A Platform for Extracting and Storing Web Data

L. Víctor Rebolledo and Juan D. Velásquez

vireboll@gmail.com, jvelasqu@dii.uchile.cl

Abstract. Web data or data originated on the Web contain information and knowledge which allows to improve web site efficiency and effectiveness to attract and retain visitors.

However, web data have many irrelevant data inside. Consequently, it is necessary to preprocess them to model and understand the web user browsing behavior inside them. Further, due to frequent changes in the visitor's behavior, as well as in the web site itself, the discovered knowledge may become obsolete in a short period of time.

In this paper, we introduce a platform which extracts, preprocesses and stores web data to enabling the utilization of web mining techniques. In other words, there is an Information Repository (IR) which stores preprocessed web data and it facilitates the patterns extraction. Likewise, there is a Knowledge Base (KB) for storing the discovered patterns which have been validated by a domain expert.

The proposed structure was tested using a real web site to prove the effectiveness of our approach.

Keywords: Platform, Web mining, Knowledge Base.

1 Introduction

Web data let extract information and knowledge that suggests changes to become a web site more effective and efficient [8][11]. This potential is determined by a site's content, its design and structure [6]. Indeed, through web log analysis is possible to understand the visitor's behavior, and further, together with web page content processing, it is possible extract which the visitor's content preferences are [9][10].

Nevertheless, web data have to be cleaned, consolidated and transformed in an ad-hoc structure for the application of web mining [1][8]. Particularly, web logs have many irrelevant data that should be cleaned; web site text content should be processed by removing HTML tags and by taking it to an appropriate structure, for example *Vector Space Model* [5]. As a result, it implies high costs in time and resources.

In addition, the visitor's interests change in the time, as well as in the web site itself. Furthermore, the discovered knowledge may become obsolete in a short period of time [8]. In other words, we periodically have to incur in the cost of preprocess web data to analyze the web site.

The proposed platform allows keeping web data ready to apply web mining. On that point, we can extract knowledge whenever we want and without incur in costs associated to preprocess web data.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 838–845, 2009.

2 The Platform

The proposed platform is formed by different components such as:

- **Web site**: composed for web logs, web page text content and other objects like images, files, etc[1].
- **Data Staging Area or DSA**: It is the area where web data are cleaned and preprocessed. For this paper, we use a relational database as DSA.
- **Extraction Process**: Through it, the data are periodically extracted from web site content and web logs to the DSA
- **Information Repository**: it is a repository, built with Data Warehouse architecture[2], where the preprocessed web data are stored. It facilitates the extraction of feature vectors which are the input for web mining algorithms.
- **Transformation and Load Process**: Corresponds to sessionization process (see subsection 3.1) and transformation of web site text content into a Vector Space Model (see subsection 3.2). The transformation happen in DSA and the data are loaded in Information Repository.

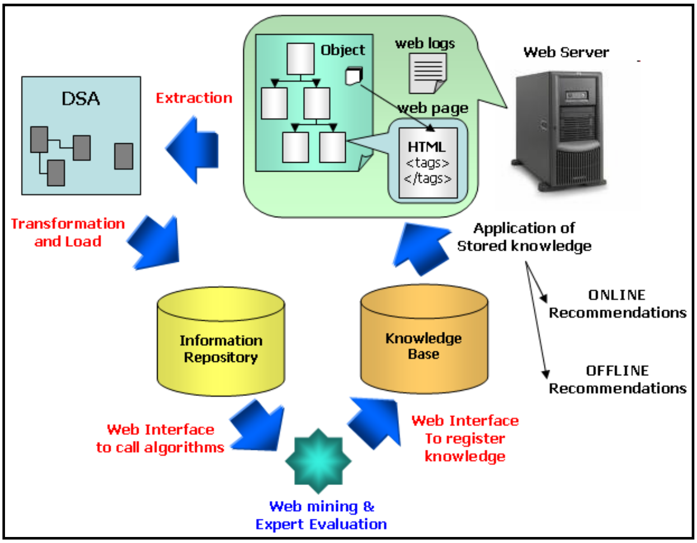


Fig. 1. Platform components

- **Web interface to generate vectors**: Through them, the data miners can create specific vectors for their web mining studies.
- **Web mining algorithms**: Corresponds to the techniques used to extract knowledge from web data (see subsection 3.5)

[1] In order to assure web data provision, we use Plone, a Content Management System, to manage the web site content. Indeed, this tool let add metadata to each web object and keep records about web site's changes.

