Web Site Structure and Content Recommendations

Juan D. Velásquez, Hiroshi Yasuda, Terumasa Aoki
Research Center for Advanced
Science and Technology, University of Tokyo
{jvelasqu,yasuda,aoki}@mpeg.rcast.u-tokyo.ac.jp

Abstract

Web sites have become necessary channels for effective marketing and efficient operation for almost every company. To maintain and update a site is, however, a non-trivial task due to the complexity of the underlying business and the dynamic visitor behavior. Tools for the reconfiguration of web sites offer help for such tasks. We propose a methodology for the reconfiguration of web structure and content. Its application to the web site of a Chilean bank shows its benefits.

1 Introduction

There is no doubt about it: In order to stay competitive a company needs a web site that has to be up-to-date, it should offer the information visitors are looking for, this information should be easily accessible, and - if that were not enough - all this should be performed in the most efficient way, hopefully automatically and online. Here the word "visitor" refers to the occasional user of a web site, when no personal information is available about her/him.

While we still need research in web intelligence and related areas in order to reach this goal, some achievements on this road have already been made [3, 8]. In this paper we contribute to the necessary developments presenting a methodology for the reconfiguration of web sites concerning two elements: improving the structure of the site by rearranging links between pages and improving the text content by identifying the most relevant keywords.

Section 2 of this paper provides an overview on related work. In section 3 we describe the preprocessing of web data before we present our methodology for web site reconfiguration in section 4. Section 5 contains its application to a bank's web site. Section 6 concludes this paper and points at future developments.

2 Related work

We propose a methodology for web site reconfiguration combining web usage mining and web content mining [2, 3]. In the following sub-sections we will sketch the current status of works related to the above mentioned areas.

2.1 Web site personalization

Personalization is the process for dynamically customizing the content (pages, items, browsing recommendations, etc.) shown to the user, based on information about his/her behavior in a web site [2, 4]. This is different to another related concept called "customization" where the visitor interacts with the web server using an interface to create her/his own web site, e.g., "My Banking Page".

2.2 Offline recommendations

While there are papers proposing online special guidance, e.g. [4], online recommendations are still not commonly used for web site reconfiguration.

Offline recommendations can be obtained following the same method, i.e., analyzing the visitor behavior in order to propose structural and content modification recommendations for the web site, which must be checked by a team of experts in the business (web master, managers, etc.)

3 Preprocessing of web data

We suggest reconfigurations based on the combined analysis of content and usage of the respective web site.

3.1 Preprocessing of web usage data

Based on the available log files we have to determine for each visitor the sequence of web pages visited in his/her session. This process is known as **sessionization** [1]. It considers a maximum time duration given by a parameter, which is usually 30 minutes in the case of total session time.

Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)
0-7695-2100-2/04 $ 20.00 IEEE

3.2 Representing the web site

We represent the web site by a vector space model, in particular by vectors of words. Let R be the number of different words in a web site and Q the number of web pages. A vectorial representation of the web site is a matrix $M(m_{ij})$ of dimension RxQ that contains the vectors of words in its columns, where m_{ij} is the weight of word i in document j.

In order to determine these weights, we use a variation of the *tfxidf-weighting* [6], defined by equation 1.

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

where f_{ij} is the number of occurrences of word i in page j and n_i is the number of pages containing word i. Additionally, we propose to increase the word importance if a particular word shows special characteristics, such as when visitors search for the word. This is done by sw (special words), which is an array of dimension R.

Definition 1 (Page Vector) $\mathbf{WP^k} = (wp_1^k, \ldots, wp_R^k) = (m_{1k}, \ldots, m_{Rk})$ with $k = 1, \ldots, Q$

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

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

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

3.3 Representing a web site visit

For both approaches presented in this paper we need a representation of the visitor behavior as introduced in the following definition.

Definition 2 (Visitor Behavior Vector) A visit to a web site is represented by $v = [(p_1, t_1) \ldots (p_n, t_n)]$, being (p_i, t_i) parameters that represent the page in the i^{th} visit and the time spent on it, respectively [8].

4 The proposed methodology for web site reconfiguration

Our methodology provides offline recommendations for web site improvements that are twofold. We propose an improved structure by recommending links between pages according to the results of our analysis. On the other hand we identify keywords, which gives the opportunity to improve the content of the site, e.g. by using the most attractive words.

