Intelligent web site: Understanding the visitor behavior

Juan D. Velásquez[1], Pablo A. Estévez[2,3], Hiroshi Yasuda[1],
Terumasa Aoki[1], and Eduardo Vera[4]

[1] Research Center for Advanced Science and Technology, University of Tokyo
{jvelasqu,yasuda,aoki}@mpeg.rcast.u-tokyo.ac.jp
[2] Center for Collaborative Research, University of Tokyo,
pestevez@vp.ccr.u-tokyo.ac.jp
[3] Department of Electrical Engineering, University of Chile
[4] AccessNova Program, Department of Computer Science,
University of Chile, esvera@accessnova.cl

Abstract. Intelligent web site is a new portal generation, able to improve its structure and content based on the analysis of the user behavior. This paper focuses on modeling the visitor behavior, assuming that the only source available is his/her browsing behavior. A framework to acquire and maintain knowledge extracted from web data is introduced. This framework allows to give online recommendations about the navigation steps, as well as offline recommendations for changing the structure and contents of the web site. The proposed methodology is applied to the web site of a commercial bank.

1 Introduction

Since the creation of the world wide web, researchers have been looking for friendlier ways of navigating web sites [3]. However, the creation of preferences is still a problem without a complete solution. Here the word visitor refers to the occasional user of a web site, where no personal information is available about her/him. It is a difficult task due to the lack of data to characterize the visitor of a web site, in contrast to work with identified users, where additional variables like sex, age, last purchase, etc., are known.

A new generation of web sites is appearing, the so-called intelligent web sites, i.e, *"sites that automatically improve their organization and presentation by learning from visitor access patterns"* [7]. Its implementation has been addressed by several initiatives [3]. A consensus approach [6] is to combine artificial intelligence, user modeling and web mining algorithms for creating intelligent web sites. The intelligent capacity suggests the ability of modifying the web site structure and its contents based on the individual user behavior.

In this paper the visitor behavior in a web site is modeled. A framework to acquire and maintain information and knowledge from web data is introduced. This framework constitutes the core of our intelligent web site proposal.

This paper is organized as follows. In section 2 the main concepts about intelligent web sites are presented. Section 3 describes the proposed framework for intelligent web site. The techniques for processing web data are explained in section 4. A real world application is shown in section 5, and the conclusions are presented in section 6.

2 Main concepts

The intelligent web site has two kinds of potential users:

- Individual users, visitors in our case, that receive personalized recommendations based on their own interests.
- Web site operators, mainly web masters, that receive recommendations about changes to the web site.

The success or failure of an intelligent web site depends on the users' satisfaction. The challenge is not minor, because there are several aspects in the environment of intelligent web sites that may affect their feasibility and performance.

There are two categories of changes in a web site: structural and content. The structural ones include the addition and elimination of links. The content ones are mainly free text modifications, although the variation of other objects like colors, pictures, etc, can be considered too.

Due to the risk of applying directly the changes proposed by an automatic system, it is preferred to give recommendations to the web users. Recommendations can be grouped into two categories: Online and Offline. The Online recommendations principally consists in navigation suggestions displayed at the bottom of the web page [4]. It is a non-invasive scheme where the user has the possibility of following the suggestion or not. Offline recommendations are targeted to the web master. These include the addition or elimination of links, and changes in the web content. It is a non-invasive scheme too, where the web master can accept or reject the recommendations. These recommendations are based on the analysis of the visitor browsing behavior and her/his content preferences. This analysis is performed by processing web logs registers and web pages. Both sources of data require preprocessing such as cleaning irrelevant data and consolidating data for the application of web mining techniques. The final goal is to get meaningful patterns that describe the visitor behavior in a web site.

