# Towards the Identification of Keywords in the Web Site Text Content: A Methodological Approach

JUAN D. VELÁSQUEZ[1], SEBASTÍAN RÍOS, ALEJANDRO BASSI[2],
HIROSHI YASUDA AND TURUMASA AOKI
*Research Center for Advanced Science and Technology, University of Tokyo, Tokyo, Japan*
*Email: {jvelasqu,srios,yasuda,aoki}@mpeg.rcast.u-tokyo.ac.jp*

***Abstract*— Since the creation of the web, the designers are looking for friendlier ways of make web page contents, which pictures, sounds, movies and free texts attract the users' interest. Special attention receive the text content, because is the most frequently parameter used to retrieve information from the web. A simple way in order to understand the user's text preferences, could be collect the words used in a searching. However, this information is only well-know for the owner of the specific searching engine. In this paper we introduce a methodology in order to extract the most interest words for a user in a particular web site, based of the user browsing behavior and the web page text content. The methodology was tested using data originated in a bank web site showing the effectiveness of our approach.**

***Index Terms*— web site keywords, text preferences, web usage mining**

## I. INTRODUCTION

A correct content in a web site allows its users to find the information they are looking for [6]. This is a non trivial problem because users' content preferences change frequently and is not clear which is the best way to define the correct content of a site [3].

Among the different kinds of content, we are interested in the free text of each page. The main idea is to define a methodology to find words in the text that attract the users' interest, i.e., the keywords of the web site [10].

A search engine like Google or Altavista, receives every day millions of queries, each one of them with keywords used to find a specific topic in the Web [4]. Usually is not possible to access the information about these keywords in the search engine, so an alternative way is necessary.

In this paper we propose to extract web site keywords from the user behavior in a web site, in order to improve the web page content [11]. We assume there is a correlation between the interest of a visited page and the time spent on it by a user in his/her session. By using clustering algorithms we group similar user behaviors and pages preferences. Based in this information, it is possible to extract the approximate keywords for a group of users.

[1] Also at the Department of Industrial Engineering, University of Chile.

[2] Center for Collaborative Research, University of Tokyo, Tokyo, Japan. Email: abassi@vp.ccr.u-tokyo.ac.jp. Also at the Department of Computer Science, University of Chile.

This paper is organized as follow. Section 2 provides an overview on related work. In section 3 we describe the methodology used to find web site keywords. Section 4 describes the application of our work for a Chilean bank. Section 5 concludes this work and points at future extensions.

## II. RELATED WORK

A methodology developed to extract important words from a web site must try to understand what text content is most significant given the user preferences [5].

In this paper, a combination of information retrieval techniques with web usage mining is proposed in order to infer the user text preferences in a web site.

### A. Extracting important words

Usually, the notion of "important words" in a web site have been related with the "most frequently used words".

In [1] these words are collected from a searching engine. It shows global words preferences from a web community, but no details about a particular web site.

In [3] a method to extract important words from a huge set of web pages is introduced. The technique is based on assigning importance to words, depending on their frequency in all documents. In this approach a vector space processing is applied, i.e., cleaning of stop words and stemming reduction.

### B. Web site keywords

In [10] the web site keywords concept was defined as *" a word or possibly a set of words that is used by web users in their search process and characterizes the content of a given web page or web site"*.

In order to find the web site keywords it is necessary to select the web pages whose text content is more significant for the users. The assumption is that there exists a correlation between the time that the user spent in a page and his/her interest in its content.

### C. Applying the vector space model on web pages

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

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

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

This weight must capture the fact that a given word can be more important than another one. For instance, if the word $i$ appears in $n_i$ documents, the expression $\frac{n_i}{Q}$ gives a sense of its importance in the complete set. The "inverse document frequency" $IDF = \log(\frac{Q}{n_i})$ can be used like a weight.

The last expression is known as TF*IDF (term frequency times inverse document frequency). A specially adapted variation of it [10] is shown in equation 2.

