Web Site Keywords: A Methodology for Improving Gradually the Web Site Text Content

Juan D. Velásquez

Department of Industrial Engineering, University of Chile,
E-mail: jvelasqu@dii.uchile.cl

Abstract

The construction of a web site is a great challenge that integrates different elements such as the hyperlink structure, colors, pictures, movies and textual contents. In the latter, the correct textual content can be the key to attracting users to visit the site. In fact, many users visit a web site by using a web search engine such as, for example, Google or Yahoo!, and continue exploring the site if it contains the information that they are looking for. In this paper a methodology will be identified in which pages in a web site can further attract the users attention when he/she is browsing the site .These words are called web site keywords and by using them in the site textual content, significant improvements, from the point of view of the user, can be achieved.

A web users browsing behavior can be classified as two categories: amateurs and experienced. The former is a user with little or no experience in using web-based systems. Their browsing behavior is normally erratic and it can take them a considerable amount of time to find what they are looking for. The latter is a user with a greater amount of experience with web-based systems whose behavior is more controlled and purpose driven and thus takes them less time in determining whether the site contains worthwhile information. What is important regarding the experienced web users is there is a correlation between the amount of time spend on a webpage during a session and the extent to which they are interested in the page content. By using this characteristic, a feature vector is created with relation to the time spend on each page during the users session. The described vectors are the input for two clustering algorithms: SOFM and K-means, which allow one to extract significant patterns about users with similar or identical browsing behavior and content preferences. Next the patterns form the basis for identifying the web site keywords. In order for validating the proposed methodology, the data originated in a complex web site belonging to a Chilean bank, were used. From the clusters identified, a set of web site keywords were identified and their utility was tested on a group of human being users, thus illustrating the effectiveness of the proposed methodology.

Key words: Web site keywords, web site text content, web usage mining, web content mining

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 differentiates between e-business success and failure lies in the respective websites potential in both attracting and retaining users. This potential heavily depends on the site content, design, and technical aspects, such as the time to download the pages from the web site to the users web browser. In terms of the content, the words used in the free text of a web site pages are very important, as the majority of the users make term-base queries in a search engine to seek out information on the Web. These queries are formed by keywords, i.e., a word or a set of words Lawrie et al. (2001) that characterize the content of a given web page or web site.

The suitable utilization of words in the web page improves user appeal, boosts effective information search, while at the same time, attracts new users and retains current users through persistent updating of page text content. So the challenge is to identify which words are important for users. Most keywords are selected from "most frequently used words". Some commercial tools [1] help to identify target keywords which customers are likely to use while web searching Buyukkokten et al. (2000).

By identifying the most relevant words in the sites pages, from the point of view of the user, improvements can be achieved within the entire web site. For instance, the site could be restructured creating a new hyperlink related with the keyword, and furthermore the text content could be modify by using the keywords related to a 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, which is also called web data, specify web log registers and web site text content.

This 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" Velásquez et al. (2005) and can be used for the creation of further web page text contents related to a specific topic. Assuming that there is a correlation between user interest and the maximum time spent

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

on each page during the users session, the analysis of word site keywords has adopted two principal approaches:

(1) For experienced users, i.e., users with some experience using web-based systems, the amount of spent time on a page has a direct relation to their interest in the page content, specifically in the textual content, which can be represented through particular words in the page.
(2) The web site designer defines some words as "special" due to the fact that different fonts are used or a particular tag is applied, for example the $< title >$ tag.

This paper is organized as follow. Section II introduces the main approaches to analyzing web data. The minimum steps for preparing the web data is for it to be inputted in a web mining algorithm, which is introduced in section III. The proposed methodology for identifying web site keywords is explained in section IV and the application for this methodology on web data originated from a real web site, which is shown in section V. Finally, section VI highlights the main outcome of this work.

2 Related work

When a user visits a web site, data regarding the pages visited are stored in the web log files. Then pages visited are determined from those that were not, and the information includes the time spent by a user on each page. Due to the fact that the pages contain information regarding a specific topic, it is possible to extract the users precise information preferences. In this sense the interaction between user and site is like an electronic inquest, that provides us with the necessary data for analyzing the user content preferences in a particular web site.

Analyzing the user text preferences in the site is an enormous challenge. First, the amount of web log registers usually can be huge, most of which is irrelevant information regarding the user browsing behavior in the site. Second, the free text within the web pages is commonly plain, i.e., lacking the additional information that allows us to determine which words attract the users' attention.

In this section the main approaches to analyzing web data for the purpose of extracting significant patterns which is related to the users' text preferences in a web site are reviewed.

2.1 Mining web data

Web mining techniques emerged directly from the application of data mining theory to pattern discovery from web data Chang et al. (2003); Linoff and Berry (2001); Spiliopoulou (1999). 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 Srivastava et al. (2000).

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 within web log files.
- User profile. A collection of information concerning a userss personal information name, age, etc.), usage information (e.g. visited pages) and interest.

Taking into consideration the above definitions and depending on the web data to be processed, web mining techniques can be grouped into three areas: Web Content Mining (WCM), Web Structure Mining (WSM), and Web Usage Mining (WUM) Velásquez and Palade (2008).