- **Knowledge Base KB**: It is a repository, as Information Repository, where the discovered knowledge will be stored.
- **Web interface to register knowledge**: The extraction of knowledge is not an automatic process because it requires the domain expert's interpretation and validation [7]. For this reason, this web interface will let data miners manually register the discovered knowledge in KB (see subsection 4.4).

3 Modeling and Preprocessing Web Data

In order to apply web mining techniques, web logs and web pages have to be preprocessed by using specific models to representing them.

3.1 Preprocessing of Web Logs

For each visitor, it is necessary to determine the sequence of pages during a session based on web logs entries. This process is called **sessionization[9]** and it can be performed by using tables and program filters. We consider a maximum time duration of 30 minutes per session and we use only web logs registers with non-errors codes chose URL parameters link to web page objects.

3.2 Preprocessing of Web Site

A good representation to web site text content is the Vector Space Model[5]. Indeed, 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 $i = 1, \dots, R, j = 1, \dots, Q$ and m_{ij} is the weight of the ith word in the jth page. To calculate these weights, we use a variant of the *tfxidf-weighting[5]*, defined as follows:

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

Where f_{ij} is the number of occurrences of the ith word in the jth page, *sw(i)* is a factor to increase the importance of special words and n_i is the number of documents containing the ith word. A word is special if it shows special characteristics, e.g., the visitor searches for this word.

Definition 1 (Page Vector). It is a vector $WP^j = (wp_1^j, \dots, wp_R^j) = (m_{1j}, \dots, m_{Rj})$ with $j = 1, \dots,$ Q, that represent a list of words contained within a web page. It represents the jth page by the weights of the words contained in it, i.e., by the jth column of M. The angle's cosine is used as a similarity measure between two page vectors:

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

3.3 Modeling the User Browsing Behavior

Our visitor behavior model 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 (User Behavior Vector UBV). It is a vector $v = [(p_1, t_1), \ldots, (p_n, \ldots, t_n)]$, where the pair$(p_i, t_i)$ represents the *i*th page visited p_i and the percentage of time spent on it within a session t_i, respectively.

3.3.1 Comparing User Behavior Vectors

Let α and β be two visitor behavior vectors of dimension C^α and C^β, respectively. Let $\Gamma(.)$ 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 between the *k*th page of vector α and the *k*th page of vector β. The term $\tau_k = min\{t_{\alpha,k}/t_{\beta,k}, t_{\beta,k}/t_{\alpha,k}\}$ is an indicator of the visitor's interest in the visited pages. The term *d*G is the similarity between sequences of pages visited by two visitors [9].

3.4 Modeling the User'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 [10]. 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 the web page content and by 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. By sorting the visitor behavior vector according to the percentage of time spent on each page, the first ι pages will correspond to the ι -most important pages.

Definition 3 (ι-most Important Page Vector IPV): It is a vector $\vartheta(v)[(\rho_1, t_1), \ldots, (\rho_\iota, t_\iota)]$, where the pair (ρ_ι, t_ι) represents the ιth most important page and the percentage of time spent on it within a session.

3.4.1 Comparing Important Page Vector

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

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

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

In (4), the similarity 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.

3.5 Applying Web Mining Techniques

Due to most of times the visitors are anonymous, there is no previous idea about visitor behavior, and hence clustering techniques are useful [7][10]. In that sense, we use 2 clustering algorithms to validate the obtained patterns. Moreover, we extract association rules to find correlations between the pages visited into a session.