3 Intelligent web site framework

In the figure 1 a framework for acquiring, maintaining and using knowledge about the visitor behavior is shown. On the left hand side of this figure, three repositories can be observed: Information Repository (IR), Pattern Repository (PR) and Rule Repository (RR). The IR stores the data to be analyzed, the PR keeps the results of these analyses, and the RR contains domain knowledge

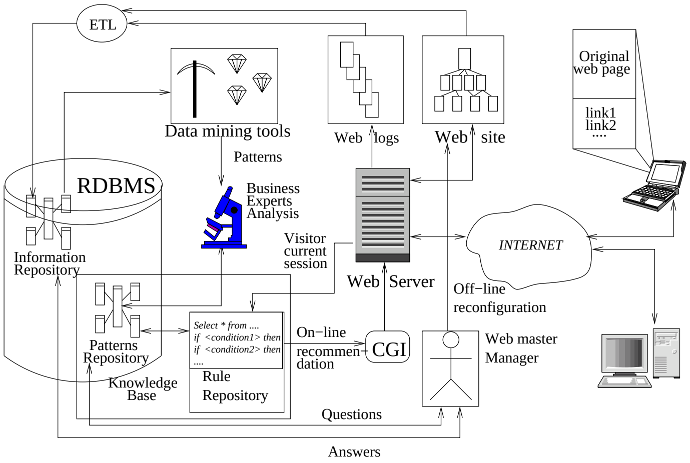

Fig. 1. An intelligent web site framework

drawn from experts. These two final structures conform the Knowledge Base about the visitor behavior. This framework allows to suggest online navigation steps, as well as offline changes to the web site structure and the web contents.

The IR can be implemented under the data mart technology applying the star model. It contains information extracted from the web data, such as the visitors sessions (page visited, spent time, page sequence, etc.) and the web page contents. By construction, the repository stores historical information, allowing the direct application of a web mining tool in any period of time. Applying web mining techniques to the IR, it is possible to discover unknown and hidden knowledge about the visitor browsing behavior and his/her preferences [10].

The behavior patterns extracted by the web mining tools should be first validated by a business expert and then loaded into the PR. The specific uses of the behavior patterns are implemented as rules, and loaded into the RR. The PR and RR constitute the complete structure of the Knowledge Base [5], which is used to give recommendations. Because both repositories are historical, the future impact of a set of web changes could be extrapolated from what happened with the visitor behavior when similar changes were made in the past.

The approach introduced has two kinds of potential users: human beings and artificial systems. The human beings consult the Knowledge Base as a Decision Support System and propose changes in the web site. These changes are usually made manually, although part of them can be automated. In the second case, the artificial systems use the PR and return navigation recommendations as a set of links to web pages. In the figure 1, the CGI[5] represents the typical interface between the web server and other system. Dynamic web pages can incorporate these links in the information to be sent to the visitors.

[5] Common Gateway Interface

4 Visitor behavior patterns extracted from web data

Before applying web mining techniques the data is transformed into behavior patterns, using a specific model about the visitor behavior.

4.1 Preprocessing of web logs

The task is to determine for each visitor, the sequence of web pages visited during a session based on the available web log files. This process is known as **sessionization** [2]. A maximum time duration of 30 minutes per session is considered. The transactions that belong to a specific session can be identified using tables and program filters. We consider only web log registers with non-errors codes whose URL parameters link to web page objects.

4.2 Preprocessing of web site

The web site is represented by a vector space model [1]. 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 of dimension RxQ, $M = (m_{ij})$ where $i = 1, \ldots, R$, $j = 1, \ldots, Q$ and m_{ij} is the weight of the i^{th} word in the j^{th} page. To calculate these weights, we use a variant of the *tfxidf-weighting* [1], defined as follows,

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

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 word is special if it shows special characteristics, e.g. the visitor searches for this word.

Definition 1 (Page Vector). $\mathbf{WP^j} = (wp^j_1, \ldots, wp^j_R) = (m_{1j}, \ldots, m_{Rj})$ *with* $j = 1, \ldots, Q$.

It represents the j^{th} page by the weights of the words contained in it, i.e., by the j^{th} column of M. The angle's cosine is used as a similarity measure between two page vectors,

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

4.3 Modeling the Visitor Browsing Behavior

Our model of the visitor behavior uses three variables: the sequence of visited pages, their contents and the time spent on each page. The model is based on a n-dimensional visitor behavior vector which is defined as follows.

Definition 2 (Visitor Behavior Vector).
$\upsilon = [(p_1, t_1) \ldots (p_n, t_n)]$, *where the pair* (p_i, t_i) *represent the* i^{th} *page visited* (p_i) *and the percentage of time spent on it within a session* (t_i)*, respectively.*

4.4 Comparing Visitor Sessions

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 [9]:

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

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 [8].