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

where $f_{ij}$ is the number of occurrences of the $i^{th}$ word in the $j^{th}$ page, $sw(i)$ is a factor to increase the importance of special words and $n_i$ is the number of documents containing the $i^{th}$ word. A page $p_j$ is represented by the column $j$ in $M$, i.e., $p_j \rightarrow (m_{1j}, \ldots, m_{Rj})$ and the distance between pages is

$$pd(p_i, p_j) = \frac{\sum_{k=1}^{R} m_{ki} m_{kj}}{\sqrt{\sum_{k=1}^{R} (m_{ki})^2} \sqrt{\sum_{k=1}^{R} (m_{kj})^2}}. \quad (3)$$

### D. Web log processing

A web log file contains information on the access of all users to a particular web site in chronological order. In a common log file, each access to one of its pages is stored together with the following information: IP address and agent, Time stamp, Embedded session Ids, Method, Status, Software Agents, Bytes transmitted, Objects required (page, pictures, etc).

Based on such log files we have to determine for each user the sequence of web pages visited in his/her session. This process is known as *sessionization* [2], [8]. It considers a maximum time duration given by a parameter, which is usually 30 minutes in the case of total session time. Based on this parameter we can identify the transactions that belong to a specific session using streams and program filters.

## III. A METHODOLOGY TO DISCOVER WEB SITE KEYWORDS

After the sessionization process it is possible to reconstruct the user session and extract the pages where the user spent more time during his/her session.

### A. A relation between page interest and spent time

The users can be grouped in two classes: amateur and experienced. The first one corresponds to persons not familiarized with a particular web site and probably with the web technology. Their behavior is characterized by an erratic browsing and sometimes they don't find what they are looking for.

The second one are users with experience in the site or other related sites and with the web technology. Their behavior is characterized by spending few time in pages with low interest and concentrating in the pages they are looking for, where they spend a significant amount of time.

As amateurs gain experience, they slowly become experienced users. Only experienced users are aware of the features of a particular web site, therefore any recommendations must be based on them.

The clustering techniques allow to group users with similar behavior, it is then possible to find clusters with amateur and experienced users, and with the support of a business expert, parasite clusters can be filtered out.

### B. Important web pages and web user session

To select the most important pages, it is assumed that the degree of importance is correlated with the percentage of time spent on each page within a session.

Using the information extracted from the sessionization process, we sort the sessions according to the percentage of time spent on each page, the first $\iota$ pages correspond to the $\iota$ most important pages.

*Definition 1 (Vector of the $\iota-$Most Important Pages):* $\vartheta_\iota = [(\rho_1, \tau_1), \ldots, (\rho_\iota, \tau_\iota)]$, where the pair $(\rho_\iota, \tau_\iota)$ represents the $\iota^{th}$ most important page and the percentage of time spent on it within a session.

Let $\alpha$ and $\beta$ be two most important page vectors. A similarity measure between two vector of the $\iota-$most important pages is defined as:

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

where the term $\min\{\cdot, \cdot\}$ indicates the ratio between the percentages of time spent on the pages visited by user $\alpha$ and $\beta$, and the term $dp$ is the similarity measure (3). The time factor allows us to distinguish between pages with similar contents, but corresponding to different users' interests.

### C. Discovering keywords from clusters

A clustering technique can be applied to find groups of similar user sessions. The most important words for each cluster are determined by identifying the cluster centroids. The importance of each word with respect to each cluster is calculated by

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

for $i = 1, \ldots, R$, where $kw$ is an array containing the geometric mean of the weights of each word within the pages contained in a given cluster. Here $\zeta$ is the set of pages contained in the cluster, and $m_{ip}$ is expressed by (2), by sorting $kw$ in descendent order the most important words for each cluster can be selected.

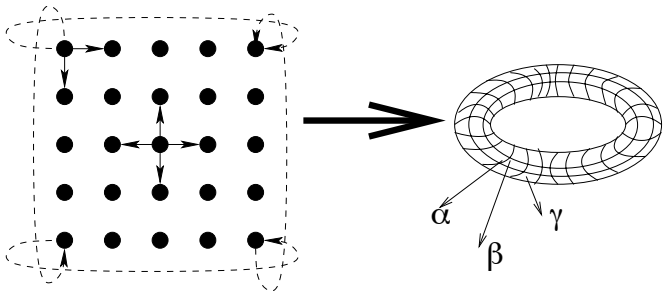


Fig. 1. Proximity of important page vectors in a SOFM whit thoroidal topology.