- **Identifying Association Rules:** By using the classic algorithm Apriori[4], we can validate or reject the patterns obtained with another technique like clustering

- **Clustering UBV:** We apply Self Organizing Feature Maps SOFM and K-means on UBV by using the similarity measure (3). Firstly, we use SOFM which requires vectors of the same dimension. Let H be the dimension of the UBV. If a user 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. Later, we use K-means algorithm by setting the number of clusters as the amount of validated clusters obtained with SOFM.

- **Clustering IPV:** A SOFM y K-means are used 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 according to each cluster is calculated by:

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

for $i = 1, \dots, R$, 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 descending order, the most important words for each cluster can be selected.

4 Real-World Application

The above described methodology was applied to the web site of University of Chile's Web Intelligence Research Group (http://wi.dii.uchile.cl). This site was built by using Plone and it is formed by 42 static web pages in Spanish and English, and by 102 objects as files and pictures. We analyzed 72.481 all the visits done in the period from January to May, 2008. Approximately, 80 thousands of raw log registers were collected.

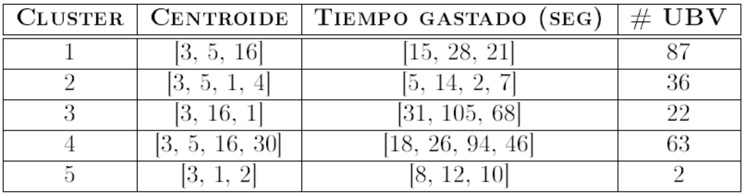
4.1 Knowledge Extracted from Association Rules

Due to we identify a little amount of sessions, we need to adjust the confidence and support levels in order to get interesting association rules. With a confidence level of 16%, we found high correlation between pages: *Bienvenido (Welcome), Estudiantes (Students), Investigación (Research)* and *Investigadores (Researchers).* Indeed, three of them appear together in 16,1% of sessions.

4.2 Knowledge Extracted from Visitor Browsing

After applying SOFM to UBV, five clusters were found. These clusters are presented in **Table 1**. Indeed, 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 in each page and the fourth column, the amount of vectors that belongs to each clusters. The last centroid only represents 2 vectors, furthermore, we only considerate four relevant clusters.

Table 1. Clusters found with SOFM



Cluster	Centroide	Tiempo gastado (seg)	# UBV
1	[3, 5, 16]	[15, 28, 21]	87
2	[3, 5, 1, 4]	[5, 14, 2, 7]	36
3	[3, 16, 1]	[31, 105, 68]	22
4	[3, 5, 16, 30]	[18, 26, 94, 46]	63
5	[3, 1, 2]	[8, 12, 10]	2

In that sense, we set K-means to obtain the following four clusters:

- **Cluster 1**: Visitors which were searching information about the members' publications of WI Group.
- **Cluster 2**: Visitors which were searching academic information about the members of WI Group.
- **Cluster 3**: Visitors which were interested on activities of teaching and seminars imparted by WI Group
- **Cluster 4**: Visitors which were searching information about the studies made by WI Group which are of public interest

4.3 Knowledge Extracted from Visitor Preferences

After applying the SOFM to the 3-most important pages vectors, five clusters were found, however only four clusters were considered. These clusters can be seen in **Fig. 2**. Applying (5), we obtained the keywords and their relative importance in each cluster. For example, the cluster 1 = {3, 16, 1}, and $kw[i] = \sqrt[3]{m_{i3}m_{i16}m_{i1}}$ with $i = 1, \dots, R$. By sorting kw[i], the group of most important words for each cluster were selected. Some of the keywords found were: *web, development, technology, data, professor, graduated, information, base*, etc.

