Towards the Identification of Important Words from the Web User Point of View

Juan D. Velásquez and José I. Fernández
Department of Industrial Engineering, University of Chile,
Av. República 701,office 301, Santiago, CHILE, P.C. 837-0720
E-mail: jvelasqu@dii.uchile.cl, josferna@ing.uchile.cl

Abstract

We introduce a methodology for identifying approximately which words attract the users attention when they are visiting the pages in a web site. These words are called "web site keywords" and can be used for creating further web page text contents related with an specific topic. Through the utilization of the correct words, we can help to the users to find what they are looking for. By applying a clustering algorithm, and under the assumption that there is a correlation between the time spent in a page and the user interest, we segment the users by navigation behavior and content preferences. Next we identify the web site keywords. This methodology was applied on data originated in a real web site, showing its effectiveness.

1 Introduction

For many companies and/or institutions, it is no longer sufficient to have a web site and high quality products or services. What often makes the difference between e-business success and failure is the potential of the respective web site to attract and retain users. This potential depends on the site content, design, and technical aspects, such as the time to download the pages from the web site to the user's web browser, among others. In terms of content, the words used in the free text of a web site pages are very important, as the majority of the users perform term-base queries in a search engine to find information on the Web. These queries are formed by keywords, i.e., a word or a set of words [14] that characterize the content of a given web page or web site.

The suitable utilization of words in the web page improves user appeal, helps effective information search, while attracting new users and retaining current users by continuous updating of page text content. So the challenge is to identify which words are important for users. Most keywords are calculated from "most frequently used words". Some commercial tools[1] help identify target keywords that customers are likely to use while web searching [6].

By identifying the most relevant words in the sites pages, from the user point of view, improvements can be performed in the entire web site. For instance, the site could be restructured putting a new hyperlink related with the keyword and of course the text content could be modify by using the keywords related with an specific topic to enrich the free text in a web page.

In this paper a methodology for analyzing the user browsing behavior and text preferences is introduced, through the application of web mining algorithms on data originated in the Web, also called web data, specify web log registers and web site text content.

The methodology aims to identify approximately which words attract the users attention when they are visiting the pages in a web site. These words are called "web site keywords" [31] and can be used for creating further web page text contents related with an specific topic.

This paper is organized as follow. Section 2 introduces a short revision about the related work. The preparation process for transforming the web data in feature vectors to be used as input in web mining algorithm, is shown in section 3. In section 4, the methodology for identifying the web site keywords is explained and applied in section 5. Finally, the section 6 shows the main conclusions of this paper.

2 Related work

When a user visit a web site, data about the pages visited are stored in the web log files. Then it is direct to know which pages were visited and which not, and inclusive the time spent by a user in each one of them. Because usually the pages contain information about an specific topic, it is possible to know approximately the users information preferences. In that sense the interaction between user and site is like an electronic inquest, given us the necessary data for

[1] see e.g. http://www.goodkeywords.com/

analyzing the user content preferences in a particular web site.

The challenge for analyzing the user text preferences in the site is double. First the amount of web log registers usually come be huge, and an important part of them with irrelevant information about the user browsing behavior in the site. Second, the free text inside of the web pages is commonly plain, i.e., without additional information that allow us to know directly which words attract the user attention.

In this section the main approaches to analyze web data for extracting significant patterns related with the users text preferences in a web site are reviewed.

2.1 Mining web data

Web mining techniques emerged as a result of the application of data mining theory to pattern discovery from web data [8, 16, 25]. Web mining is not a trivial task, considering that the web is a huge collection of heterogeneous, unlabelled, distributed, time variant, semi-structured and high dimensional data. Web mining must consider three important steps: preprocessing, pattern discovery and pattern analysis [27].

The following common terminology is used to define the different types of web data:

- Content. The web page content, i.e., pictures, free text, sounds, etc.
- Structure. Data that shows the internal web page structure. In general, they have HTML or XML tags, some of which contain information about hyperlink connections with other web pages.
- Usage. Data that describes visitor preferences while browsing in a web site. It is possible to find such data inside web log files.
- User profile. A collection of information about a user: personal information (name, age, etc.), usage information (e.g. visited pages) and interest.

With the above definitions and depending of the web data to be processed, web mining techniques can be grouped in three areas: Web Content Mining (WCM), Web Structure Mining (WSM), and Web Usage Mining (WUM).

2.2 Identifying words for creating an automatic web page text summarization

The goal is to automatically construct summaries of a natural-language document [11]. In this case,a relative semi-structure is created by the application of HTML tags from web page text content, which examines topics without restriction to a domain. In many cases, the pages might only contain few words and non-textual elements (e.g. video, pictures, audio, etc.) [1].

