# Improvement of a Methodology for Website Keyobject Identification through the Application of Eye-Tracking Technologies

Juan D. Velásquez Silva
*Web Intelligence Consortium Chile Research Centre*
*Department of Industrial Engineering*
*University of Chile*
*Santiago, Chile*
*jvelasqu@dii.uchile.cl*

***Abstract*—This paper introduces the utilization of data originated in the web user ocular movement to improve the methodology for identifying Website Keyobjects that was designed by Velásquez and Dujovne through the use of eye tracking tools. Given a website, this methodology takes as input the request register (*Web log*) of the website, the pages that compose it and the interest of users in the web objects of each page, which is quantified using a survey. Subsequently, the data is transformed and preprocessed before finally applying Web mining algorithms that allow the extraction of the Website Keyobjects. In this paper, a novel application of the eye tracking technology is proposed, in order to dispense with the survey, that is to say, using a more precise tool to achieve an improvement in the classification of the Website Keyobjects. It was concluded that eye tracking technology is useful and accurate when it comes to knowing what a user looks at and therefore, what attracts their attention the most. Finally, it was established that there is an improvement of between 15% and 16% when using the information generated by the eye tracker.**

***Keywords*-Web Mining, Eye Tracking, Website Keyobjects.**

## I. Introduction

In recent years, there has been a new research area called Web Mining, which studies different ways of extracting information from data generated on the Web. With this knowledge, it is possible to develop techniques and algorithms to attract and retain users on a website. Specifically, this area applies data mining techniques to data originated on the Web with the aim of obtaining valuable information to continuously improve a website in terms of design and content, among other aspects.

An interesting result that has been achieved in this field are the Website Keywords [1], defined as a word or set of words used by users in their information search process, and which characterize the content of a website or page. After finding them, the sites can be redesigned according to the needs and requirements of its users and, thus, be in the vanguard. Although identifying the Website Keywords of a site helps to know the preferences of users, the methodology that discovers them only focuses on the textual content, leaving out the analysis of the multimedia content of websites [1]. For this reason, Velásquez and Dujovne [2] extended this methodology and integrated both the textual content and the multimedia content in their analysis.

Velásquez and Dujovne defined a Web Object as "any structured group of words or a multimedia resource that is present on a web page that has metadata which describe its content". Also, they characterized a Website Keyobject as a Web Object that captures the attention of the users and that characterizes the contents of a web site [3]. From the above definitions it is possible to deduce that every website consists of a set of Web Objects and that the set of Website Keyobjects it has is a subset of the former.

Velásquez and Dujovne [2] designed a methodology that allowed the identification of the Website Keyobjects of a site. These, like the Website Keywords, give guidelines for the websites to be redesigned according to user requirements. This methodology requires knowledge of the time spent by users on the Web Objects, i.e. how long a user spends looking at each Web Object. To determine the permanence time they propose two steps:

By mixing the two processes, Velásquez and Dujovne [2] estimated the time spent by users on the Web Object. However, thanks to the eye tracking technology it is possible to dispense with this survey. Eye tracking technology allows knowing what a person looks at as a function of time [4]. By applying this technology to users browsing a site it is possible to measure the time spent on each Web Object. In this sense, if it was possible to quantify the permanence time of a control group of users on Web Objects, an improvement to the methodology developed could be made, i.e., determining the Website Keyobjects more accurately.

## II. Related Work

In this section a conceptual framework about the traditional web mining process, which use data originated in the Web, and the possibility to use other kind of data, for instance generated by application of eye tracking tools, will be introduced.

*1) Data Originated on the Web:* The data that are originated on the Web are classified into three types: content, structure and usability.

978-0-7695-4880-7/12 $26.00 © 2012 IEEE
DOI 10.1109/WI-IAT.2012.182

*2) Web Mining:* Web mining is the application of data mining to data originated on the Web [5], [6]. It is conceived as a product of the intersection of several research areas such as databases, information retrieval, artificial intelligence, especially the sub-areas of machine learning and language processing. Research in this field is experiencing significant growth because of the large amount of data available for analysis [7]. This is no small task, considering that the Web is a large collection of heterogeneous, declassified, distributed, time-varying, semi-structured high-dimensional data [8].