In WCM, the objective is to find useful information from web documents. In this sense, it is similar to Information Retrieval (IR) techniques. However, web content is not only limited to the text, but includes other objects like pictures, sounds and movies. But of course, due to the majority of the users perform queries in search engines to find relevant textual information in the Web, the web page text analysis applies a large degree of the effort on WCM.

A web text content can be unstructured as free text, semi-structured as HTML or fully structured as a table or database. Each one of these formats has different levels of metadata, which could to be adopted for identifying significant information for the web user, in particular, which texts contain meaningful words or concepts that attract the user attention, which is constantly updated to improve the web site content Ríos et al. (2006); Velásquez et al. (2005).

WSM uses the hyperlink structure Chakrabarti et al. (1999) to analyze hyperlinks within web pages from a set of web sites. Web sites are considered, in this analysis, as a directed graph, in which the pages are the vertex and the

hyperlinks pointing to others vertex. A means of measuring the popularity of the page in the website could be by counting the number of hyperlinks. So in this sense it work like a bibliography citation system, i.e., a frequently cited paper is an important paper. This evaluation could provide us with important insight into the web page content quality, under the assumption that if a web page contains links to other pages or websites then it contains valuable information for the web community and because some search engines use this assumption to rank pages, it allows the user to extrapolate a page with a good ranking, i.e., a website entitled "very popular in the community", will undoubtedly attract more web users.

The interest in WUM is growing quickly in the scientific and commercial communities, possibly due to its direct application to web personalization and the growing complexity of web sites Velásquez and Palade (2008). In general, WUM uses traditional data mining methods to deal with usage data. However, some modifications are necessary due to the different types of usage data.

The aim of WUM is to discover patterns using different kinds of data mining techniques, (statistical analysis, association rules, clustering, classification, sequential patterns and dependency modelling Joshi and Krishnapuram (2000); Srivastava et al. (2000). Each WUM technique needs a model of user behavior per web site in order to define a feature vector to extract behavior patterns. Usually, the model contains the sequence of pages visited during the user session and some usage statistics, like the time spent per session, or page, etc.

2.2 Identifying words for creating an automatic web page text summarization

The aim here is to automatically construct summaries of a natural-language document Hahn and Mani (2000). 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.) Amitay and Paris (2000).

In text summarization research, the three major approaches are Mani and Maybury (1999): 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 Mitra et al. (2002) 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 renders 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 Chuang and Yang (2000); Zechner (1996). It is clear that the resulting text may not be cohesive, but the aim of the technique 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 Buyukkokten et al. (2000). The final approach is a discourse model based on extraction and summarization Lawrie et al. (2001); Liddy et al. (1993) by using natural language cues such as proper name 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 of 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, or word that contain significant information about a particular topic, from the web site user‘s point of view. The identification of these components can be useful for improving the website text content.

Usually, the keywords in a web site are correlated with “most frequently used words”. In Buyukkokten et al. (2000), 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 Baeza-Yates (2004), collects keywords from a search engine. This indicates the global word preferences of a web community, but contains no details about a particular web site. Finally, instead of analyzing words, in Loh et al. (2000) 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 allow the extraction of 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

Velásquez et al. (2003a). The web log registers contain information concerning the page navigation sequence and the time spent at each page visited, through applying the **sessionization** process. The web page source is the web site itself. Each web page is defined by its content, in particular by its free text. To study user behavior both data sources - web logs and web pages have to be prepared by using filters and by estimating real user sessions. The preprocessing stage involves, firstly, a cleaning process and, secondly, 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** Cooley et al. (1999). It is based on web log registers and is due to the problems mentioned above, the process is not exempt of errors Spiliopoulou et al. (2003). Sessionization uses the data contained in the web logs registers, then it is not possible to determine whether the visitor has pressed the "back" button on the browser. If the page is in the browser cache and the visitor returns 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 capturing the exact visitor browsing have been proposed Berendt et al. (2002); Cooley et al. (1999). However, this scheme could be easily avoided by the visitor.

Many authors Berendt et al. (2002); Cooley et al. (1999); Mobasher et al. (1999) 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 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 this sense, whatever the chosen strategy adopted 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 also belong to the same group.

There are several techniques for sessionization, which can be grouped in two major strategies: *proactive* and *reactive* Spiliopoulou et al. (2003).

Proactive strategies aim to identify the user using identification methods like cookies and these consist in 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 visitors privacy. First, if the site is revisited after several hours, the session will be considered too long, it will actually be considered as a new session. Secondly, some aspects of the cookies seem to be incompatible with the principles of the data protection, like in the European Union Spiliopoulou et al. (2003). 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 only contained in the web log files and process the registers to generate a set of reconstructed sessions.

In web site analysis, the general scenario is that the web sites usually do not implement identification mechanisms. The utilization of reactive strategies can be more useful. They can be classified into two main groups Berendt and Spiliopoulou (2001); Cooley et al. (1999):

- 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 Catledge and Pitkow (1995). 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, which is 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 tasks 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 there are restrictions. This method is used in Tonella et al. (2001) for comparing pages written in different languages with similar HTML structure. The comparison is enriched by applying a text content matching process Tonella et al. (2002), which considers a translation task to be completed first. The method is highly effective when the same language is used in the pages in comparison. A short survey of algorithms for comparing documents by using structural similarities is found in Buttler (2004).