In text summarization research, the three major approaches are [18]: paragraph based, sentence based and using natural language cues in the text.

The first approach consists in selecting a single paragraph of a text segment [19] that addresses a single topic in the document, under the assumption that there are several topics in the text. The application of this technique in a web page is not obvious; web site designers have a tendency to structure the text by paragraph per page. Therefore a document contains only a single topic, which makes the application of this technique difficult.

In the second approach, the most interesting phrases or key-phrases are extracted and assembled in a single text [9, 37]. It is clear that the resulting text may not be cohesive, but the techniques goal is to provide maximum expression of the information in the document. This technique is suitable for web pages, since the input may consist of small pieces of text [6]. The final approach is a discourse model based on extraction and summarization [14, 15] by using natural language cues such as proper names identification, synonyms, key-phrases, etc. This method assembles sentences by creating a collage text with information about the entire document. This technique is most appropriate for documents with a specific domain and thus its implementation for web pages is difficult.

2.3 Web page key-text extraction and applications

The key-text components are parts of an entire document, for instance a paragraph, phrase and up to a word, that contain significant information about a particular topic, from the web site user point of view. The identification of these components can be useful for improving the web site text content.

Usually, the keywords in a web site are correlated with "most frequently used words". In [6], a method for extracting keywords from a large set of web pages is introduced. The technique is based on assigning importance to words, depending on their frequency, in all documents. Next, the paragraph or phrases that contain the keywords are extracted and their importance is validated through tests with human users.

Another method, in [2], collects keywords from a search engine. This shows the global word preferences of a web community, but no details about a particular web site.

Finally, instead of analyzing words, in [17] a technique to extract concepts from web page texts is developed. The concepts describe real-world objects, events, thoughts, opinions and ideas in a simple structure, as descriptor terms. Then, by using the vector space model, the concepts are

transformed into feature vectors, allowing the application of clustering or classification algorithms to web pages and so extract concepts.

3 Web data preparation process

Of all available web data, the most relevant for the analysis of user browsing behavior and preferences, are the web log registers and the web pages [33]. The web log registers contain information about the page navigation sequence and the time spent at each page visited, by applying the **sessionization** process. The web page source is the web site itself. Each web page is defined by its content, in particular free text. To study user behavior both data sources - web logs and web pages - have be prepared by using filters and by estimating real user sessions. The preprocessing stage involves, first, a cleaning process and, second, the creation of the feature vectors as input of the web mining algorithms, within a structure defined by the patterns sought.

3.1 The session reconstruction process

The process of segmenting the visitors activities into individual visitor sessions is called **sessionization** [10]. It is based on web log registers and due to the problems mentioned above, the process is not free of errors [26]. Sessionization assumes that a session has a maximum time duration and it is not possible to know if the visitor pressed the "back" button in the browser. If the page is in the browser cache and the visitor comes back to it in the same session, it would not be registered in the web logs. Thus the use of invasive schemes such as sending another application to the browser and capture the exact visitor browsing have been proposed [3, 10]. However, this scheme could be easily avoided by the visitor.

Many authors [3, 10, 20] have proposed using heuristics to reconstruct sessions from web logs. In essence, the idea is to create subsets with the users visits and apply mechanisms over the web log registers generated that allow to define a session as a series of events interlaced during a certain period.

The session reconstruction aims to find the real user sessions, i.e., which pages were visited by a physical human being. In that sense, whatever the chosen strategy to discover real sessions, it must satisfy two essential criteria: the activities performed by a real person can be grouped together and the activities that belong to the same visit (others object required for the visited page) also belong to the same group.

There are several techniques for sessionization, which can be grouped in two major strategies: *proactive* and *reactive* [26].

Proactive strategies aim to identify the user using identification methods like cookies and these consist in a piece of code associated with the web site. When a visitor visits the site for the first time, a cookie is sent to the browser. Next, when the page is revisited, the browser shows the cookie content to the web server, and an automatic identification takes place. The method has problems from a technical point of view and also with respect to the visitor's privacy. First, if the site is revisited after several hours, the session will be considered too long, it will actually be a new session. Secondly, some aspects of the cookies seem to be incompatible with the principles of data protection of some communities, like the European Union [26]. Finally, the cookies can be easily detected and deactivated by the visitor.

Reactive strategies are noninvasive with respect to privacy and they make use of the information contained in the web log files only and consist in processing the registers to generate a set of reconstructed sessions.

In web site analysis, the general scenario is that the web sites usually don't implement identification mechanisms. The utilization of reactive strategies can be more useful. They can be classified into two main groups [4, 10]:

- Navigation Oriented Heuristics: assume that the visitor reaches pages through hyperlinks from others pages. If a page request is unreachable through pages previously visited by the visitor, a new session is initiated.