*3) Eye Tracking:* Eye tracking is a technique by which an individual's eye movements are measured. Thus, a researcher can know what a person is looking at in each moment and the sequence in which their eyes move from one place to another [9].

*4) Eye Movements:* When looking at any scene, the eyes of a person move between points that capture their attention, which are able to recreate a brain image of the scene [10]. While there are models of complex eye movements consisting of five steps [11], the typical model (which is enough for studies of eye tracking) consists of two concepts, fixation and saccades. Fixation is defined as the moment in which the eyes are fixed on an object and it is possible to appreciate it in detail, while the saccades correspond to rapid eye movements between two fixations.

*5) Visual Attention:* Visual attention is a phenomenon that has been studied for about a hundred years and still cannot be understood. Early studies were limited by technology, and corresponded only to observation and introspection. At present, this field is studied by different disciplines such as psychophysics, cognitive neuroscience and computer science, to name a few [**?**].

*6) The Eye-mind Hypothesis:* Considering the theories described in the previous section, Duchowski [11] posited the following model:

1) Given a stimulus such as an image, the scene is seen largely in parallel through peripheral vision and therefore, at low resolution. At this stage, the interesting features of the image can "appear".
2) In this moment, attention is disconnected from the foveal vision (high resolution), but the eyes quickly move toward the first region of attraction.
3) Once the eyes are positioned, the fovea is aligned to the region of interest and attention is linked to perception, i.e. the user's attention has been captured and, therefore, it is possible to observe in high resolution.

Nielsen and Pernice [10] proposed a simpler model, which in its essence is the same. Their hypothesis states that "People are usually thinking about what they are looking at. Although they do not always understand what they see or are not fully focused on it; if they are observing something, then they are paying attention, especially when they are concentrated on a particular task".

## III. Methodology to Find Website Keyobjects

In [2] was defined a Web Object as "*a structured group of words or multimedia content that is present on a Web page, and which has metadata describing its content.*" In the previous definition, the metadata are fundamental as they are the basis of the information to construct the vector representing the content of the page. In addition, two multimedia files can be compared by their metadata, a problem which can be dealt with more easily than comparing files directly, because it only compares text. Along with this, they defined the Web Site Key Object as "*one or a group of Web Objects that attract user attention and which characterize the contents of a web page or site*". They provide knowledge about the content and format of most interest to users of a website, therefore finding them may be useful to improve the site both in presentation and content.

The sub-processes that allow for the identification of the Website Keyobject are [3]:

1) **Sesionization**: the objective of this stage is to complete the sequence of pages visited by the different web users of the site. In addition, this sequence must have associated with it the residence times of each web user per page visited during his/her session.
2) **Addition of Metadata**: the first step of this stage is to identify the objects that make up the pages of the site. Once identified, the concepts that describe each of the objects must be defined. Then, this information must be stored in a database. The gathering of these data should be done together with the web master to ensure that the concepts accurately reflect the content of objects.
3) **Web Object Similarity**: To compare two web objects a similarity measure $do : |\mathbf{c}| \times |\mathbf{c}| \rightarrow [0, 1]$ was defined by using a given distance measure (e.g. an edit distance).
   Given two objects $x_i$ and $x_j$ such that $|x_i| = N$ and $|x_j| = M$, where $N, M \geq 0 \wedge N \leq M$, and given $x_i \rightarrow c_k^1, k \in \{1, \ldots, M\}$ as the $k^{th}$ concept of the object $x_i$. We need to find a representation for which all WebObjects can be compared. A compare function that could be used, but not limited to, is based on the edit distance between the pairwise alignment between WebConcepts of objects $x_i$ and $x_j, \forall x_i, x_j \in \mathbf{x}, i \neq j$. Once all the concepts are paired, the objects concepts are ordered in such a way that every concept is in the same relative position in relation to each object. Then a string that consists of a symbol representing each concept is created to represent each object. This string has the structure shown in the expression 1,

$$x = WebConcept_1, \ldots, WebConcept_N \Rightarrow x = c_1, \ldots, c_N \quad (1)$$

where $c_k \in \{x_i \to c\}$ represents a set of all concepts in object $x_i$.