### D. A SOFM as clustering technique

An artificial neural network of the Kohonen type (Self-organizing Feature Map; SOFM) has been applied to the preprocessed data originated in the web transactions. Schematically, a SOFM is represented as a two-dimensional array of neurons. Each neuron is constituted by an n-dimensional vector, whose components are the synaptic weights. By construction, all the neurons receive the same input at a given moment.

The notion of neighborhood among neurons defines diverse topologies. In this case a toroidal topology was used [9], which means that the neurons located on an edge are close to those located on the opposite edge (see Figure 1). This characteristic has shown its advantages when it is necessary to maintain the continuity of the clusters or when the data corresponds to sequences of events, like the user behavior in a web site.

## IV. Practical Application

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

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

The web site selected[1], belongs to the first Chilean virtual bank, i.e., it doesn't have physical branches and all the transactions are made using electronic means, like e-mails, portals, etc.

We have the following information about the web site:

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

### A. Sessionization process

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

[1]http://www.tbanc.cl/

Only 16% of the users visit 10 or more pages and 18% less than 4. The average number of visited pages is 6, thus we fixed 6 as the cardinally of the User Behavior vector.

We chose 3 as the maximum number of components of the Important Page vector, i.e., the parameter $\iota = 3$.

Finally, applying the above described filters, approximately 300,000 vectors were identified.

### B. Web page content processing

Applying web page text filters, we discovered that the complete web site contains $R$=4023 different words for our analysis.

To calculate $sw_i$ in equation 2, we have three sources in the particular web site in the study:

1) E-mails. The web site offers the option to send e-mails to the call center. The text sent is a source to identify important words. Let $ew_i = \frac{w^i_{e-mail}}{TE}$ be the array of special words within e-mails, where $w^i_{e-mail}$ is the frequency of $i^{th}$ word and $TE$ is the total number of words in the complete set of e-mails.
2) Marked words. A web page contains words with special tags, e.g., a different font like italic or a word belonging to the title. Let $mw_i = \frac{w^i_{marks}}{TM}$ be the array with the marked words inside the web site, where $w^i_{mark}$ is the frequency of the $i^{th}$ word and $TM$ is the total number of words in the whole web site.
3) Searched words. The web site offers also a search engine, i.e., a system by which the users can search for specific subjects, typing in keywords. Let $aw_i = \frac{w^i_{ask}}{TA}$ be the array with the words used in this search engine, where $w^i_{ask}$ is the frequency of $i^{th}$ word and $TA$ is the total number of words in the complete set of words.
4) Related web site. Usually a web site belongs to a market segment, in this case the banks market segment. Then it is possible to collect web site pages belonging to the others web site in the same market. Let $rw_i = \frac{w^i_{rws}}{RWS}$ be the array with the words used in the web sites of the market, where $w^i_{rws}$ is the frequency of $i^{th}$ 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 using the above described methods.

### C. Applying Self-Organizing Feature Maps

We used a SOFM with three input neurons and 32*32 neurons on the feature map. The toroidal topology maintains the continuity of the cluster space, which allows to study the transition among the preferences of the users from one cluster to another.

We trained the neural network on a Pentium IV, with 1 Gb in RAM and running Linux Operating System, distribution Redhat 8.0. The necessary time was 25 hours and the epoch parameter was 100.

TABLE I
IMPORTANT PAGE CLUSTERS CENTROIDS

| Cluster | Pages Visited |
|---|---|
| 1 | (7,15,186) |
| 2 | (110,130,45) |
| 3 | (91,154,101) |
| 4 | (115,102,1) |
| 5 | (3,9,147) |
| 6 | (108,131,62) |
| 7 | (81,201,144) |
| 8 | (161,172,191) |
| 9 | (87,178,141) |
| 10 | (161,175,209) |

TABLE II
PAGES AND THEIR CONTENT

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

### D. Results

In order to obtain the relevant clusters, the Web pages were labeled according to its main topic. Next, they were checked to verify that the pages associated to a cluster follow a unique topic. Applying this criterion and with the collaboration of the business expert, ten clusters were accepted. They are shown in table I. The second column contains each cluster's centroid neurons (winner neurons), representing the most important pages visited.