- Time Oriented Heuristics: set a maximum time duration, which is usually 30 minutes for the entire session [7]. Based on this value we can identify the transactions belonging to a specific session by using program filters.

3.2 Processing web page text content

There are several methods for comparing the content of two web pages, here considered as the free text inside the web pages. The common process is to match the terms that make up the free text, for instance, by applying a word comparison process. A more complex analysis includes semantic information contained in the free text and involves an approximate term comparing task as well.

Semantic information is easier to extract when documentation includes additional information about the text content, e.g., market language tags. Some web pages allow document comparison by using the structural information contained in HTML tags, although with restrictions. This method is used in [28] for comparing pages written in different languages with similar HTML structure. The comparison is enriched by applying a text content matching process [29], which considers a translation task to be completed

first. The method is highly effective when the same language is used in the pages under comparison. A short survey of algorithms for comparing documents by using structural similarities is found in [5].

Comparisons are made by a function that returns a numerical value showing the similarities or differences between two web pages. This function can be used in the web mining algorithm to process a set of web pages, which might belong to a web community or an isolated web site. The comparison method must consider efficiency criteria in the web page content processing [13]. Here the vector space model [24], allows a simple vectorial representation of the web pages and, by using distance for comparing vectors, provides a measure of the differences or similarities between the pages.

Web pages must be cleaned before transforming them into vectors, both to reduce the number of words - not all words have the same weight - and make the process more efficient. Thus, the process must consider the following types of words:

- HTML Tags. In general, these must be cleaned. However, the information contained in each tag can be used to identify important words in the context of the page. For instance, the <title> tag marks the web page central theme, i.e., gives an approximate notion of the semantic meaning of the word and, is included in the vector representation of the page.
- Stop words (e.g. pronouns, prepositions, conjunctions, etc.)
- Word stems. After applying a word suffix removal process (word stemming [22]), we get the word root or stem.

For vector representation purposes, let R be the total number of different words and Q be the number of pages in the web site. A vectorial representation of the set of pages is a matrix M of size RxQ,

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

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

Based on *tfxidf-weighting* introduced in [24] the weights are estimated as,

$$m_{ij} = f_{ij}(1 + sw(i)) * \log(\frac{Q}{n_i}). \quad (2)$$

Here, f_{ij} is the number of occurrences of word i in page j and n_i is the total number of times that the word i appears in the entire web site. Additionally, a words importance is augmented by the identification of special words, which correspond to terms in the web page that are more important than others, for example, marked words (using HTML tags), words used by the user in search of information and, in general, words that imply the desires and the needs of the users. The importance of special words is stored in the array sw of dimension R, where $sw(i)$ represents an additional weight for the i^{th} word.

The array sw allows the vector space model to include ideas about the semantic information contained in the web page text content by the identification of special words.

The common sources of special words are:

1. E-Mails. The offer to send user e-mails to the call center platform. The text sent is a source to identify the most recurrent words. Let $ew_i = \frac{w^i_{e-mail}}{TE}$ be the array of words contained in e-mails, which are also present in the web site, where w^i_{e-mail} is the frequency of the i^{th} word and TE is the total amount of words in the complete set of e-mails.
2. Marked words. Within a web page, there are words with special tags, such as a different font, e.g., italics, or a word belonging to the title. Let $mw_i = \frac{w^i_{marks}}{TM}$ be the array of marked words inside web pages, where w^i_{mark} is the frequency of the i^{th} word and TM is the total amount of words in the whole web site.
3. Asking words. A bank, for example, has a search engine through which the users can ask for specific subjects, by introducing key words. Let $aw_i = \frac{w^i_{ask}}{TA}$ be the array of words used by the user in the search engine and also contained in the web site, where w^i_{ask} is the frequency of the i^{th} word and TA is the total amount of words in the complete set.
4. Related web site. Usually a web site belongs within a market segment, in this case the financial institutions market. Then, it is possible to collect web site pages belonging to the other sites in the same market. Let $rw_i = \frac{w^i_{rws}}{RWS}$ be the array with the words used in the market web sites including the web site under study, where w^i_{rws} is the frequency of the i^{th} word and RWS is the total number of words in all web sites considered.

The final expression $sw_i = ew_i + mw_i + aw_i + rw_i$ is the simple sum of the weights described above.

In the vectorial representation, each column in the matrix M is a web page. For instance the k^{th} column m_{ik} with $i = 1, \ldots, R$ is the "k^{th}" page in the entire set of pages.