Comparisons are made by a function that returns a numerical value highlighting 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 rely on efficiency criteria in the web page content processing Jr and Ziviani (2004). Here the vector space model Salton et al. (1975), allows a simple vectorial representation of the web pages and, by using a 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 to render 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 then 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 Porter (1980)), we acquire 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,

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

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

Based on *tfxidf-weighting* introduced in Salton et al. (1975) the weights are estimated as,

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

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. Fig. 1, shows the special words detection for marked words using HTML tags.

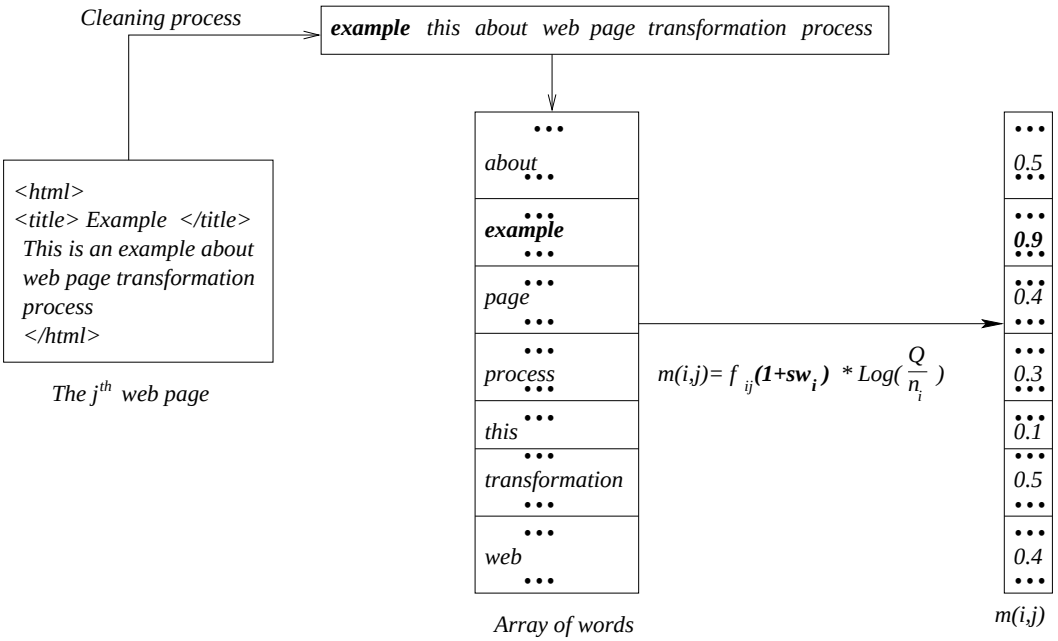

Fig. 1. Including the importance of special words in the vector space model

The common sources of special words are:

(1) E-Mails. The offer of sending 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.

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* $\mathbf{WP^k} = (wp^k_1, \ldots, wp^k_R) = (m_{1k}, \ldots, m_{Rk})$, $k = 1, \ldots, Q$ *, 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

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

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 searching for.

By using mainly a combination of a WUM and WCM approaches, it is essential to analyze the web user text preferences in a web site and to identify which words attract the users attention throughout their navigation of the site.

Prior to the application of a web mining tool, the data related to web user behavior must be processed to create feature vectors, whose components will

depend on the implementation of the mining algorithm that are to be used and the preference patterns that are 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. In Xiao et al. (2001), given a web site S and a group of users $U = \{u_1, \ldots, u_m\}$ who visit a set of pages $P = \{p_1, \ldots, p_n\}$ in a certain period of time, the feature vector is characterized by a usage function $use(p_i, u_j)$ that associates a usage value between a page p_i and a user u_j, such us

$$use(p_i, u_j) = \begin{cases} 1 \; if \;\; p_i \; has \; been \; visited \; by \; u_j \\ 0 \; otherwise, \end{cases}$$

The feature vector is $v = [use(p_1, u_k), \ldots, use(p_n, u_k)]$ for $k = 1, \ldots, m$.

Along the same lines, in Joshi and Krishnapuram (2000), each URL in the site is designed by a unique number $j \in \{1, \ldots, N_U\}$, where N_U is the total amount of real URLs. A user session is represented as $v = [s_j^{(1)}, \ldots, s_j^{(N_U)}]$ where

$$s_j^{(i)} = \begin{cases} 1 \; if \; the \; user \; visited \; the \; j^{th} \; URL \; during \; the \; i^{th} \; session \\ 0 \; otherwise. \end{cases}$$

More advanced models also consider the time spent in each page visited by the user. In Mobasher et al. (2000), the entire web site is considered as a set of URLs $U = \{url_1, \ldots, url_n\}$, and the users' transactions is the set $T = \{t_1, \ldots, t_m\}$. Then, for the user *"i"*, the associated transactions are $t_i \in T / t_i = \{u_1^{t_1}, \ldots, u_n^{t_n}\}$ where

$$u_j^{t_i} = \begin{cases} 1 \; if \;\; url_j \in t_i \\ 0 \; otherwise. \end{cases}$$

The result is a vector of "0" or "1" that represents the URLs visited per user.