4.1 Applying clustering techniques

A simple analysis of the visitor's browsing behavior in a web site shows that in a session only a subset of the total number of pages are visited. Then it is feasible to identify groups of comparable visitors sessions. In that sense, the browsing is similar to a purchase behavior, since the visitors are looking for products, services or information [4, 7, 8].

4.2 Improving the link structure

The analysis of the regularity of web site visits provides information on the behavior and priorities of the majority of the visitors. This information is crucial when reconfiguring the site's structure. To extract it, we define a similarity measure able to capture the navigation behavior of web site visitors. Then we show how to find clusters of similar visits using this measure. Finally, we present a set of generic offline suggestions for an improved structure of the site.

4.2.1 Similarity measure for structure

Let α and β be two visitor behavior vectors of dimension C^α and C^β, respectively. Let $\Gamma(\cdot)$ be a function that returns the navigation sequence corresponding to a visitor vector. A similarity measure has been proposed elsewhere to compare visitor sessions as follows [8]:

$$sm(\alpha, \beta) = dG(\Gamma(\alpha), \Gamma(\beta))\frac{1}{\eta}\sum_{k=1}^{\eta} \tau_k * dp(p_{\alpha,k}, p_{\beta,k}) \quad (3)$$

where $\eta = \min\{C^\alpha, C^\beta\}$, and $dp(p_{\alpha,k}, p_{\beta,k})$ is the similarity (2) between the k^{th}page of vector α and the k^{th}page of vector β. The term $\tau_k = \min\{\frac{t_{\alpha,k}}{t_{\beta,k}}, \frac{t_{\beta,k}}{t_{\alpha,k}}\}$ is an indicator of the visitor's interest in the pages visited. The term dG is the similarity between sequences of pages visited by two visitors [5, 8].

4.3 Identifying web site keywords

We first cluster the most important web pages of all visits in order to identify the most important pages of the site. As a second step we determine the most important words for each cluster [6]. For this purpose we need a similarity measure that is introduced in the following subsection.

4.3.1 A content based similarity measure

From the visitor behavior vector (see above) we want to select the most important pages, assuming that the importance is correlated to the relative time spent on each page [6]. This is done as follows.

We take the visitor behavior vector and sort it according to the percentage of total session time spent on each page. Then we select the ι most important pages, i.e. the first ι pages.

Definition 3 (Important Pages Vector) Let υ be a visitor behavior vector. $\vartheta_\iota(\upsilon) = [(\rho_1, \tau_1), \ldots, (\rho_\iota, \tau_\iota)]$ is called Important Pages vector, (ρ_k, τ_k) being the component that represents the k^{th} most important pages and the percentage of total session time spent on it, respectively.

Let α and β be two visitor behavior vectors. The proposed content based similarity measure of the two visitors is shown in equation 4:

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

The first element indicates the visitors interest in the pages visited. If the percentage of total session time spent by visitors α and β on the k^{th} page are close, the value of the expression will be close to **1**, else it will be close to **0**.

The second element is dp, the similarity of the two pages in the vectorial representation introduced in equation 2.

4.3.2 Identifying the keywords for each cluster found

Using the vector page model and equation 1, we propose the following method to determine the most important keywords and their importance in each cluster. Equation 5 shows a measure (geometric mean) used in order to calculate the importance of each word relative to each cluster.

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

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

5 Application of the proposed methodology

We applied the described methodology to the web site of the first Chilean virtual bank, where all transactions are made using electronic means, like e-mails, portals, etc. (see www.tbanc.cl). We analyzed the visits during the period January to March, 2003 when the site had 217 static web pages with texts written in Spanish and approximately eight million raw web log registers.

The pages were numbered to facilitate the analysis, see Table 1.

Only 16% of the visitors visited 10 or more pages and 18% less than 4. The average number of visited pages is 6, thus we fixed 6 as the cardinality of the Visitor Behavior vector, i.e., the parameter $H = 6$ and we eliminated visits with 3 or less pages. 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.

Table 1. 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

5.1 Extracting visitor browsing patterns

A SOFM was used for clustering web site visits, it has 6 input neurons and 32*32 output neurons in the feature map.

The resulting clusters are presented in more detail in table 2. The second and third column of this table contain the center neurons (winner neuron) of each of the clusters, representing the sequence of the visited pages and the time spent in each one of them, respectively.

