A Methodology to Find Web Site Keywords

Juan D. Velásquez† Richard Weber§ Hiroshi Yasuda†
Terumasa Aoki†
†Research Center for Advanced Science and Technology, University of Tokyo, E-mail: {jvelasqu,yasuda,aoki}@mpeg.rcast.u-tokyo.ac.jp
§Center for Collaborative Research, University of Tokyo, E-mail: weber@vp.ccr.u-tokyo.ac.jp
on leave from: Department of Industrial Engineering, University of Chile

Abstract

For many companies and/or institutions it is no longer sufficient to have a web site and high quality products or services. What in many cases makes the difference between success and failure of e-business is the potential of the respective web site to attract and retain visitors. This potential is determined by a site’s content, its design, and technical aspects, such as e.g. time to load the pages among others.

In this paper, we concentrate on the content represented by free text of each of the web pages. We propose a method to determine the set of the most important words in a web site from the visitor’s point of view. This is done combining usage information with web page content arriving at a set of keywords determined implicitly by the site’s visitors.

Applying self-organizing neural networks to the respective usage and content data we can identify clusters of typical visitors and the most important pages and words for each cluster. We applied our method to a bank’s web site in order to show its benefits.

Institutions that perform consequently and regularly the proposed analysis can design their web sites according to their visitors’ needs and requirements and this way stay ahead of their competitors.

1. Introduction

Internet provides one of the most powerful tools for effective marketing and efficient operations for many institutions. The data provided in and generated by a web site contains crucial information for the success of the related business.

In order to analyze such data, web mining [10, 12, 14] offers various techniques that are classified into web structure mining, web content mining, and web usage mining.

This paper presents a methodology for web site improvement combining the latter two areas using information from two sources: content and usage of web sites. We use web page content, especially free text together with pattern from web usage as input for clustering of visitor sessions [16]. Analyzing the pages that belong to each one of the clusters found, we can determine the most important words for each cluster and consequently for each type of visitor.

Determining clusters of similar user sessions is, however, a non-trivial process due to the complex structure of today's web sites [17]. In previous work, we presented a Data Mart we have constructed to store efficiently the available data and developed a measure to determine the similarity of visitor sessions using data from this Data Mart [19].

Based on these tools we applied a self-organizing feature map for clustering of sessions of a bank’s web site. The clusters found and the relevance of the different words inside each page allowed us to identify the most important words for the respective clusters, i.e. groups of visitors. This information has clearly a tremendous value for companies where free text is important in order to attract and retain visitors.

Section II. of this paper provides an overview on related work. In section III. we describe the data preparation process, which is necessary for the comparison of user sessions based on the content of web pages. This comparison is presented in section IV. In section V, we show how a self-organizing feature map was used for session clustering using the previously introduced similarity measure. Section VI describes the application of our work for a Chilean bank. Section VII concludes this work and points at future extensions.
2. Related work

Since our goal is to identify words from a particular web site that attract the majority of the visitors and retain them on the site we have to combine information from web usage and web content. This way we can determine the importance of pages as well as of single words inside the pages for our current visitors.

To understand which content of a certain web page catches a visitor’s attention is, however, a non-trivial problem, since such pages typically contain a collection of heterogeneous, unlabelled, distributed, time variant, semi-structured and high dimensional data [14]. This problem has motivated on one hand the development of models to represent the respective data and store the associated information [13, 17].

On the other hand, new algorithms to mine web data have been suggested [10]. These are categorized in three sub-areas: Web Structure Mining (WSM), Web Content Mining (WCM), and Web Usage Mining (WUM) [2].

In this paper, we propose a combination of WCM and WUM techniques. In the following sub-sections we will examine the current state-of-the-art in these areas in relation to our work. But first we will explain what we understand as ‘keywords’ in the context of this paper.

2.1. Web Site Keywords: The Concept

To find the most appropriate keywords for a web site is an important undertaking in order to get good rankings in search engines and to be found by potential customers. Web spiders are software programs that search relevant information (e.g. important words in web pages) [5]. Several commercial tools¹ support this task to find the perfect target keywords customers are likely to use while searching the web [3]. A keyword is a word or possibly set of words [9] that is used by visitors in their search process and characterizes the content of a given web page or web site.

Our approach also tries to determine the most important words from a given site. It differs, however, from the aforementioned proposals in the sense that we conclude the words that attract and retain visitors to a particular web site from their usage. With other words we involve current and past visitors into a continuous process of keyword determination.