In Cooley et al. (1999), $t =< ip_t, uid_t, \{(l_1^t.url, l_1^t.time), \ldots, (l_m^t.url, l_m^t.time)\} >$ is a general user transaction where ip_t is the IP address, uid_t a user identifier, $l_i^t.url$ is the url of i^{th} page visited and $l_i^t.time$ the time stamp of the transaction *"t"*. A similar approach is used in Wong et al. (2001), where the term "User Transaction" is defined by the set of URLs visited and the

time spent in each of them during the user session. Here the feature vector is $U = [(URL_1, d_1), \ldots, (URL_n, d_n)]$, with d_j the time spent on URL_j.

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, as it is used to locate interesting information related to a particular topic through a search engine. Hence, it is necessary to include a new variable as part of the web user behavior feature vector information concerning 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 book, 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.

Figure 2 shows an eight page web site, where the user browsing data is stored in the associated web log file. After a session reconstruction process, the visitors usage data associated to the IP address **1.2.3.4** allows to create the vector $\upsilon_1 = [(1, 3), (2, 40), (6, 5), (5, 16), (8, 15)]$.

As the time-stamp parameter in the web log only shows the precise moment when the web object is requested, the time spent by the user on each page visited is calculated by the difference in time spent between the visits to two consecutive pages in the same session. This calculation can be problematic for the final page visited, as the user may have exited the site, which cannot be determined exactly. There are a number of solutions to this problem. First, apply the average time spent at other pages visited during the user session; second, set up a maximum duration with the user in inactivity status after which it is assumed to be a finished session status. For example if there is a security protocol for acessing a web site and there is no interaction between the user and the site for a determined period, then the site finishes the session, by sending a special page to the user as shown in Figure 2, (on the right side).

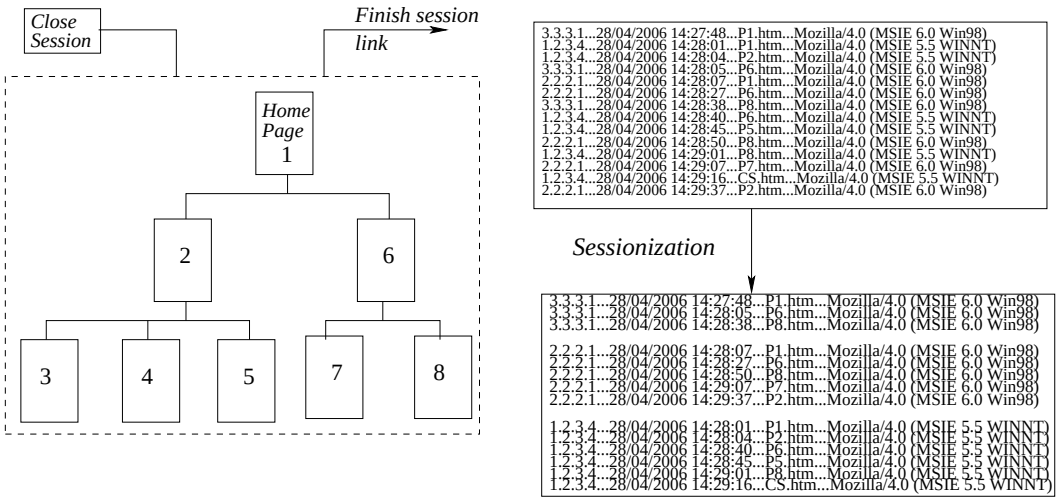

Fig. 2. User behavior vector's creation

It is helpful to examine the "finish session link", i.e., a link used voluntarily to finish his session by the user, (see Figure 2, with the page "close session").

Now it is clear that the length of a session depends on user browsing navigation - and this generates vectors of different lengths. Some mining algorithms require vectors with the same cardinality, and then the UBV will need to set the length, i.e., the n value. One solution is to set n to the average number of pages visited by the users. If so, the vectors whose length is less than n are classified as having null values; if not, vectors will be trimmed to the first n values in the user session.

The UBV can be used as input of web mining algorithms, preferentially those related to clustering and classification techniques. In both cases, the comparison between vectors is essential, i.e., it is necessary to have a similarity measure that shows how similar or different two vectors are and compare the three elements - page sequence, page content and time spent - that form a UBV.

4.2 Analyzing the user text preferences

The aim is to determine the most important words for users at a given web site by comparing the user text preferences through the analysis of pages visited and the time in which it takes on both Velásquez et al. (2003b). 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, Velásquez et al. (2003b,a, 2004b) 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:

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

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 $\mathbf{1}$. In the extreme opposite case, it will be near $\mathbf{0}$. 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"* Velásquez et al. (2004a). It is interesting to note that the same web site keywords may be used by the user in a search engine one 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 is a relation between the time spent by the user on a page and ones interest in

its content Velásquez et al. (2005). This relation is collected by the Important Page Vector (IPV), given that one acquires 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 the former users 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 the user is looking for Mobasher et al. (1999); Runkler and Bezdek (2003); Velásquez and Palade (2007).