Definition 1 (Word Page Vector) *It is a vector* $\mathbf{WP^k} = (wp^k_1, \ldots, wp^k_R) = (m_{1k}, \ldots, m_{Rk}), \quad k = 1,\ldots,Q$ *, is the vectorial representation of the* k^{th} *page in the set of pages under analysis.*

With the web pages in vectorial representation, it is possible to use a distance measure for comparing text contents. The common distance is the angle cosine calculated as

$$dp(WP^i, WP^j) = \frac{\sum_{k=1}^{R} wp_k^i wp_k^j}{\sqrt{\sum_{k=1}^{R}(wp_k^i)^2}\sqrt{\sum_{k=1}^{R}(wp_k^j)^2}}. \quad (3)$$

The Eq. (3) allows to compare the content of two web pages, returning a numerical value between $[0, 1]$. When the pages are totally different, $dp = 0$, and when they are the same, $dp = 1$. Another important aspect is that the Eq. (3) complies with the requirement of being computationally efficient, which makes it appropriate to be used in web mining algorithms.

4 Extracting user web page content preferences

Different techniques are applied to analyze web site user behavior ranging from simple web page use statistics to complex web mining algorithms. In the last case, research concentrates on predictions about which page the user will visit next and the information they are looking for.

By using mainly a combination of a WUM and WCM approaches, it is propose to analyze the web user text preferences in a web site and for this way, to identify which words attract the user attention during their navigation in the site.

Prior to the application of a web mining tool, the data related to web user behavior has to be processed to create feature vectors, whose components will depend on the particular implementation of the mining algorithm to be used and the preference patterns to be extracted.

4.1 Modelling the web user behavior

The majority of the web user behaviour models examine the sequence of pages visited to create a feature vector that represents the web user's browsing profile in a web site [12, 21, 36]. These models analyze web user browsing behavior at a web site by applying algorithms to extract browsing patterns. A next step is to examine user preferences, defined as the web page content preferred by the user; and it is the text content that captures special attention, since it is used to find interesting information related to a particular topic by search engine. Hence, it is necessary to include a new variable as part of the web user behavior feature vector - information about the content and time spent in each web page visited.

Definition 2 (User Behavior Vector (UBV)) *It is a vector* $\upsilon = [(p_1, t_1) \ldots (p_n, t_n)]$*, where* (p_i, t_i) *are the parameters that represent the* i^{th} *page from a visit and the percentage of time spent on it in the session, respectively. In this expression,* p_i *is the page identifier.*

In Definition 2, the user behavior in a web site is characterized by:

1. Page sequence; the sequence of pages visited and registered in the web log files. If the user returns to a page stored in the browser cache, this action may not be registered.

2. Page content; represents page content, which can be free text, images, sounds, etc. For the purposes of this paper, the free text is mainly used to represent the page.

3. Spent time; time spent by the user in each page. From the data, the percentage of time spent in each page during the user session can be directly calculated.

4.2 Analyzing the user text preferences

The aim is to determine the most important words at a given web site for users, by comparing the user text preferences, through the analysis of pages visited and the time spent on each of them [34]. It differs, however, from the previously mentioned approaches, as the exercise is to find the keywords that attract and retain users from the user web usage data available. The expectation is to involve current and past users in a continuous process of keywords determination.

User preferences about web content are identified by content comparison of pages visited, [34, 33, 35] by applying the vector space model to the web pages, with the variation proposed in section 3.2, Eq. (2). The main topics of interest can be found by using a distance measure among vectors (e.g. Euclidean distance),

From the user behavior vector (UBV), the most important pages are selected assuming that degree of importance is correlated to the percentage of time spent on each page. The UBV is sorted according to the percentage of total session time spent on each page. Then the ι most important pages, i.e. the first ι pages, are selected.

Definition 3 (Important Pages Vector (IPV)) *It is a vector*
$\vartheta_\iota(\upsilon) = [(\rho_1, \tau_1), \ldots, (\rho_\iota, \tau_\iota)]$*, where* (ρ_ι, τ_ι) *is the component that represents the* ι^{th} *most important page and the percentage of time spent on it by session.*

Let α and β be two UBVs. The proposed similarity measure between the two IPVs is introduced in equation 4 as:

$$st(\vartheta_\iota(\alpha), \vartheta_\iota(\beta)) = \frac{1}{\iota} \sum_{k=1}^{\iota} \min\{\frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha}\} * dp(\rho_k^\alpha, \rho_k^\beta) \quad (4)$$

The first element in (4) indicates the users interest in the visited pages. If the percentage of time spent by users α and β on the k^{th} page visited is close to each other, the value of the expression $\min\{\cdot, \cdot\}$ will be near **1**. In the extreme opposite case, it will be near **0**. The second element in (4) is dp, the distance between pages in the vectorial representation introduced in (3). In (4) the content of the most important pages is multiplied by the percentage of total time spent on each page. This allows pages with similar contents to be distinguished by different user interests.

