelligent Systems* , pages 480-489, Santiago, Chile, December, 2002

Veláquez, Juan D. He is doctoral student of the RCAST, University of Tokyo. He received the B.E. in Electrical Engineering and B.E. in Computer Science in 1995, the P.E. in Electrical Engineering and P.E. in Computer Engineering in 1996, Master in Computer Science and Master in Industrial Engineering in 2001 and 2002 respectively from the University of Chile, Chile. From 1997, he is adjunct professor at Computer Science and Industrial Engineering Departments, University of Chile. In 1998, he received the Oracle Data Base Administrator Certification. In the professional area, he has been consultant in several ministries of the Republic of Chile and software companies in Latin America. His research interests include Data Mining, Web Mining and Very Large Data Bases.

Yasuda, Hiroshi He received the B.E., M.E. and Dr.E. from the University of Tokyo, Japan in 1967, 1969, and 1972 respectively. Since joining the Electrical Communication Laboratories of NTT, in 1972, he has been involved in works on Video Coding, Image Processing, Telepresence, B-ISDN Network and Services, Internet and Computer Communication Applications. After served twenty-five years (1972-1997), with last position of Vice President, Director of NTT Information and Communication Systems Laboratories at Yokosuka, he left NTT and has joined the University of Tokyo. He is now Director of The Center for Collaborative Research (CCR). He had served as the Chairman of ISO/IEC JTC1/SC29 (JPEG/MPEG Standardization) from 1991 to 1999. He had also served as the President of DAVIC (Digital Audio Video Council) from September 1996 to September 1998. He received 1987 Takayanagi Award, 1995 the Achievement Award of EICEJ, 1995-1996 The EMMY from The National Academy of Television Arts and Science and also 2000 Charles Proteus Steinmetz Award from IEEE. He is Fellow of IEEE, EICEJ, and IPSJ, a member of Television Institute.

Aoki, Terumasa He is lecturer with the Research Center for Advanced Science and Technology, the University of Tokyo. He received his B.S., M.E. and Ph.D. in Information and Communication from the University of Tokyo, Japan in 1993, 1995, and 1998 respectively. His current research interests are in the fields of Terabit IP router, access control of Gigabit LAN/WAN, next-generation video conferencing system, high efficient image coding and management of digital content copyrights. He has received various academic excellent awards such as the 2001 IPSJ Yamashita award, the FEEICP Inose award for 1994, and the other 4 awards.

Weber, Richard He is assistant professor at the Department of Industrial Engineering of the University of Chile and academic director of the Master and Ph.D. program of Operations Management. He received a B.S. in Mathematics, a M.S. and a Ph.D. in Operations Research from Aachen Institute of Technology, Germany. His research interests include Data Mining, Dynamic Data Mining, and Web Mining. He is member of IEEE, ACM and the editorial board of the international journal IDA-Intelligent Data Analysis and has been visiting professor at The Center for Collaborative Research (CCR) at the University of Tokyo.