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 user 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 theme and in otherwise the cluster is rejected. It is necessary to consider which page on the site is closest to the vectors in the cluster. Due to the fact that we know the web site pages are vectorial representation and by using the Eq. (3) we can identify the closet page to a given clusters vector and to this respect to acquire the associate pages of the cluster and to review it if the pages share a common main theme.

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

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

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 author signed a non-disclosure agreement with the bank but are not in a position to provide its name.

The main characteristics of the bank web site are listed as followed; presented in Spanish, with 217 static web pages and approximately eight million raw web log registers for the period under consideration, January-March. The bank web site was designed for two types of users - visitors and customers. A visitor is defined as an anonymous user, normally looking for information concerning credits, investments, information on how to open an account, etc. Regarding the visitor, the bank web site consists of information pages which can be accessed by any user. A bank customer has access to a special, hidden part of the web site, via https, an encrypted version of the http.

The confidentiality agreement does not allow us to reveal the structure of the customer zone. The other two options are accessible to visitors. The third levels shows information about products and services and the fourth contains specific information on production.

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.

Before applying any task, it is helpful to recall the dynamic nature of a web site. Often when web page content has changed, the pages physical web site name is kept unchanged. For instance, the “index.html page may have had ten changes in the last month but continues being called “index.html. This situation can complicate tracking page content changes but it is important for the understanding of user information preferences. This problem can be avoided by maintaining the older page versions and recording the date when any change took place.

5.1 Session reconstruction process

Fig 3 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. 3 line 4, where a robot that belongs to Google is detected.

#	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

Fig. 3. A raw web log file from the bank web site

The raw web logs data covers four months of transactions, with approximately eight millions registers. Only registers related to web pages are considered for session reconstruction and user behavior analysis purposes; information that points to other objects, like pictures, sounds, etc., will be cleaned. Finally approximately 30.000 real user sessions were identified for creating the IPV to be used as input of the web mining algorithms.

5.2 Web page content preprocessing

Each web page in the site is related with a very specific topic. The pages in the web site contain a very specific main topic and they can grouped as show the Table 1.

By applying web page text filters, it was found that the complete web site contains R=2,035 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) The e-mails received in the call center platform. During the period under analysis, 2128 e-mails were received, containing a total of 8125 words

Table 1
Web site pages and the main theme

Pages	Main theme
1,2,14,18,24,79, 26,28	Home page (the complete versions)
25,30–45, 84,132,143,144,153,182, 190,191,194	Voluntary retirement savings
62,74,85,160	Commercial agreements
19–23	Bank account premium customers
3-11,129,170	Bank advertisements
51,66,114,115	Special promotions
90,91,116,128,137,161,166,212	Bank account plain customers
31,117,138,147,167,200	Credit card point discount
29,48,60,61,67–71,80–83,92-102,118–125,130,,148,155–158,162,163,176	Raising services using different products
103-106,198,204,210	Car credit, home credit, simple credit
12,13,16,17,27,46–50,57–59,65,75–78, 86-89.95,96,113,126,127,133–136,145, 154,164,165,173–175,192	Online services (electronic transactions, bank statement, credit card statement, insurance, other related accounts)
52-55,63,140,141,183,184,187–189, 195,197,199,201–208	National and international investments, savings economic indicators, stock market information
15,72,73, 107,108,149–152, 168 169,171,177–179,180,185,193	Complementary services
64,131,159,211	Credits simulation
109-112,142,181,196,207,209	Credit Card

following preprocessing and cleaning. From this set of words, only 1037 are to be found in the web site text content.

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

(3) Related web sites. Four web sites belonging to others bank institutions were considered, each of them with approximately 300 pages. The total number of different words was 9253, with 1142 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

Simply put, the purpose of web mining algorithms is to extract patterns about the user navigation behavior and text content preferences. So a prior operation is required that creates user behavior vectors as an input into the algorithms. The vectors are created from user sessions and which are reconstructed by using data contained in the web logs. In order to construct the Important Page Vectors (IPV), a number of three elements were fixed, i.e., the ι parameter is equal to 3.

For extracting significant patterns which allow to identify the web site keywords, and also for comparing results purpose, two clustering were used: Self Organizing Feature Maps (SOFM(and K-Means. Next sections show the results of applying the mentioned algorithms.

5.3.1 Extracting web user text preferences by using SOFM

SOFM with 3 input neurons and 32 output neurons was used to find clusters of Important Page Vectors. Figure 4 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 4, shows 7 main clusters which contain the information about the most important web site pages. However, 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 2. The second column contains the center neurons(winner neuron) of each cluster and represents the most important pages visited.