4.3 Identifying web site keywords

A web site keyword is defined as *"a word or possibly a set of words that make a web page more attractive for an eventual user during his/her visit to the web site"* [32]. It is interesting to note that the same web site keywords may be used by the user in a search engine, when he or she is looking for web content.

In order to find the web site keywords, it is necessary to select the web pages with text content that is significant for users. The assumption is that there exists a relation between the time spent by the user in a page and his/her interest in its content [31]. This relation is collected by the Important Page Vector (IPV), given the necessary data for extracting the web site keywords through the utilization of a web mining tool.

Among these web mining techniques, special attention should be paid to the clustering algorithms. The assumption is that, given a set of clusters extracted from data generated during former user sessions in the web site, it is possible to extract the users preferences by analyzing the clusters content. The patterns in each cluster detected would be sufficient to extrapolate the content that he or she is looking for [20, 23, 30].

In each IPV, the page component has a vectorial representation introduced by the Eq. (2). In this equation, an important step is the calculus of the weights consider in the special words array sw_i. The special words are different of a normal word in the site because they belong to an alternative and related source or they have an additional information showing their importance in the site, for instance and HTML tag emphasized a word .

The clustering algorithm is used for grouping similar IPVs by comparing the time and page components of each vector, being important to use the the similarity measure introduced in the Eq. (4). The results should be a set of clusters whose quality must be checked by using an accept/reject criterion. A simple way is to accept the clusters whose pages share a same main them and in otherwise to reject the cluster. In that point, it is necessary which page in the site is closet with the vectors in the cluster. Because we know the web site pages vectorial representation and using the Eq.(3) we can identify the closet page to a given cluster's vector and for this way to get the associate pages of the cluster and to review if the pages share a common main theme.

For each accepted cluster and remembering that the centroids contain the pages where the users spent more time during their respective sessions and in vectorial representation the special words have the highest weights, the procedure for identify the web site keywords is to apply a measure, described in the Eq. (5) (geometric mean) to calculate the importance of each word

$$kw[i] = \sqrt[\iota]{\prod_{p \in \zeta} m_{ip}} \quad (5)$$

where $i = 1, \ldots, R$, kw is an array containing the weights for each word relative to a given cluster and ζ the set of pages representing this cluster. The web site keywords are the result of sorting kw and to detect the words with highest weights, for instance the ten words.

5 Extracting patterns from data originated in a real web site

For experimental purposes, the selected web site should be complex with respect to several features: number of visits, periodic updating (preferably monthly in order to study the user reaction to changes) and rich in text content. The web page of a Chilean virtual bank (no physical branches, all transactions undertaken electronically) met these criteria. As noted, the authors signed a non-disclosure agreement with the bank and are not in a position to provide its name.

The main characteristics of the bank web site are the following; presented in Spanish, with 217 static web pages and approximately eight million raw web log registers for the period under consideration, January-March 2003.

The behavior of a user at the bank web site is analysed in two ways. First, by using web log files which contain data about visitor and customer browsing behavior. This data requires prior reconstruction and cleaning before web mining tools are applied. Second, web data is the web site itself, specifically the web page text content - this also needs preprocessing and cleaning.

5.1 Session reconstruction process

Fig 1 shows part of the bank's web log registers and includes both identified customers and anonymous vistors.

Customers access the site through a security connection, using a SSL protocol that allows the storage of an identification value in the authuser parameter in the web log file. Another way of identifying users is by cookies, but sometimes these are disactivated by users in their browsers. In this case it will be necessary to reconstruct the visitor session.

During the session reconstruction process, filters are applied to the web logs registers. In this particular case, only the registers requesting web pages are used to analyze the site specific user behavior. It is also important to clean abnormal sessions, for example web crawlers, as is shown in Fig. 1 line 4, where a robot that belongs to Google is detected.

The raw web logs cover four months of transactions, with approximately eight millions registers. Only registers related to web pages are considered for the session reconstruction and user behavior analysis purposes; information that points to other objects, like pictures, sounds, etc., will be cleaned.

5.2 Web page content preprocessing

By applying web page text filters, it was found that the complete web site contains R=2,034 different words to be used in the analysis. Regarding the word weights and the special words specification, the procedure introduced in section 3.2 was used, in order to calculate sw_i in equation 2. The data sources were:

1. Marked words. Inside the web pages, 743 different words were found after applying the preprocessing and cleaning step.