Table 2. Visitor browsing behavior clusters

Cluster	Visited Page Sequences	Time spent in seconds
A	(72,90,155,188,141,97)	(3,65,38,6,70,95)
B	(158,170,184,110,105,1)	(5,85,115,110,35,10)
C	(1,5,10,151,147,190)	(10,70,180,110,191,151)
D	(100,108,122,128,41,64)	(20,75,44,64,102,22)

5.1.1 Web site structure recommendations

Together with an expert from the bank we developed a set of propositions, suggesting links to be added and/or to be eliminated from the current site, based on the above introduced generic scheme and the clusters of visitor behavior we found.

Adding links inside each cluster. Improving the accessibility of pages within each cluster from other pages belonging to the same cluster.

Adding links between clusters. Improving the accessibility of pages belonging to different clusters that have many visitors in common.

Eliminating links. Rarely used links between clusters with few visits in common can be eliminated.

Currently, these propositions are under evaluation in the bank for their final implementation on the web site.

5.2 Extracting visitor text preferences

Applying web page text filters, we found that the complete web site contains R=4,096 different words for our analysis. In order to calculate sw_i in equation 1, we used three sources in the particular web site in study.

1. The web site offers the option to send e-mails to the call center platform. The text sent is a source to identify the most frequently used words.

2. Marked words. Web pages contain words with special tags, e.g., a different font like italic or a word belonging to the title phrase.

3. Words used in the search engine. The web site offers also a search engine, i.e., a system by which the visitors can search for specific subjects, typing in their keywords.

5.2.1 Clustering important pages with a Self-Organizing Feature Map

A SOFM with 3 input neurons and 32*32 neurons on the feature map was used. In order to validate clusters, each page contains a description of its main topic. Then a cluster is accepted if it contains only related pages.

Applying this filter, 8 clusters that contain the information about the most important pages in the web site were identified and they are shown in table 3. The second column contains each cluster's center neurons (winner neurons), representing the most important pages visited.

Table 3. Important page clusters

Cluster	Pages Visited	Cluster	Pages Visited
A	(186,71,135)	B	(155,97,90)
C	(180,157,169)	D	(8,150,5)
E	(7,8,200)	F	(2,110,105)
G	(130,61,102)	H	(45,114,120)

Applying equation 5, we obtained the keywords and their relative importance in each cluster. For instance, if ζ is the set of pages representing cluster F, then $\zeta = \{2, 110, 105\}$, and $kw[i] = \sqrt[3]{m_{i2} m_{i110} m_{i105}}$, with $i = 1, \ldots, R$.

Finally, sorting $kw[i]$ we selected a group of the most important words in each cluster, for instance the 7 words with the highest weight. From the keywords we found, we randomly selected 6 examples which are (in Spanish): Seguro (assured), Descuento (discount) , Crédito (credit), Ahorro (save money), Cuenta (account), Inversiones (investments).

6 Conclusions and future work

We presented a methodology for the reconfiguration of web sites focussing on improving both structure and content. The results from usage analysis provide insights for structure improvements, whereas identifying the most important words helps to improve the content of a site. We applied this methodology to the case of a bank's web site where we showed its potential.

Future work consists in adding content other than free text to our analysis and applying the proposed methodology to different sites in order to get hints on further developments.

References

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

[2] Z. Lu, Y.Y. Yao and N. Zhong, Web Intelligence, *Springer-Verlag*, Berlin,2003.

[3] R. Kosala and H. Blockeel, Web Mining Research: A Survey, *SIGKDD Explorations*, Vol. 2, no 1, pages 1-15, 2000.

[4] B. Mobasher, H. Dai, T.Luo and M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, Vol. 6, pages 61-82, 2002.

[5] T. A. Runkler and J. Bezdek, Web Mining with Relational Clustering, *Int. Journal of Approximate Reasoning*, Vol. 32, No. 2-3, pages 217-236, 2003.

[6] J. Velásquez, R. Weber and H. Yasuda, A method to find Web Site Keywords, *Procs. of the IEEE Int. Conf. on e-Technology, e-Commerce and e-Service*, pages 285-292, Taipei, Taiwan, March, 2004.

[7] J. Velásquez, H. Yasuda, T. Aoki and Richard Weber, Using Self Organizing Feature Maps to acquire knowledge about visitor behavior in a web site, *Lecture Notes in Artificial Intelligence*, Vol. 2773, No. 1, pages 951-958, 2003.

[8] J. Velásquez, H. Yasuda, T. Aoki and R. Weber, A new similarity measure to understand visitor behavior in a web site, *IEICE Transactions on Information and Systems*, Vol. E87-D, No. 2, February, 389-396, 2004.