4) **Permanence time on objects**:after defining the objects, a survey is carrying out a control group of users so that each respondent distributed a total of 10 points of interest among all objects on a page. With these data, they estimated the percentage spent by each user on the objects on each page.
5) **Vector of user behavior**: finally, for each session identified, the $n$ objects that captured more attention from the user were selected, thus defining the Important Object Vector (IOV) according to equation 2.

$$v = [(o_1, t_1)...(o_n, t_n)] \qquad (2)$$

6) **Clustering Algorithms**: once all the cleaning and transformation of data has been done, the following step is the processing of clustering algorithms on user sessions, represented by Important Object Vector. To run these algorithms, it is of crucial importance to define a measure of distance, or similarity, between these vectors.
7) **Similarity measures for sessions**: The similarity defined for comparing two IOV $\alpha, \beta$ is calculated by using the expression 3.

$$st(\alpha, \beta) = \frac{1}{i} * (\sum_{k=1}^{i} min(\frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha}) * do(o_k^\alpha, o_k^\beta)) \qquad (3)$$

where $do : |\mathbf{c}| \times |\mathbf{c}| \to [0, 1]$ is the similarity between two WebObjects (e.g. an edit distance) and $\tau_k^\alpha, \tau_k^\beta$ the time spent for the user seeing a WebObject.

## IV. Experiment and Results

On a site, the methodology designed by Velásquez and Dujovne [3] hereinafter the original methodology will be executed, and also the modified methodology will be executed, to which the permanence times calculated with an eye tracker will be incorporated.

### A. Experiment

To capture the interest of users in the different objects of the pages, two procedures will be used, using an eye tracker and the application of a survey. In this sense, experiments were conducted on the site http://www.mbauchile.cl, belonging to the Master in Business Administration and Management program of the Department of Industrial Engineering of the University of Chile. This site consists of 124 pages and 163 different objects that appear on the site 2,047 times. Thus, the average number of objects per page is 12.55. With regard to visits, each month 4,158 different people access the site, 6,111 sessions are recorded and 26,589 pages are seen.

In the other hand, before selecting the control users who would participate in the experiment, the business expert was asked for information on the target of the site under study with the objective of choosing a representative sample of individuals who visit the site. Considering this information, 33 people were selected. Of these, 16 were male and 17 female, and their average age was 24.3 years. Regarding the knowledge and use of the Web, 15 of them defined themselves as expert Web users, 12 considered that their knowledge is average and only 6 of them defined themselves as basic users.

Finally, for the development of the experiments eye tracking software and hardware tools were used. Regarding the hardware, the Tobii T120 Eye Tracker[1] was used, which consists of a 17-inch monitor to which two infrared emitters and a light sensor are incorporated. This hardware has a time resolution of 120 Hz and has a margin of error of 0.5. With respect to software, the Tobii Studio Enterprise Edition was used, a solution which easily allows mapping what is displayed on the monitor with the same location that users observe.

*1) Capturing of Data:* To discover the composition of the site http://www.mbauchile.cl,, a crawler that generated the list of pages that make up the site was implemented. This crawler was developed using the Beautiful Soup Python library. Once the list of pages to be analyzed was captured, the following procedure was identifying the objects of the site. To separate each page into the objects that compose them, two criteria were considered, differences in the content (concepts) and the spatial separation between objects of each page. In order to know the coordinates of the objects within the pages, the Python Imaging Library (PIL) was used, which allows working with images on the Python interpreter. A script was used, which from the images of the pages of the site generated the coordinates of the various objects belonging to each page. After defining the objects, the concepts that describe their contents were generated. The procedures were similar to what Velásquez and Dujovne [3] did, that is to say, each concept was created manually.

Regarding the visits to the site, it was possible to recover the web log from the server that hosts the site together with the system administrator. It was not possible to recover all of the requests as the historic records were not stored. However, the requests corresponding to the month of August 2011 were obtained. During this month, 3,031 different people visited the site, in 5,480 sessions. In total 28,832 pages were viewed, there were 156,259 requests and a traffic of 3.20 GB was reached. In order to measure the interest of users in the web objects, their permanence time on each was measured and estimated. This was done in two ways, by using an eye tracker and by applying the survey used by Velásquez and Dujovne [3].