The cluster content analysis is as follow:

- **Cluster 1**. It is related with information concerning credit cards. The respective main theme per page in the centroid are: how to use the promotions given by the credit card (use of the credit card discount points); the use of credit card in national/international purchase and information about special services on the credit card. The cluster is accepted because the information contained in each page is related with the same main theme.
- **Cluster 2**. This cluster contain pages related with information regarding retirement programs, for example, investment funds for pensions, personal

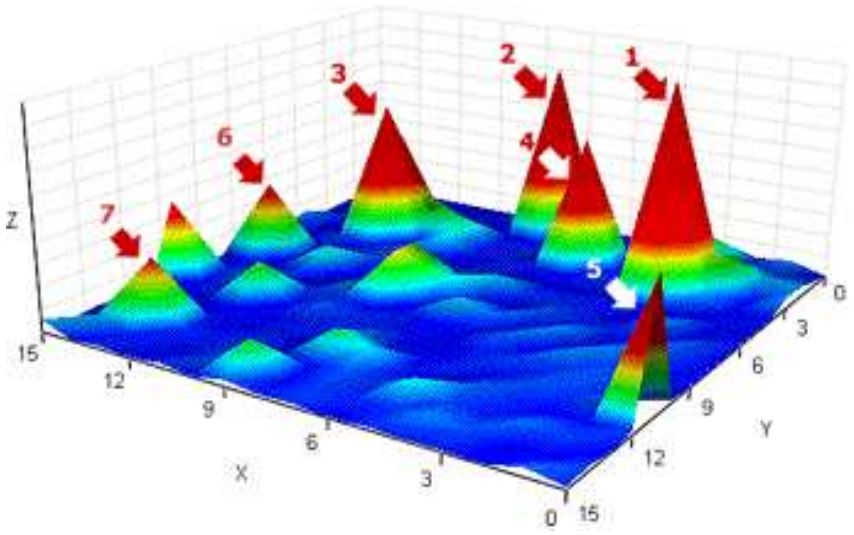

Fig. 4. Clusters of important page vectors

Table 2
Important page vectors clusters detected by using the SOFM algorithm

Cluster	Pages in cluster
1	(117, 192, 129)
2	(130, 39, 64)
3	(126, 59, 76)
4	(58, 64, 10)
5	(195, 193, 198)
6	(171, 159, 199)
7	(192, 155, 208)

money saving and long term credit simulation. The cluster is clearly accepted because it contains highly related information.

- **Cluster 3**. It is related to summaries regarding online services in the bank, for instance the account statement, electronic transactions, other related accounts (for instance a bi-personal or family account) etc. If the pages in the cluster are correctly related, then the cluster is accepted.
- **Cluster 4**. The clusters pages are related to different ways of obtaining a credit and of providing the user with a respective estimation of how much they will be expected to pay per month. Also information is provided concerning how to pay past debts from other institutions through the repay bank credit, i.e., the bank pays the user debt, and what follows is that the user pays back the bank. The cluster is accepted because the information contained is highly related.
- **Cluster 5**. This cluster contains pages associated with investment and housing credit. Because the contain relation between the pages in the cluster is

not high, the cluster is rejected.

- **Cluster 6**. The pages in this cluster are definitively unrelated, because the main topics are: complementary services (for instance, medical insurance), raising services and online services. Here, the cluster is rejected.
- **Cluster 7**. The contents associated to the cluster can be interpreted as the user's intention for exploring foreigner markets. In the pages appear information concerning the prospect of opening foreigner account, the possible investments, some economics indicators and the main entrance requirements for dealing with foreigner accounts. Because the information inside each page in the cluster is related with the same main theme, the cluster is accepted.

After applying the accept/reject criteria, a final step is required to get the web site keywords; to analyze which words in each accepted 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 $\zeta = \{130, 39, 64\}$, then $kw[i] = \sqrt[3]{m_{i130} m_{i39} m_{i64}}$, with $i = 1, \ldots, R$.

Finally, by sorting the kw in descendending order, we can select the l most important words for each cluster, for example $l = 8$.

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 3 shows the keywords found by the proposed method.

Cluster	Keywords	kw sorted by weight
1	$(w_{2023}, w_{1233}, w_{287}, w_{4087}, w_{594}, w_{587}, w_{1575}, w_{257})$	(2.35,1.93,1.56,1.32,1.03,0.92,0.83,0.76)
2	$(w_{1003}, w_{449}, w_{895}, w_{867}, w_{1667}, w_{1456}, w_{767}, w_{458})$	(2.54,2.14,1.98,1.58,1.38,1.03,0.91,0.83)
3	$(w_{1005}, w_{2048}, w_{505}, w_{1675}, w_{1545}, w_{556}, w_{543}, w_{654})$	(2.72,2.12,1.85,1.52,1.31,0.95,0.84,0.74)
4	$(w_{501}, w_{733}, w_{385}, w_{684}, w_{885}, w_{1326}, w_{1434}, w_{1564})$	(2.84,2.32,2.14,1.85,1.58,1.01,0.92,0.84)
7	$(w_{8}, w_{1254}, w_{64}, w_{878}, w_{238}, w_{126}, w_{1338}, w_{343})$	(2.51,2.12,1.41,1.22,0.98,0.95,0.9,0.84)

Table 3
The 8 most important words per cluster

5.3.2 Extracting web user text preferences by using K-means

The web site keyword methodology, is based on a clustering algorithms for identifying clusters, and for this way to extract significant words candidate to be web site keywords. Next by applying the above explained accept/reject criteria, the cluster is validate and the web site keywords extracted.

In order to validate the web site keyword methodology, another clustering

algorithm was applied on the same web data. It was the well know K-means and because the number of clusters identified by using the SOFM were 7, the parameter k in K-means was fixed on 7 too.