2.2. Web content mining

The goal of Web Content Mining (WCM) is to find useful information from web contents. In this sense, it is comparable to Information Retrieval (IR) techniques [1].

In this paper we are concerned about IR applied to a collection of documents (e.g. web pages) of free text. Traditionally, such documents are represented using the vector space model [1]. A step previous to this model is the word representation based on a tokenized process, using simple syntactic rules, and token stemmed to their canonical form (e.g., ‘writing’ to ‘write’, ‘are’, ‘is’, ‘were’ to ‘be’).

Let $R$ be the number of different words in the entire collection of documents and $Q$ the number of documents. In our case a document would be a web page and the collection of documents the respective web site. A vectorial representation of the web site would then be a matrix $M$ of dimension $R \times Q$ with:

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

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

The weight must capture the fact that different words can have different degrees of importance. For instance, if $n_i$ is the number of documents containing word $i$, the expression $\log(\frac{N}{n_i})$ gives a notion of its importance in the complete set. The “inverse document frequency” $IDF = \log(\frac{Q}{n_i})$ can be used like a weight.

A variation of the last expression is to apply a factor to the IDF part, known as TF*IDF (term frequency inverse document frequency) and shown in equation 2.

$$ m_{ij} = f_{ij} \times \log(\frac{Q}{n_i}) \quad (2) $$

where $f_{ij}$ is the number of occurrences of word $i$ in document $j$.

In the vectorial representation, a document can be compared to another one using their distance in the vector space. Below, we will use this distance as input for data mining algorithms.

In WCM we have two main strategies: mining of document contents (web page content mining) and improvement of content search in tools like search engines (search result mining). In this paper we focus on web page content mining. It can be unstructured as free text, semi structured as HTML or fully structured as tables or databases [14].

Beyond analyzing text contained in web pages, extracting concepts from hypertexts is also part of WCM [4]. Hypertext documents like e.g. web pages are modelled in different levels of detail depending on the application. The simple model indicates two elements $(D, L)$, where $D$ represents the content document, for instance by a vector representation, and $L$ the links information.

Since the representation of the hypertext depends on the particular application, it is necessary to identify first what elements are desired to model. In the case of a web page, aside from free text, the information contained in tags referring to the content, for example, the $<title>$ tag shows the

¹ http://www.goodkeywords.com/
main theme of a page, and should be incorporated in the final expression of the hypertext vector representation [19].

2.3. Web usage mining

Web usage mining (WUM) uses traditional data mining methods in order to work with usage data. However, modifications are necessary due to the data's nature.

As first step, data incompleteness calls for preprocessing of usage data [16]. Some possible scenarios are:

- Single IP address/Multiple Server Sessions. Usually, the Internet Services Providers (ISP) have mechanisms to accelerate the visitor request to a web site, for instance Proxy Servers. In this case we have a single IP address accessing to a web site. The real situation, however, regarding visits to the site is different, because there are more real sessions in the requests.
- Multiple IP addresses/Single Server Sessions. For privacy reasons or ISP configuration, it is possible to assign a random IP address to visitor request.
- Multiple IP address/Single Visitor. A visitor that accesses a web site from different machines, but has the same behavior each time.
- Multiple Agent/Single User. As before, when a visitor uses different machines that may have different agents.

Applying the sessionization process corrects partially the problems mentioned above by identifying the real sessions per visitor. Thus the preprocessing step returns the input data for the mining process.

The goal of WUM is pattern discovery using different kinds of data mining techniques, such as statistical analysis, association rules, clustering, classification, sequential patterns and dependency modelling [12, 16].

Applications of WUM can be grouped in two main categories: user modelling in adaptive interfaces, known as personalization, and identification of user navigation pattern, in order to improve the web site structure.

After applying web mining it becomes necessary to analyze the discovered patterns. In this step, the participation of an expert in the business in study is recommended.

2.4. Combining WUM and WCM

Combining the philosophies behind WCM and WUM provides a complemented vision of both techniques [13]. Applying WUM we can understand the visitor browsing behavior, but we cannot discover which content is interesting for the visitor. This analysis is possible using WCM [2].

Comparing the content of visited pages provides information on visitors preferences regarding web page content [17, 18, 19]. In this sense, it has been proposed to apply the vector space model introduced in section 2.2 to the web pages. Using a distance among vectors (e.g. Euclidean distance), we can find the main topics of interest in the visited pages.