Fifteen of the control users were posed the following situation: "*You have an intention to apply for an MBA program,*

[1]http://www.tobii.com/en/eye-tracking-research/global/

*but have not yet made a final decision, therefore your first step will be to get informed. In the search for information you have arrived at the website http://www.mbauchile.cl, which provides relevant data about the MBA program offered by the University of Chile. Starting from the home site, navigate freely until you can make a decision or decide to take a new step*". In this way, it was attempted to emulate the typical navigation of users at home. The remaining 18 individuals did not navigate freely, but were instructed to look at the pages shown without following any links. Each of them was presented semi-random pages of the site. Users could move to the next page when deemed convenient, but if they spent more than a minute on a page, they were automatically redirected to the following page. The number of pages displayed to users was not higher than 30.

On the other hand, after facing the eye tracker, no matter how their eye movements were captured, subjects were asked to answer a survey in which they had to indicate the objects that most captured their attention. To measure the interest of the users, they had to distribute 10 points, as they chose, over the objects on each page, considering that the more points an object had, the more interest he or she had.

*2) Selection, Data Cleaning and Transformation:* Once the 163 objects on the site were identified, the business expert validated this separation, but also brought together, broke up and removed some pre-selected objects. Following this validation, as some objects changed, it was necessary to re-calculate their positions on the pages. To do this, the script that generated the location (in pixels) of objects on the page was modified and re-executed. Then the coordinates of the pixels of the objects were normalized according to the size of the stimulus, so that the coordinates of the objects came from being stored as integers between 0 and the length of the object, to a double-precision value between 0 and 1.

In the other hand, the results of the survey corresponded to files where the page, the object, and points of interest assigned by the user were registered. These data were transformed so that the 10 points assigned by each user, corresponded to 100% of user interest in the stimulus. For example, if a person assigned 5 points to object X on page Y, it was assumed that the interest of the user in object X was 50% on page Y. Similarly to the previous case, 0 was assigned to the interest of the user in objects that did not get points. Afterwards these results were averaged and stored. Starting from web requests recovered, the next step was to run the process of sessionization. The result was a set of sessions, in which each of these contained a list of pages associated with the amount of time spent on these. Then, for each register, the page was replaced by the objects that made it up, and the time spent on the page was weighed by the percentage of permanence on the object.

*3) Application of Data Mining Algorithms:* The methodology to find Website Keyobjects designed by Velásquez and Dujovne [3] grouped the vectors of user behavior using three techniques: Self-Organizing Feature Maps, K-means and Association Rules. The results of these techniques were sets of vectors where the elements were similar to each other, but different when taking elements from different sets. The criteria used to determine if an object was a Website Keyobject was to select the objects that appeared more times in the clusters shown by the three algorithms.

### *B. Results*

After several tests regarding the number of neurons of the SOFM network, it was found that one constituted by 12X12 neurons gave the best results. This network gave 8 clusters for each experiment. The output of this algorithm was modified to deliver lists of session identifiers belonging to each cluster, in order to count the number of occurrences of objects in different clusters. Regarding the K-means technique, the number obtained by the previous technique, i.e. 8, was used as K. On the other hand, for the implementation of Association Rules, the Weka platform was used. Only objects in the IOVs that were processed according to the platform were considered. The Apriori algorithm was used and it was asked to generate only 30 rules with a maximum confidence of 0.9. Both the algorithm and the platform to be used were the ones used by Dujovne [3] when he implemented this methodology.