Table 4
Important page vectors clusters detected by using the K-means algorithm

Cluster	Pages in cluster
1	(211, 72, 1)
2	(104,112,205)
3	(110,9,117)
4	(48,126,21)
5	(28, 211, 191)
6	(63,188,135)
7	(104,87,64)

The cluster content analysis is as follow:

- **Cluster 1**. The pages in the centroid are not well related. The main themes are credit simulation, complementary services and the information contained during the analysis period in the home page. Because the relation between these pages is not clear, the cluster is rejected.
- **Cluster 2**. In this cluster, two different contents share the same words, but within different context, i.e., the word credit. The first page is concerning the credit card, the second one is regarding normal credit. Finally, the third is related to investments. The relation among the clusters pages are unclear and thus the cluster is rejected.
- **Cluster 3**. This cluster contains pages highly related to credit card information and therefore the cluster is accepted.
- **Cluster 4**. In this cluster, the user is focused on obtaining information related to the products contracted with the bank regarding the specification of bank account statements, credit card statements, etc. The pages are well related by content, therefore the cluster is accepted.
- **Cluster 5**. The cluster content is related with retirement programs. The users are interested in how the Chilean retirement system works and in which way the bank can help them. In the Chilean retirement system, part of the money saved by an individual can be invested on the stock market and the other part in long term deposits. The users wish to know what the bank can offer them with regarding to investing this money.
- **Cluster 6**. In this cluster, the pages are related to information concerning national and international investments, the stock market indicators and bank products statements, such as account statements, investments statements, etc. The cluster page content is well related and the cluster is accepted.

- **Cluster 7**. This cluster is related mainly to the banks credits, specifying the home credit and the plain credit. The users also are performing some simulations for estimating how much they must pay in the case of credit. Because the pages are related by content, the cluster is accepted in this case.

5.3.3 Comparing clustering algorithms and results

The methodology introduced for extracting web site keywords require a clustering algorithms for identifying first which pages contains interesting text that corresponds with the web users tastes. In both cases, K-means and SOFM allow the identification of clusters in the Important Page Vectors data set. The clusters identified are composed of different pages from the web site under observation. However, the web pages textual content is comparable, which allows for the comparison of clusters generated from K-means and SOFM respectively.

Each cluster generated by using SOFM is compared with the set of clusters produced by using K-means. The similarity between two clusters (generated separately by each algorithm) is produced when the textual content of the pages in a cluster is closer to the content of another cluster's pages. Then we discover each cluster generated by the SOFM locates its homologous in a cluster produced by the K-mean, such as show the table 5.

Table 5
Comparing clusters generated by SOFM and K-mean

Clusters	SOFM	K-mean
1-3	(117, 192, 19)	(110,9,117)
2-5	(130, 39, 64)	(28, 211, 191)
3-4	(126, 59, 76)	(48,126,21)
4-7	(58, 64, 10)	(104,87,64)
7-6	(192, 155, 208)	(63,188,135)

By using the textual information contained in each cluster and by applying the web site keyword extraction step, a set of these words are shown in Table 6. However keywords that stand on their own do not make sense, They need a web page context where they could be employed as key words, e.g. marked words to emphasize a concept or as link words to other pages.

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.

Table 6
A part of the discovered keywords

#	Keywords	
1	Cuenta	Account
2	Ahorro	Saving
3	Promoción	Promotion
4	Tarjeta	Credit Card
5	Hipotecario	House credit
6	Seguro	Insurance
7	Puntos	Points
8	Crédito	Credit

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 3, 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. When a user is looking for a specific page in the search engine, the probability of locating the web site increases drastically.

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.

5.5 Testing the text content recommendation effectiveness

The application of any recommendation will need the web site owner's agreement as some users may dislike the changes, which could become a potential risk for the business. Unless carefully handled there is the danger that "*the cure might be worse than the disease*" and users migrate to other web sites.

In the case of a virtual bank and others where the web site is the core business, customer loss due to web site modifications can only be tolerated within a narrow range and only if it can be shown to retain existing customers and attract new customers in a very short time period. So the loss potential must be estimated by some a priori test, discussed in the following sections.

As noted above web site keywords must be seen in context. To test the effectiveness of web site keywords, understood as the capacity to attract user attention during a web page session, a textual fragment such as a paragraph, should be created. These texts are a data source for web site keyword identification and although it is possible to use fictional examples, it was decided to use texts belonging to the web site itself as alternatives could unwittingly exaggerate attention to the test's detriminent. Therefore web texts were used so that a similar stylistic and information environment were maintained.

Five paragraphs were selected from the bank web site. Two contained the greatest number of web site keywords while the others were extracted randomly. All examples were shown to the same amateurs and experienced users.

Table 7 shows the results of the web site keyword effectiveness test. The users showed a good receptivity toward paragraphs that contained the keywords, considering them interesting and with relevant information. So for the user, the particular words contain important information - the words, particularly web site keywords attract user attention.

Web site keywords can guide the web site designer about the specific text content. Of course, the utilization of the keywords does not guarantee the success of the paragraph, for it must be combined with elements such as semantic content, style and the paragraph meaning, all important for transmitting the message to the users.