2. Related web sites. Four web sites were considered, each of them with approximately 300 pages. The total number of different words was 9253, with 1842 of them contained in the web site text content.

After identifying the special words and their respective weights, it is possible to calculate the final weight for each word in the entire web site, by applying the Eq. 2. Then, the vector representation for all the pages in the site is obtained.

5.3 Analyzing user text preferences

We fixed at 3 the vector's maximum dimension. Then, a SOFM with 3 input neurons and 32 output neurons was used to find clusters of Important Page Vectors.

Figure 2 shows the neurons positions within the SOFM on the x, y axes . The z axis is the normalized winning frequency of a neuron during training.

Figure 2, shows 8 main clusters which contain the information about the most important web site pages. However,

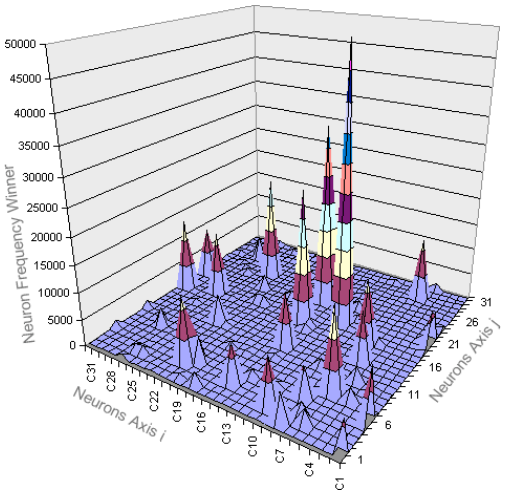

Figure 2. Clusters of important page vectors

only 5 were accepted. The accept/reject criterion is simple; if the pages in the cluster centroid have the same main theme, then the cluster is accepted - otherwise it is rejected.

The cluster centroids are shown in Table 1. The second column contains the center neurons (winner neuron) of each cluster and represents the most important pages visited.

Table 1. Important page vectors clusters

Cluster	Pages Visited
1	(16, 159,173)
2	(148,16,71)
3	(16,159,204)
4	(16,113,27)
5	(16,113,49)

A final step is required to get the web site keywords; to analyze which words in each cluster has a greater relative importance in the complete web site.

The keywords and their relative importance in each cluster are obtained by applying the equation 5. For example, if the cluster is $\zeta = \{16, 159, 173\}$, then $kw[i] = \sqrt[3]{m_{i16} m_{i159} m_{i173}}$, with $i = 1, \ldots, R$.

Finally, by sorting the kw in descendending order, we can select the k most important words for each cluster, for example $k = 5$.

We are not able to show the specific keywords because of the confidentiality agreement with the bank. For this reason the words are numbered. Table 2 shows the keywords found by the proposed method.

Table 3 shows a selected group of keywords from all clusters. Keywords on their own, however, do not make much sense. They need a web page context where they

#	IP	Id	A	Time	Method/URL/Protocol	Statu	Byte	Referer	Agent
1	164.77.129.50	-	-	12/Apr/2003:23:47:44	GET /img/tab.gif HTTP/1.1	200	89	http://www.thebank.cl	MSIE 6.0; Windows 98
2	200.28.206.200	-	20	12/Apr/2003:23:48:31	GET transa/info.htm HTTP/1.1	200	144	/infoeco/info.html	MSIE 4.01; Windows 95
3	200.86.248.170	-	-	12/Apr/2003:23:48:37	GET /img/gen.gif HTTP/1.1	304	0	/ofert/wines/	MSIE 6.0; Windows 98
4	66.249.65.97	-	-	12/Apr/2003:23:48:41	GET /index.htm HTTP/1.1	200	88	-	Googlebot/2.1; google.com/bot.html
5	216.241.8.179	-	31	12/Apr/2003:23:50:03	GET /tx/infoeco/card.htm HTTP/1.1	200	210	/tx/infoeco/prom/	MSIE 6.0; Windows NT 5.1
6	164.77.129.50	-	-	12/Apr/2003:23:48:34	GET /tx/infoeco/ HTTP/1.1	200	186	/tx/infoeco/card.htm	MSIE 6.0; Windows 98
7	200.28.206.200	-	20	12/Apr/2003:23:51:13	GET transa/account.htm HTTP/1.1	200	180	/transa/info.htm	MSIE 4.01; Windows 95
8	216.241.8.179	-	31	12/Apr/2003:23:51:23	GET /tx/infoeco/ind.htm HTTP/1.1	200	300	/tx/infoeco/card.htm	MSIE 6.0; Windows NT 5.1
9	200.86.248.170	-	-	12/Apr/2003:23:51:41	GET /prom/wine.html HTTP/1.1	404	0	/ofert/wines/	MSIE 6.0; Windows 98
10	164.77.129.50	-	44	12/Apr/2003:23:52:04	GET /tx/infoeco/ind.htm HTTP/1.1	200	186	/tx/infoeco/	MSIE 6.0; Windows 98