*1) Accuracy, Recall and F-Measure:* To compare the results of both experiments, three ranges were selected, taking 10, 20 and 30 Website Keyobjects. Table I shows the accuracy achieved by experiments in the three ranges selected. Table II shows recall, and Table III shows the F-measure indicator.

| Experiment | 10 | 20 | 30 |
|---|---|---|---|
| First | 70% | 72% | 68% |
| Second | 70% | 93% | 96% |

Table I
*Precision*

| Experiment | 10 | 20 | 30 |
|---|---|---|---|
| First | 21% | 43% | 58% |
| Second | 21% | 61% | 77% |

Table II
*Recall*

| Experiment | 10 | 20 | 30 |
|---|---|---|---|
| First | 32.3% | 53.8% | 61.6% |
| Second | 32.3% | 73.7% | 85.5% |

Table III
*F-measure*

An increase (15% to 16%) in accuracy can be noticed by comparing both experiments, which validates the fact that this technology is useful for measuring the interest of users. On the other hand, one can note that when choosing 30 Website Keyobjects the accuracy obtained decreases. This is because for this number a selection of objects that are not relevant begins. At this point, if we consider as threshold n = 24 it is concluded that the accuracy of the second experiment is 93%. Regarding recall, when taking 10 or 20 Website Keyobjects, this indicator is considerably smaller. This is expected since some Keyobjects are no longer being considered. However, in the second experiment, better results for all ranges observed are obtained. Finally, F-measure is a measure that combines both accuracy and recall indicators, therefore it was expected to achieve better results in the second experiment.

## V. Conclusion and Future Work

In the present study it was proved that using an eye tracker to measure the amount of time users spend looking at the different objects on a web page instead of conducting a survey to estimate these values improve the accuracy of finding the Website Keyobjects of a site. To achieve this result, a comprehensive study on data originating from the web, on the mathematical models used to describe the behavior of web users and on existing eye-tracking tools was carried out.

One of the limitations that eye tracking has is that it only determines what a person looks at. This is insufficient when what is sought is to qualify what a person looks at. In other words, these tools cannot determine if what is observed is liked or disliked. Therefore, the results provided by the eye tracker must be considered as the measure (always positive) of the interest of a person.

Finally, it could be possible to analyze the way in which this methodology operates, to investigate whether changing the structure of it can achieve better results than using 3 different data mining algorithms. Thus, creating a new way of ranking the objects may help in this goal.

## Acknowledgement

This work was supported partially by the FONDEF project D10I- 1198 entitled, WHALE: Web Hypermedia Analysis Latent Environment and the Millennium Institute on Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16).

## References

[1] J. D. Velásquez, “Web site keywords: A methodology for improving gradually the web site text content,” *Intelligent Data Analysis*, vol. 15, no. 4, pp. 327–348, 2011.

[2] L. E. Dujovne and J. D. Velásquez, “Design and implementation of a methodology for identifying website keyobjects,” in *Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems: Part I*, ser. KES ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 301–308.

[3] J. D. Velásquez, L. E. Dujovne, and G. L’Huillier, “Extracting significant website key objects: A semantic web mining approach,” *Eng. Appl. Artif. Intell.*, vol. 24, no. 8, pp. 1532–1541, 2011.

[4] N. F. Ali-Hasan, E. J. Harrington, and J. B. Richman, “Best practices for eye tracking of television and video user experiences,” in *Proceeding of the 1st international conference on Designing interactive user experiences for TV and video*, ser. UXTV ’08. New York, NY, USA: ACM, 2008, pp. 5–8.

[5] G. Chang, M. Healey, J. McHugh, and J. Wang, *Mining the World Wide Web*. Kluwer, 2001.

[6] M. Spiliopoulou, “Data mining for the web,” *Principles of Data Mining and Knowledge Discovery*, pp. 588–589, 1999.

[7] J. D. Velásquez and L. C. Jain, *Advanced Techniques in Web Intelligence - Part 1*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[8] S. K. Pal, V. Talwar, and P. Mitra, “Web mining in soft computing framework: Relevance, state of the art and future directions,” *Neural Networks, IEEE Transactions on*, vol. 13, no. 5, pp. 1163–1177, 2002.

[9] A. Poole and L. Ball, “Eye tracking in human-computer interaction and usability research: current status and future prospects,” *Encyclopedia of human computer interaction*, pp. 211–219, 2005.

[10] J. Nielsen and K. Pernice, *Eyetracking web usability*. New Riders Pub, 2009.

[11] A. Duchowski, *Eye tracking methodology: Theory and practice*. Springer Verlag, 2003.