Finally, a similarity measure has been suggested [19] that allows to compare the behavior of different visitors, through the analysis of visitor preferences. This measure has been used in a cluster algorithm in order to find groups of similar visitor sessions [18, 19] and using this information, predict preferences of future web site visitors.

3. Data preparation process

Our method for keyword identification requires web data from the following two sources that are easily available:

- Web log registers.
- Web pages.

Web log registers contain information about the browsing behavior of web site visitors, in particular the page navigation sequence and the time spent in each page. 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. Here, we are only interested in the free text of the pages.

In order to study the visitor 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 [6]. 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.

2 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
3.2. Web page processing

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 preprocess the text applying the following filters:

- HTML Tags. Some tags show interesting information about the page content, for instance, the <title> tags mark the web page’s central theme. We use this information to identify special words inside the text.
- Stop words (e.g. pronouns, prepositions, conjunctions, etc.) are removed.

After filtering, we represent a document (web site) by a vector space model. Let $R$ be the number of different words in a web site and $Q$ be the number of its pages. Based on equation 2, we propose a variation incorporating the influence of special words, i.e., words that have different levels of importance for a visitor. Some examples are words associated to page title, specially marked word (e.g. using italic font), a referrer word (used by a visitor in order to find a topic), etc.

$$m_{ij} = f_{ij}(1 + \alpha w_i) \cdot \log\left(\frac{Q}{n_i}\right)$$

In equation 3, $\alpha w_i$ (special words) is an array with dimension $R$ reflecting the mentioned words. Its component $i$ is the weight of a particular word in the web site.

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)**

$$WP_k = (wp_{k1}, \ldots, wp_{kQ}) = (m_{1k}, \ldots, m_{Rk}) \; k = 1, \ldots, Q$$

Based on this definition, we used the angle’s cosine as similarity measure between two page vectors:

$$dp(WP^i, WP^j) = \frac{\sum_{k=1}^{R} wp_{ki} wp_{kj}}{\sqrt{\sum_{k=1}^{R} (wp_{ki})^2} \sqrt{\sum_{k=1}^{R} (wp_{kj})^2}} \quad (4)$$

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

### 4. Comparing visited web pages

We propose a measure to compare the behavior of two visitors through the analysis of their preferences [18], in particular analyzing the visited pages regarding their content and the interest for each of these pages. Consequently, our model for visitor behavior is based on the following two variables: the content and the time spent in each one of them [17, 18]. The model is represented by a vector with dimension $n$ and two parameters in each component.

**Definition 2 (Visitor Behavior Vector)** Let $v$ be the Visitor Behavior Vector for a certain visitor. Then we can write $v$ as:

$$v = [(p_1, t_1), \ldots, (p_n, t_n)]$$

being $(p_i, t_i)$ parameters that represent the page and the percentage of total session time spent on the $i^{th}$ page of the respective visit (relative time). In this expression, $p_i$ is the page identifier.

From the visitor behavior vector we want to select the most important pages, assuming the importance being correlated to the relative time spent on each page. This is done by the following definition.

**Definition 3 (Important Pages Vector)**

$$\vartheta_i(v) = [(\rho_1, \tau_1), \ldots, (\rho_i, \tau_i)]$$

being $(\rho_k, \tau_k)$ the component in $v$ that represents the $k^{th}$ most important page and the relative time spent on it.

We take the visitor behavior vector and sort it according to the relative time spent on each page. Then we select the $i$ most important pages, i.e. the first $i$ pages.

Let $\alpha$ and $\beta$ be two visitor behavior vectors. The proposed similarity measure between the two visitors is introduced in equation 5:

$$st(\vartheta_i(\alpha), \vartheta_i(\beta)) = \sum_{k=1}^{i} \min\left(\frac{\tau_{\alpha}^k}{\tau_{\alpha}^\beta}, \frac{\tau_{\beta}^k}{\tau_{\beta}^\alpha}\right) \cdot dp(\rho_{\alpha}^k, \rho_{\beta}^k) \quad (5)$$
The first element is indicating the visitor’s interest in the pages visited. If the percentage of time spent by visitors $\alpha$ and $\beta$ on the $k^{th}$ page visited are close to each other, the value of the expression will be near 1. In the opposite case, it will be near 0.

The second element is $dp$, the distance between pages in the vectorial representation introduced in equation 4. This distance is used because two visitors may visit different web pages in the web site with similar content, e.g., one page contains information about classic rock and another one about progressive rock. In both cases the visitors have interest in music, especially in rock.