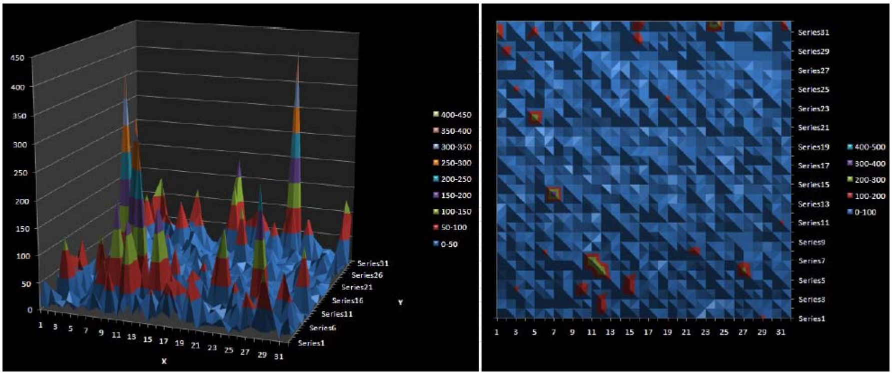


Fig. 2. Identifying clusters by using SOFM

At the same way and by using the above IPV, we apply K-means to get four clusters (we set it a priori). Some of the keywords found were: *information, investigation, title, graduated, professor, system, courses, knowledge, intelligent*, etc.

4.4 Loading the Knowledge Base

Through a web interface, we can introduce the patterns interpreted with the expert help. The Knowledge base will be formed by a Fact Table with the studies' results and Dimension tables with attributes which characterize the discovered patterns. In that sense, the KB stores the web mining technique used (WMT), the date when the technique was applied, the found pattern and its interpretation, etc. For example:

- **Time:** (2008, July, 23, 05:30 hrs.)
- **Browsing_Behavior:** The clusters centroids discovered by the mining process.
- **WMT:** SOFM with thoroidal neighbor and 32*32 neurons
- **Text_Preference:** The keywords context of the discovered clusters.

This information can be used by a human for changes in web site structure and content and by a system which makes navigation recommendations to the user when they present a behavior pattern that coincide with some stored in the repository [8].

5 Conclusions

The proposed platform preprocesses web data periodically in order to always have available vectors to enter to web mining algorithms. Moreover it allows storing discovered patterns in a repository, which can be used by a computational system through a set of rules that make online navigation recommendations and by humans for offline changes in the web site content and structure. Consequently, the web site can be modified in order to make it more efficient and effective to attract and retain users. In future works, other web mining techniques will be applied in order to provide new patterns and rules for the KB.

Acknowledgements

The authors thank to the Millennium Scientific Nucleus on Complex Engineering Systems (Chile), which has partially funded this work.

References

[1] Cooley, R.W.: Web usage mining: discovery and application of interesting patterns from web data, Dissertation for degree of Doctor of Philosophy. University of Minnesota, Faculty of the Graduate School, Minnesota, USA (2000)

[2] Kimball, R., Merx, R.: The Data Webhouse Toolkit. Wiley Computer Publisher, Chichester (2000)

[3] Kosala, R., Blockell, H.: Web mining research: a survey. SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining 2(1), 1–15 (2000)

[4] Larose, D.T.: Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Chichester (2005)

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

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

[7] Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press, London (1999)

[8] Velasquez, J.D., Palade, V.: Adaptive Web site: A Knowledge Extraction from Web Data Approach. IOS Press, Amsterdam (2008)

[9] Velasquez, J.D., Yasuda, H., Aoki, T., Weber, R.: 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 (2004)

[10] Velasquez, J.D., Yasuda, H., Aoki, T., Weber, R., Vera, E.: Using self organizing feature maps to acquire knowledge about visitor behavior in a web site. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS (LNAI), vol. 2773, pp. 951–958. Springer, Heidelberg (2003)

[11] Yao, Y.Y.: Web intelligence: New frontiers of exploration. In: Proceedings of the 2005 International Conference on Active Media Technology (AMT 2005), Takamatsu, Kagawa, Japan, May 19-21 2005, pp. 3–8 (2005)