Table 7
Testing the web site keyword effectiveness

#	Including the web site keyword?	Acceptability opinion				
		Irrelevant	Moderately irrelevant	Some information	Moderately relevant	Relevant
1	Yes				3	2
2	Yes			1	2	2
3	No	2	2	1		
4	No		3	2		
5	No		4	1		

6 Conclusions

When experienced customers visit a web site, there is a correlation between the maximum time spent per session in a page and its free text content. Hence the concept of web site keyword defined as a word or set of words that attract the visitor. These keywords then convey information concerning visitor web site preferences. The most important pages by session are identified by arranging visitor behavior vectors by time component. So the "Important Page Vector (IPV) can be defined and a new similarity measure applied to find clusters among these vectors.

Clusters of IPVs were found by using SOFM and K-means separately. The results of both algorithms were comparable showing the effectiveness of the approach used for identifying web users with similar content preferences. In that sense, amateur web users, due to their erratic browsing behavior, do not create clusters. On the other hand, experienced web user, are grouped in content preferences.

To understand what information the experienced web users are looking at, one must allow for the improvement the web site textual content for both sets of users. Indeed the experienced users can find the information more quickly and the amateur users have an efficient and oriented way of finding information. Something like "learning from the experienced user to help the amateur users.

Web site isolated keywords do not make sense, they these need a context to be applied. For instance if the cluster is regarding credit cards, the web site keywords extracted are a means to creating a textual content on pages related to that topic.

The methodology proposed is based on a representation of the web page textual content in the vector space model. Because it is term-based, two words

used within a different content can signify the same. For instance "the course finished explained how to prepare the main course", in this case the word "course" for vector representation purposes is the same, but through the content it serves a different meaning. In a future work, the inclusion of semantic models on a web page representation could be a potential solution to this problem and would clarify the specific textual content in a page.

Acknowledgment

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

References

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

Baeza-Yates, R., 2004. Web usage mining in search engines. Idea Group, Ch. Web Mining: Applications and Techniques, pp. 307–321.

Berendt, B., Hotho, A., Stumme, G., 2002. Towards semantic web mining. In: Proc. in First Int. Semantic Web Conference. pp. 264–278.

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

Buttler, D., 2004. A short survey of document structure similarity algorithms. In: Procs. Int. Conf. on Internet Computing. pp. 3–9.

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

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

Chakrabarti, S., Dom, B., Kumar, S., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., Kleinberg, J., August 1999. Mining the web's link structure. IEEE Computer 32 (8).

Chang, G., Healey, M., McHugh, J., Wang, J., 2003. Mining the World Wide Web. Kluwer Academic Publishers.

Chuang, W., Yang, J., 2000. Extracting sentence segment for text summarization? a machine learning approach. In: Procs. Int. Conf. ACM SIGIR. Athens, Greece, pp. 152–159.

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

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

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

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

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

Liddy, E., McVearry, K., Paik, W., Yu, E., McKenna, M., 1993. Development, implementation and testing of a discoursemodel for newspaper texts. In: Procs. Int. Conf. on ARPA Workshop on Human Language Technology. Princeton, NJ, USA, pp. 159–164.

Linoff, G., Berry, M., 2001. Mining the Web. Jon Wiley & Sons, New York.

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

Mani, I., Maybury, M., 1999. Advances in automatic text summarization. MIT Press, Cambridge, Mass.

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

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

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

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

Ríos, S., Velásquez, J., Yasuda, H., Aok, T., 2006. Using a self organizing feature map for extracting representative web pages from a web site. International Journal of Computational Intelligence Research 1 (2), 159–16.

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

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

Spiliopoulou, M., 1999. Data mining for the web. In: Principles of Data Mining and Knowledge Discovery. pp. 588–589.

Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M., 2003. A framework for the evaluation of session reconstruction heuristics in web-usage analyis. INFORMS Journal on Computing 15, 171–190.

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

Tonella, P., Ricca, F., Pianta, E., Girardi, C., 2001. Recovering traceability links in multilingual web sites. In: Procs. Int. Conf. Web Site Evolution. IEEE Press, pp. 14–21.

Tonella, P., Ricca, F., Pianta, E., Girardi, C., 2002. Restructuring multilingual web sites. In: Procs. Int. Conf. Software Maintenance. IEEE Press, pp. 290–299.

Velásquez, J., Palade, V., 2007. A knowledge base for the maintenance of knowledge extracted from web data. Journal of Knowledge-Based Systems (Elsevier) 20 (3), 238–248.

Velásquez, J. D., Palade, V., 2008. Adaptive Web Sites: A Knowledge Extraction from Web Data Approach. IOS Press.

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

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

Velásquez, J. D., Yasuda, H., Aoki, T., November 2003a. 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. Melbourne, Florida, USA, pp. 669–672.

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

Velásquez, J. D., Yasuda, H., Aoki, T., Weber, R., February 2004b. 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.

Wong, C., Shiu, S., Pal, S., 2001. Mining fuzzy association rules for web access case adaptation. In: In Workshop on Soft Computing in Case-Based Reasoning Research and Development, Fourth Int. Conf. on Case-Based Reasoning (ICCBR 01).

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

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