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

Applying equation 5, we obtained the keywords and their relative importance in each cluster. For instance, if $\zeta$ is the set of pages representing cluster 4, then $\zeta = \{115, 102, 1\}$, and $kw[i] = \sqrt[3]{m_{i115} m_{i102} m_{i1}}$, with $i = 1, \ldots, R$.

Finally, sorting $kw[i]$ we were able to select a subset of the most important words of each cluster.

From the keywords found we randomly selected eight, which are shown in Table III. Given the confidentiality agreement, we are not allowed to show their respective weights.

The web site keywords represent a set of concepts that could motivate the user interest to visit the web site. Their use as isolated words don't make much sense, since the clusters represent different context through a set of keywords.

The specific recommendation is to use the keywords as "word to write" in a web page, i.e., the paragraphs written in the page should includes some keywords and even some of them may be used as links to other pages.

The keywords could also be used as index words in a search engine, i.e., some of them could be used in the customization of the crawler that visit the web site and load the pages. Then, when a user is looking for a specific page in the search engine, the probability of getting the web site will increase.

TABLE III
SOME OF THE IDENTIFIED KEYWORDS

| # | Keywords |
|---|---|
| 1 | Credit |
| 2 | House-credit |
| 3 | Points |
| 4 | Card |
| 5 | Contest |
| 6 | Promotions |
| 7 | Concourse |
| 8 | Account |

### E. Comparing with an alternative approach

In [3] (see section II-A) a method to extract important words is introduced. It was applied applied on 20 million web pages and the final result was 22,390 special words candidates to be keywords.

The testing process consist on identified sentences in the set of pages that contain special words. Next, a set of sentences is extracted and shown to a group of users.

As a practical result, 70% of the sentences were accepted by the users, extrapolating a similar percentage of keywords identified.

The web site keywords introduced in this paper combine two important concepts:

1) For experienced users, the spent time in a page has direct relation with their interest, in particular with the words contained in the page.
2) The web site designer defines some words as "special" since different fonts are used or a particular tag is applied, for instance $< title >$ tag.

Because a clustering algorithm is used to group pages with similar importance for a set of users, the keywords found can represent the user text preferences.

Twelve cluster were found, but only eight of them accepted by the business expert. Next, 64 special words were extracted and 83% of them accepted as web site keywords.

Also it is important to note that the web site keywords approach allows a more reliable user text preference classification. In average per cluster, the acceptance of the keywords is a 91%.

## V. CONCLUSIONS

We introduced a methodology to find significant words in a web site, called "web site keywords". It is based on the assumption that there exists a correlation between the visited pages as well as the time spent per session on them, and the point of view of the users, i.e., the more interesting pages are those where the user spent more time in his/her session.

Using a clustering algorithm and a distance to compare user text preferences, we find clusters whose centroids are used for extracting the web site keywords.

To use the keywords in an isolated way doesn't make sense, they represent a set of significant words for the user. In that case, the recommendation is to use a combination of the keywords to create a new web page.

As future work, it is necessary to improve the distance introduced in this work adding semantic information to compare text content between web pages.

## REFERENCES


[1] R. Baeza-Yates (2004) *Web Mining: Applications and Techniques*, Query Usage Mining in Search Engines, pp. 307–321. Idea Group, Anthony Scime (ed.).

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

[3] O. Buyukkokten, H. Garcia-Molina and A. Paepcke (2001) Seeing the whole in parts: text summarization for web browsing on handheld devices. *Procs. 10th Int. Conf. on World Wide Web*, Hong Kong, pp. 652–662.

[4] M. Chau and H. Chen (2003) *Web Intelligence*, Personalized and Focused Web Spiders, pp. 197–217, Berlin: Springer-Verlag, N. Zhong, J. Liu, Y. Yao (eds.).

[5] S. Loh, L. Wives and J. P. M. de Oliveira (2000) Concept-based Knowledge Discovery in Texts Extracted from the Web. *SIGKDD Explorations*, **2**(1): 29–39.