4.5 Modeling the visitor's text preferences

A web site keyword is defined as a word or a set of words that makes the web page more attractive to the visitor. The task here is to identify which are the most important words (keywords) in a web site from the visitor's viewpoint. This is done by combining usage information with web page content and analyzing the visitor behavior in the web site.

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. Sorting the visitor behavior vector according to the percentage of time spent on each page, the first ι pages correspond to the ι most important pages.

Definition 3 ($\iota-$Most Important Pages Vector).
$\vartheta_{\iota}(\upsilon) = [(\rho_1,\tau_1),\ldots,(\rho_{\iota},\tau_{\iota})]$, *where the pair* $(\rho_{\iota},\tau_{\iota})$ *represents the* ι^{th} *most important page and the percentage of time spent on it within a session.*

Let α and β be two visitor behavior vectors. A similarity measure between two $\iota-$most important pages vectors is defined as:

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

where the term $\min\{\cdot,\cdot\}$ indicates the visitors' interest in the pages visited, and the term dp is the similarity measure (2).

In (4) the similarity in content of the most important pages is multiplied by the ratio of the percentage of time spent on each page by visitors α and β. This allows us to distinguish between pages with similar contents, but corresponding to different visitors' interests.

4.6 Applying web mining techniques

Similar visitor behaviors are grouped into clusters with common characteristics, such as the navigation sequence or the preferred pages.

Clustering the visitor sessions For clustering the visitor sessions the Self-organizing Feature Map (SOFM) was applied using the similarity measure (3). The SOFM requires vectors of the same dimension. Let H be the dimension of the visitor behavior vector. If a visitor session has less than H elements, the missing components up to H are filled with zeroes. Else if the number of elements is greater than H only the first H components are considered.

Clustering the $\iota-$most important pages vectors A SOFM is used to find groups of similar visitor 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,

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

for $i = 1, \ldots, R$, where kw is an array containing the geometric mean of the weights of each word (1) within the pages contained in a given cluster. Here ζ is the set of pages contained in the cluster. By sorting kw in descendent order the most important words for each cluster can be selected.

5 Real-world application: Bank web site

The above described methodology was applied 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 all the visits done in the period from January to March, 2003. Approximately eight millions of raw web log registers were collected. The site had 217 static web pages with texts written in Spanish, which were numbered from 1 to 217, to facilitate the analysis. In the table 1 the web pages are grouped by their main topic.

The sessionization process was implemented in **perl**. Only 16% of the visitors visited 10 or more pages and 18% less than 4. The average number of visited pages was 6, thus we fixed in $H = 6$ the dimension of the visitor behavior vector. We chose $\iota = 3$ as the maximum number of components of the most important page vector. Approximately 300,000 visitor behavior vectors were identified. The complete web site contained R=4,096 different words. The cluster interpretation was performed by a bank expert. A cluster was accepted only if its content made sense to the business expert.

Table 1. Bank web pages and their contents

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

5.1 Knowledge extracted from visitor browsing

After applying the SOFM to the visitor behavior vectors, four main clusters were found. These clusters are presented in more detail in table 2. The second column of this table contains the centroid (winner neuron) of each cluster, representing the sequence of the pages visited. The third column contains the time spent on each page of the corresponding sequence.

Table 2. Visitor behavior clusters

Cluster	Visited Page Sequences	Time spent in seconds
1	(1,3,8,9,147,190)	(40,67,175,113,184,43)
2	(100,101,126,128,30,58)	(20,69,40,63,107,10)
3	(70,86,150,186,137,97)	(4,61,35,5,65,97)
4	(157,169,180,101,105,1)	(5,80,121,108,30,5)

Based on the clusters found we made offline recommendations for reconfigurating the link structure of the bank web site. Some of these recommendations are: a) **Add links.** The general idea is to improve the accessibility of important pages within each cluster. For instance, in cluster 3 the visitors spend long time on page 150 (35 s), then look a few seconds at the page 186 (5 s) and then move to the page 137, where they stay longer (65 s). Our recommendation was to add a direct link from page 150 to page 137. b) **Eliminate links.** Links that are rarely used can be eliminated. For instance the link from page 150 to page 186 caused "confusion" to many visitors in Cluster 3.