Figure 1. A raw web log file from the bank web site

Table 2. The 5 most important words per cluster

C	Keywords	kw sorted by weight
1	$(w_{2023}, w_{1233}, w_{287}, w_{1087}, w_{594})$	(2.35,1.93,1.56,1.32,1.03)
2	$(w_{1003}, w_{449}, w_{895}, w_{867}, w_{1567})$	(2.54,2.14,1.98,1.58,1.38)
3	$(w_{1005}, w_{948}, w_{505}, w_{1675}, w_{1545})$	(2.72,2.12,1.85,1.52,1.31)
4	$(w_{501}, w_{733}, w_{385}, w_{684}, w_{885})$	(2.84,2.32,2.14,1.85,1.58)
5	$(w_{200}, w_{1321}, w_{206}, w_{205}, w_{1757})$	(2.33,2.22,1.12,1.01,0.93)

could be used as special words, e.g. marked words to emphasize a concept or as link words to other pages.

Table 3. A part of the discovered keywords

#	Keywords	
1	Cuenta	Account
2	Fondo	Fund
3	Inversión	Investment
4	Tarjeta	Credit Card
5	Hipotecario	House credit
6	Seguro	Insurance
7	Cheques	Check
8	Crédito	Credit

The specific recommendation is to use the keywords as "words to write" in a web page, i.e., the paragraphs written in the page should include some keywords and some could be linked to other pages.

Further it is possible on the basis of this exercise to make recommendations about the text content. However, to reiterate, keywords do not work separately for they need a context. Reviewing Table 2, for each cluster, the discovered keyword could be used to rewrite a paragraph or an entire page. In addition, it is important to insert keywords to highlight specific concepts.

Keywords can also be used as index words for a search engine,i.e., some could be used to customize the crawler that visits web sites and load pages. Then, when a user is looking for a specific page in a search engine, the probability of getting the web site increases.

5.4 Improving the web site Text Content

Web site keywords are concepts to motivate the users' interests and make them visit the web site. They are to be judged within their context for as isolated words they may make little sense, since the clusters represents different contexts. The specific recommendation is to use the keywords as "words to write" in a web page.

Web site keywords can also be used as search engine index words, i.e., some of them could be used to customize crawlers that visit web sites and load pages. Then, when a user is looking for a specific page in the search engine, the probability of getting the web site increases.

As each page contains a specific text content, it is possible to associate the web site keyword to the page content; and from this suggest new content for site revision or reconstruction. For example, if the new page version is related to the "credit card", then the web site keywords "credit, points and promotions" must be designed for the rewritten page text content.

6 Conclusions

When the users visit a web site, there is a correlation between the maximum time spent per session in a page and its free text content. Then we created the "Important Page Vector (IPV)", which is the basic data structure for storing the pages where the user spent more time during his/her session. By using the IPV as input in a SOFM, we can to identify clusters that contain the user navigation and information preferences.

The cluster accept/reject criterion is simple: if the pages inside in each cluster are related with a similar main topic, then the cluster is accepted, else rejected. Applying this criterion, five cluster were accepted and the patterns contained in each on them were used for extracting the web site keywords.

The web page text contain can be improved by using the web site keywords, and for this way to attract the user attention when he/she is visiting the site. However, it is necessary to remember that these words cannot be used isolated, in fact they need a context, which usually is provided by a human being.

As future work, we will apply the methodology in other web data, for instance the images, in order to identify which elements attract the user attention in the web site.

Acknowledgment

This work was supported partially for the Millennium Institute on Complex Engineering Systems.

References

[1] E. Amitay and C. Paris. Automatically summarizing web sites: Is there any wayaround it? In *Procs. of the 9th Int. Conf. on Information and Knowledge Management*, pages 173–179, McLean, Virginia, USA, 2000.

[2] R. Baeza-Yates. *Web usage mining in search engines*, chapter Web Mining: Applications and Techniques, pages 307–321. Idea Group, 2004.

[3] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In *Proc. in First Int. Semantic Web Conference*, pages 264–278, 2002.

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

[5] D. Buttler. A short survey of document structure similarity algorithms. In *Procs. Int. Conf. on Internet Computing*, pages 3–9, 2004.

[6] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Focused web searching with pdas. *Computer Networks*, 33(1-6):213–230, June 2000.

[7] L. D. Catledge and J. E. Pitkow. Characterizing browsing behaviors on the world wide web. *Computers Networks and ISDN System*, 27:1065–1073, 1995.