Finally, we combine in equation 5 the content of the most important pages with the time spent on each of them by a multiplication. This way we can distinguish between two pages with similar content, but different percentages of time spent on them.

5. Mining the web site content and visitor transactions

We use a clustering algorithm in order to find groups of similar visitor sessions. Based on this information we determine the most important words for each cluster.

5.1. Clustering visitor sessions

An artificial neural network of the Kohonen type (Self-organizing Feature Map; SOFM) has been applied to the preprocessed data that is stored in a data mart repository. Schematically, a SOFM is represented as a two-dimensional array in whose positions the neurons are located [7]. 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 notion of neighborhood among the neurons provides diverse topologies. In this case the thoroidal topology was used [18], which means that the neurons closest to the ones of the superior edge, are located in the inferior and lateral edges (see figure 2).

Since the Important Pages vectors have two components, 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 important page vector example presented to the network. The time vector’s component is modified with a numerical adjustment, i.e., $\tau_{i+1}^N = \tau_{i}^N \ast f_{\beta}$ with $i = 1, \ldots, \epsilon$.

The page component needs another updating scheme [8]. Using the page distance, the difference between page content components is shown in expression 6.

$$D_{NE} = [dp(\rho_{1}^{N}, \rho_{1}^{E}), \ldots, dp(\rho_{\epsilon}^{N}, \rho_{\epsilon}^{E})]$$

It 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 $D'_{NE} = D_{NE} \ast f_{\epsilon}$, with $f_{\epsilon}$ 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 and its neighbor neurons is given by equation 7.

$$\rho_{i+1}^{N} = \gamma \in \Gamma / D'_{NE,i} \approx dp(\gamma, \rho_{i}^{N})$$

with $\Gamma = \{\gamma_{1}, \ldots, \gamma_{Q}\}$ the entire set of all pages in the web site, $D'_{NE,i}$ the $i^{th}$ component of $D'_{NE}$. Then given $D'_{NE,i}$ it is necessary to find the page $\gamma$ in $\Gamma$ whose $dp(\gamma, \rho_{i}^{N})$ is closest to $D'_{NE,i}$.

5.2. Identifying web site keywords

Using the vector page model and equation 3, we propose the following method to determine the most important keywords and their importance in each cluster. Equation 8 shows a measure (geometric mean) used in order to calculate the importance of each word relative to each cluster.

$$kw[i] = \sqrt{\Pi_{p \epsilon \zeta} m_{ip}}$$

with $i = 1, \ldots, R$. $kw[i]$ is an array with the weights for each word relative to a given cluster and $\zeta$ the set of pages representing this cluster. Sorting $kw[i]$ we can select a group of the most important words for each cluster.

6. Practical Application

In order to prove the effectiveness of the proposed approach for keyword determination, a web site was selected, considering the following characteristics:

- 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 the web master with pages of interest, i.e., if a page is not visited, then it will be dropped.
The web site selected\footnote{http://www.tbanc.cl/}, 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, corresponding to the period January to March, 2003.

### 6.1. Sessionization process

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

Only 16% of the visitors visit 10 or more pages and 18% less than 4. The average number of visited pages is 6, thus we fixed 6 as cardinal of the visitor behavior vector. We chose 3 as maximum number of components of the important page vector, i.e., the parameter \( \iota = 3 \). Finally, applying the above described filters, approximately 400,000 vectors were identified.

### 6.2. Web page content processing

Applying web page text filters, we find that the complete web site contains \( R = 4,096 \) different words for our analysis. Regarding word weights, especially the special words (see equation 3), we applied the following procedure. In order to calculate \( sw_i \) in equation 3, we have three sources in the particular web site in study:

1. The web site offers the option to send e-mails to the call center. The text sent is a source to identify important words. Let \( ew_i = \frac{w_{i}^{\text{email}}}{TE} \) be the array with the special words inside of e-mails, where \( w_{i}^{\text{email}} \) is the frequency of word \( i \) in the complete set of e-mails and \( TE \) is the total number of words in the e-mails.

2. Marked words. A web page contains words with special tags, e.g., a different font like italic or a word belonging to the title phrase. Let \( mw_i = \frac{w_{i}^{\text{mark}}}{TM} \) be the array with the marked words inside the web site, where \( w_{i}^{\text{mark}} \) is the frequency of word \( i \) in the whole web site and \( TM \) is the total number of words in the site.

3. Searched words. The web site offers also a search engine, where \( aw_i = \frac{w_{i}^{\text{ask}}}{TA} \) be the array with the words used in this search engine, where \( w_{i}^{\text{ask}} \) is the frequency of word \( i \) in the complete set of words used and \( TA \) is the total number of words.