5.2 Knowledge extracted from visitor preferences

After applying the SOFM to the $3-$most important pages vectors, 8 main clusters were found. These clusters are shown in table 3. The second and fourth columns contain the centroid of each cluster, representing the 3-most important pages visited.

Table 3. Clusters of the 3-most important pages

Cluster	Pages Visited	Cluster	Pages Visited
1	(6,8,190)	5	(3,9,147)
2	(100,128,30)	6	(100,126,58)
3	(86,150,97)	7	(70,186,137)
4	(101,105,1)	8	(157,169,180)

Applying (5), we obtained the keywords and their relative importance in each cluster. For instance, for cluster 1 $\zeta = \{6, 8, 190\}$, and $kw[i] = \sqrt[3]{m_{i6} m_{i8} m_{i190}}$, with $i = 1, \ldots, R$. By sorting $kw[i]$ the group of most important words for each cluster were selected. Our confidentiality agreement with the bank, does not allow us to show the specific keywords found per cluster. Some of the keywords found are (translated from Spanish): Credit, House-credit, (Credit) Card, Promotions, Contests, Points.

6 Conclusions

The proposed framework provides a methodology to process web data, store the information extracted and prepare it for the application of web mining techniques, with the aim of discovering meaningful patterns about the visitor browsing behavior and his/her preferences. The methodology was successfully applied to a real-world web site, owned by a commercial bank. In this way the most important pages and keywords were automatically found, and then validated by a business expert. This allowed us to give offline recommendations for changing the structure and contents of the bank web site. Future research is needed to test the proposed framework with other web sites, as well as to measure the effectiveness of the recommendations provided.

References

1. M.W. Berry, S.T. Dumais and G.W.O'Brien, Using linear algebra for intelligent information retrieval, *SIAM Review*, Vol. 37, pages 573-595, 1995.
2. B. Berendt and M. Spiliopoulou, Analysis of navigation behavior in web sites integrating multiple information systems, *The VLDB Journal*, Vol. 9, pages 56-75, 2001.
3. C. Bouras and A. Konidaris, Web Components: A Concept for Improving Personalization and Reducing User Perceived Latency on the World Wide Web, *Proc. Int. Conf. on Internet Computing*, Vol. 2, pages 238–244, June, 2001.
4. P. Brusilovsky, Adaptive Web-based System: Technologies and Examples, *IEEE Web Intelligence Int. Conference, Tutorial*, October,2003.
5. M. Cadoli and F. M. Donini, A Survey on Knowledge Compilation, *AI Communications*, Vol. 10(3-4), pages 137-150, 1997.
6. M. Kilfoil, A. Ghorbani, W. Xing, Z. Lei, J. Lu, J. Zhang and X. Xu, Toward an adaptive web: The state of the art and science, *In Proc. Conf. of Communication Network and Services Research*, pages 108-119, Moncton, NB, Canada, 2003.
7. M. Perkowitz and O. Etzioni, Towards adaptive Web sites: Conceptual framework and case study, *Artificial Intelligence*, Vol. 118(1-2), pages 245-275, 2000.
8. T. A. Runkler and J. Bezdek, Web Mining with Relational Clustering, *International Journal of Approximate Reasoning*, Vol. 32(2-3), pages 217-236, 2003.
9. J. D. Velásquez, H. Yasuda, T. Aoki and R. Weber, A new similarity measure to understand visitor behavior in a web site, *IEICE Trans. on Inf. and Sys.*, E87-D(2), pages 389-396, February, 2004.
10. J.D. Velásquez, H. Yasuda, T. Aoki, R. Weber and E. Vera, Using self organizing feature maps to acquire knowledge about visitor behavior in a web site, *Lecture Notes in Artificial Intelligence*, 2773(1), pages 951-958, September, 2003.
11. J. D. Velásquez, R. Weber, H. Yasuda and T. Aoki, A Methodology to Find Web Site Keywords, *Procs. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service*, pages 285-292, March, Taipei, Taiwan,2004.