[8] G. Chang, M. Healey, J. McHugh, and J. Wang. *Mining the World Wide Web*. Kluwer Academic Publishers, 2003.

[9] W. Chuang and J. Yang. Extracting sentence segment for text summarization? a machine learning approach. In *Procs. Int. Conf. ACM SIGIR*, pages 152–159, Athens, Greece, 2000.

[10] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Journal of Knowlegde and Information Systems*, 1:5–32, 1999.

[11] U. Hahn and I. Mani. The challenges of automatic summarization. *IEEE Computer*, 33(11):29–36, 2000.

[12] A. Joshi and R. Krishnapuram. On mining web access logs. In *Proc. of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 63–69, 2000.

[13] A. P. Jr and N. Ziviani. Retrieving similar documents from the web. *Journal of Web Engineering*, 2(4):247–261, 2004.

[14] D. Lawrie, B. W. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In *Proc. 24th Int SIGIR Conf. on Research and Development in Information Retrieval*, pages 349–357, New Orleans, Louisiana, USA, 2001. ACM Press.

[15] E. Liddy, K. McVearry, W. Paik, E. Yu, and M. McKenna. Development, implementation and testing of a discourse-model for newspaper texts. In *Procs. Int. Conf. on ARPA Workshop on Human Language Technology*, pages 159–164, Princeton, NJ, USA, 1993.

[16] G. Linoff and M. Berry. *Mining the Web*. Jon Wiley & Sons, New York, 2001.

[17] S. Loh, L. Wives, and J. P. M. de Oliveira. Concept-based knowledge discovery in texts extracted from the web. *SIGKDD Explorations*, 2(1):29–39, 2000.

[18] I. Mani and M. Maybury. *Advances in automatic text summarization*. MIT Press, Cambridge, Mass., 1999.

[19] S. Mitra, S. K. Pal, and P. Mitra. Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, 13(1):3–14, 2002.

[20] B. Mobasher, R. Cooley, and J. Srivastava. Creating adaptive web sites through usage-based clustering of urls. In *Procs. Int Conf IEEE Knowledge and Data Engineering Exchange*, November 1999.

[21] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.

[22] M. F. Porter. An algorithm for suffix stripping. *Program; automated library and information systems*, 14(3):130–137, 1980.

[23] T. A. Runkler and J. Bezdek. Web mining with relational clustering. *International Journal of Approximate Reasoning*, 32(2-3):217–236, Feb 2003.

[24] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM archive*, 18(11):613–620, November 1975.

[25] M. Spiliopoulou. Data mining for the web. In *Principles of Data Mining and Knowledge Discovery*, pages 588–589, 1999.

[26] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing*, 15:171–190, 2003.

[27] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[28] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. Recovering traceability links in multilingual web sites. In *Procs. Int. Conf. Web Site Evolution*, pages 14–21. IEEE Press, 2001.

[29] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. Restructuring multilingual web sites. In *Procs. Int. Conf. Software Maintenance*, pages 290–299. IEEE Press, 2002.

[30] J. D. Velásquez and V. Palade. A knowledge base for the maintenance of knowledge extracted from web data. *Journal of Knowledge Based Systems (Elsevier)*, page to appear, 2007.

[31] J. D. Velásquez, S. Ríos, A. Bassi, H. Yasuda, and T. Aoki. Towards the identification of keywords in the web site text content: A methodological approach. *International Journal of Web Information Systems*, 1(1):11–15, March 2005.

[32] J. D. Velásquez, R. Weber, H. Yasuda, and T. Aoki. A methodology to find web site keywords. In *Procs. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service*, pages 285–292, Taipei, Taiwan, March 2004.

[33] J. D. Velásquez, H. Yasuda, and T. Aoki. Combining the web content and usage mining to understand the visitor behavior in a web site. In *Procs. 3^{th} IEEE Int. Conf. on Data Mining*, pages 669–672, Melbourne, Florida, USA, November 2003.

[34] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. Using the kdd process to support the web site reconfiguration. In *Procs. IEEE/WIC Int. Conf. on Web Intelligence*, pages 511–515, Halifax, Canada, October 2003.

[35] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. A new similarity measure to understand visitor behavior in a web site. *IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization*, E87-D(2):389–396, February 2004.

[36] J. Xiao, Y. Zhang, X. Jia, and T. Li. Measuring similarity of interests for clustering web-users. In *ADC '01: Proceedings of the 12th Australasian conference on Database technologies*, pages 107–114, Washington, DC, USA, 2001. IEEE Computer Society.

[37] K. Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Procs. Int. Conf. on Computational Linguistics*, pages 986–989, 1996.