The final expression \( sw_i = ew_i + mw_i + aw_i \) is the simple sum of the weights using the above described methods.

Applying equation 3, a 3-dimensional matrix with the similarity between pairs of pages is created and shown graphically in figure 3.

![Figure 3. Similarity measure among web pages](image)

Since the matrix is triangular, figure 3 shows only its superior side. Clusters of similar page contents can be identified.

### 6.3. Applying Self-Organizing Feature Maps

We used a SOFM with 3 input neurons and 32 output neurons. The thoroidal topology maintains the continuity between clusters, which allows to study the transition among the preferences of the visitor from one cluster to another. The neural network was trained on a Pentium IV, with 1 Gb in RAM and running Linux Operating System, distribution Redhat 8.0. The time necessary was 25 hours and the epoch parameter was 100. Since training is done offline and only few times, the duration of 25 hours did not present problems in this application.

Figure 4 shows in the \( x, y \) axis the nets neurons. The \( z \) axis is a normalized version of the number of times that a neuron won during training.

### 6.4. Results

In figure 4 we can identify 8 main clusters. They contain the information about the most important pages in the web site as shown in table 1. The second column contains
the center neurons (winner neuron) of each cluster, representing the most important pages visited.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Pages Visited</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(6,8,190)</td>
</tr>
<tr>
<td>2</td>
<td>(100,128,30)</td>
</tr>
<tr>
<td>3</td>
<td>(86,150,97)</td>
</tr>
<tr>
<td>4</td>
<td>(101,105,1)</td>
</tr>
<tr>
<td>5</td>
<td>(3,9,147)</td>
</tr>
<tr>
<td>6</td>
<td>(100,126,58)</td>
</tr>
<tr>
<td>7</td>
<td>(70,186,137)</td>
</tr>
<tr>
<td>8</td>
<td>(157,169,180)</td>
</tr>
</tbody>
</table>

Table 1. Important page clusters

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

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

Table 2. Pages and their content

Applying equation 8, we get the keywords and their relative importance in each cluster. For instance, if \( \zeta = \{6, 8, 190\} \), then \( kw[i] = \sqrt{m_6m_8m_{190}} \), with \( i = 1, \ldots, R \).

Finally, sorting \( kw[i] \) we can select a group of the most important words in each cluster, for instance the 8 words with highest weight. Due to the confidentiality agreement with the bank where we apply the described techniques, we cannot show the specific keywords found per cluster. However, using a word label, we present the result in figure 5.

Some keywords from figure 5 are selected randomly and presented in table 3. Following the confidentiality agreement, their specific weights cannot be shown.

<table>
<thead>
<tr>
<th>#</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Crédito</td>
</tr>
<tr>
<td>2</td>
<td>Hipotecario</td>
</tr>
<tr>
<td>3</td>
<td>Tarjeta</td>
</tr>
<tr>
<td>4</td>
<td>Promoción</td>
</tr>
<tr>
<td>5</td>
<td>Concurso</td>
</tr>
<tr>
<td>6</td>
<td>Puntos</td>
</tr>
<tr>
<td>7</td>
<td>Descuento</td>
</tr>
<tr>
<td>8</td>
<td>Cuenta</td>
</tr>
</tbody>
</table>

Table 3. Some of the identified keywords (in Spanish)

7. Conclusions

A methodology to identify keywords inside a web site was introduced. Its particularity is that these keywords are concluded by combining visitor behavior and web page content. In the first part we proposed a way to find the most important pages for the visitor, assuming that the time spent in each page is proportional to the visitor interest.

Next a new similarity measure was defined. It is based on two characteristics derived from the above assumption: the most important pages visited, and the time spent in each one of them, ordered by time. Using this similarity in a self-organizing feature map, we found clusters from visitor sessions, which allow us to study the visitor text preferences in the web site. The similarity introduced, can be very useful to increase the knowledge about the visitor preferences in the web, in particular to identify keywords that attract and retain visitors.

As future work, it is proposed to improve the presented methodology introducing advanced variables, such as weights that allow to differentiate between pages having different numbers of words. We will also add content other than free text to our analysis, e.g. pictures. Furthermore, it will be necessary to continue applying our method-
ology to other web sites in order to get new hints on future developments.

Acknowledgement: This project was supported partially by the Nucleus Millennium Science on Complex Engineering Systems (www.sistemasdeingenieria.cl).

References