[6] J. Nielsen (1999) User interface directions for the web. *Communications of ACM*, **42**(1): 65–72.

[7] T. A. Runkler and J. C. Bezdek (2000) Automatic Keyword Extraction with Relational Clustering and Levenshtein Distances. *IEEE International Conference on Fuzzy Systems*, pp. 636–640.

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

[9] J. D. Velásquez, H. Yasuda, T. Aoki, R. Weber and E. Vera (2003) Using self organizing feature maps to acquire knowledge about visitor behavior in a web site. *Lecture Notes in Artificial Intelligence*, **2773**(1): 951–958.

[10] J. D. Velásquez, R. Weber, H. Yasuda and T. Aoki (2004). A Methodology to Find Web Site Keywords. *IEEE Int. Conf. on e-Technology, e-Commerce and e-Service*, March, Taipei, Taiwan, pp. 285–292.

[11] J. D. Velásquez, H. Yasuda and T. Aoki (2004) Web Site Structure and Content Recommendations. *IEEE/WIC Int. Conf. on Web Intelligence*, Beijing, China, pp. 636–639.



**Juan D. Velásquez** received his BE in electrical engineering and BE in computer science in 1995, PE in electrical engineering and PE in computer engineering in 1996, Masters in computer science and Masters in industrial engineering in 2001 and 2002, respectively, from the University of Chile, Chile and his PhD from the University of Tokyo, Japan in 2005. In 1996, he joined the academic staff of the Physics and Mathematics Sciences Faculty at University of Chile as a part time lecturer. In 2004 he was incorporated as a full-time academic staff member of the Industrial Engineering Department (DIE) and promoted to Assistant Professor. He also served as the Executive Director Manager of the Informatics Management and e-Business professional engineering programs from 1997 to 2001. In the professional area, he has been consultant by several ministries of the Republic of Chile and software companies in Latin America. His research interests include data mining, web mining and very large data bases.

**Sebastían A. Ríos** is a doctoral student at the RCAST of the University of Tokyo, Japan. He received the BE on Industrial Engineering on 2001, the BE on Computer Science and the PE on Industrial Engineering on 2003 from the University of Chile, Chile. He has been lecturer since 2002 on the Department of Industrial Engineering of the University of Chile. His research interests consist in Data Mining, Web Mining and Web Semantics.

**Alejandro Bassi** received his BE and Master in Computer Science from University of Chile, Chile in 1985 and 1988, respectively, and the Dr. Informatique, Université de Paris XI in 1995. Actually he is Assistant Professor with the Department of Computer Science. His research interests include natural language analysis, speech analysis and compilers.

**Hiroshi Yasuda** received his BE, ME and DrE from the University of Tokyo, Japan in 1967, 1969, and 1972, respectively. Since joining the Electrical Communication Laboratories of NTT, in 1972, he has been involved in work on video coding, image processing, tele-presence, B-ISDN network and services, Internet and computer communication applications. After serving twenty-five years (1972–1997), his final position being Vice President, Director of NTT Information and Communication Systems Laboratories at Yokosuka, he left NTT and joined the University of Tokyo. He is now Director of The Center for Collaborative Research (CCR). He had served as the Chairman of ISO/IEC JTC1/SC29 (JPEG/MPEG Standardization) from 1991 to 1999, as well as the President of DAVIC (Digital Audio Video Council) from September 1996 to September 1998. He received the 1987 Takayanagi Award, the 1995 Achievement Award of EICEJ, the 1995–1996 EMMY award from The National Academy of Television Arts and Science, and the 2000 Charles Proteus Steinmetz Award from IEEE. He is a Fellow of IEEE, EICEJ, and IPSJ, and a member of Television Institute.

**Terumasa Aoki** is a lecturer with the Research Center for Advanced Science and Technology, the University of Tokyo. He received his BS, ME and PhD in information and communication from the University of Tokyo, Japan in 1993, 1995, and 1998, respectively. His current research interests are in the fields of terabit IP router, access control of gigabit LAN/WAN, next-generation video conferencing system, high-efficiency image coding, and management of digital content copyrights. He has received various academic excellent awards such as the 2001 IPSJ Yamashita award, the FEEICP Inose award for 1994, and the four other awards.