International Journal of Engineering Research and Technology. ISSN 0974-3154 Volume 10, Number 1 (2017)

Identification of Website Key Objects by Combining Eye-tracking Technologies with Web Usage Mining

Mr. Rajesh D. Gade
PG Scholar, Department of Computer Science And Engineering,
D. Y. Patil College of Engineering and Technology,
Kolhapur,Maharashtra,India.

Prof. Suresh K. Shirgave
Professor, Department of Department of Information Technology,
DKTE Society's Textile & Engineering Institute,
Ichalkaranji, Maharashtra, India.

Abstract

Web mining, as a research area studied various ways of extracting information from data generated on the web.This paper presents the work being carried out toidentify the website key objects. It combines the knowledge extracted by analyzing web user's perception, using an eye-tracking tool with the knowledge of web user's behavior, extracted by applying web mining algorithms to the data originated on the website. The data originated by capturing the web user's ocular movements on a webpage, by using an eye-tracking tool is utilized here.Eye-tracking technology allows knowing what a person looks at, as a function of time. A researcher can know what a person is looking at in each moment and the sequence in which his eyes move from one place to another. Eye-tracking uses eye-tracking tools, a precise quantitative measure for tracking the user's eye position on the computer's screen which can be used as a measure of the user's interest in a webpage.

The work identifies the most important web objects from the user's point of view, called as Website Key Objects.Combining this data with the sequence of webpage visits registered in the weblog, significant insights about his/her behavior and interest within a website can be extracted. The webmasters can further utilize this information to redesign the website according to the needs and requirements of the web users.

Keywords: Web Mining, Session Filtering, Eye Tracking, Edit Distance, Clustering, User Behaviour, Permanence Time

Introduction

Webmasters are always looking for new ways to enhance the content of their site in order to attract and retain users. If they can gain knowledge of their user preferences, they can offer the content, users are looking for.As a part of web mining, various ways are indentified for 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 web site.The work carried out implements a novel approach for identifying the most important web objects from the web user's point of view, called as Website Key Objects. Web user's ocular movements on a webpage are captured by using an eye-tracking tool. The data originated is further combined with the sequence of webpage visits registered in the web log, to obtain significant insights about his/her behavior and interest within a website. The web data is transformed and preprocessed before applying web mining algorithms that allow the extraction of the website key objects.

As a part of web mining, identifying the website keywords helps to know the preferences of user's, but the methodology that discovers them only focuses on the textual content, leaving out the analysis of the multimedia content of websites [1]. Vel´asquez designed a methodology that allowed the identification of the website key objects which took as input the time spent by user's on the web objects. The permanence time was determined by two steps: Sesionization and Application of a survey. [2]

Application of a survey refers to the implementation of a survey on a control group, wherein web objects are sorted according to their importance within each page of the Web site. Eye-tracking technology allows knowing what a person looks at, as a function of time. A researcher can know what a person is looking at in each moment and the sequence in which his eyes move from one place to another. Eye movements of a web user can be classified into two types: 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 [3].

Understanding web user's perception is important for creating a better website design, which will finally attract and retain the users. During the web user's navigation, the websites structure is perceived as a cognitive experienced state, which is determined by "high levels of skill and control, high levels of challenge and arousal, focused attention". There needs to be qualitative and quantitative measures to evaluate the user's browsing experience. [2]

Various evaluation methods are implemented like Questionnaires, Observation, Interview, Focused group, Think-Aloud, Eye Tracking [1]. The first five methods need to apply some kind of survey for getting the user's point of view. Eye Tracking uses eye-tracking tools, a more precise quantitative measure used for tracking the user's eye position on the computer's screen which can be used as a measure of the user's interest in a Web page.

Methodology

The system that implements this novel application of eye-tracking technology, is designed in following phases:

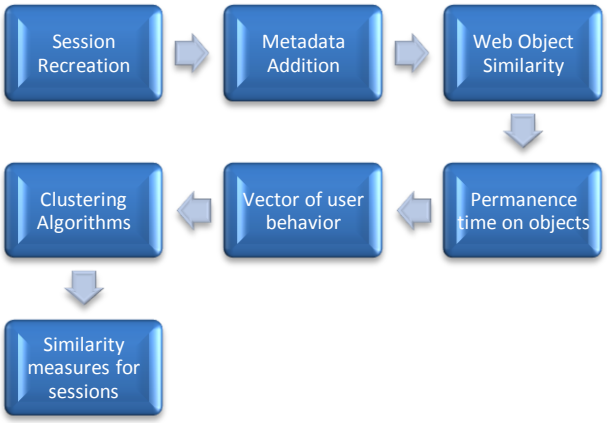


Figure 1: Thephases of implementation.

SessionRecreation:

This phase recreates the sequence of pages visited by the different web users of the site. User access logs obtained from the web server that hosts the site together with the system administrator is utilized for this purpose. A normal web log contains pages visited by web users, a host that requested the page and a timestamp for the request. By reconstructing the user's sessions, it is possible to determine the time a user spends on a given page.

Metadata Addition:

In this phase the web objects that make up the pages of the site are identified. The implementation of web objects can be made in several ways because it relies heavily on the ontology used to describe them. Here an XML document is created that describes every object in a particular page. Each object will be characterized by an identifier, its format and a list of concepts that describe its content. Metadata is used to define web object within a webpage. In this sense, a set of N objects will be defined by X={x_1,,,,x_N }, and each object is defined by a set of M concepts, C={c_1,,,,c_M }. Figure 2 shows the graphical representation of hierarchy between website, web pages, web objects and web contents [2].

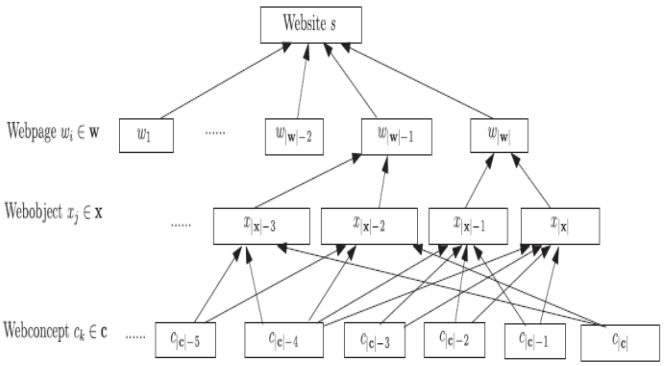


Figure 2: Directed Acyclic Graph as the hierarchy between Websites, Webpage, WebObjects, and Web Concepts.

Web Object Similarity:

Each web objects is represented by a group of concepts that defines a given object's content. To compare two web objects "Levenshtein Distance (edit distance)"$do:|c|\times|c|\rightarrow[0,1]$ is used [6]. A compare function is based on the edit distance between the pair wise alignment between web concepts of objects $x_i,x_j, \forall X_i$.

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 $X_i \rightarrow c\_k \wedge 1 \quad k \in \{1,\ldots,M\}$ as the K[th]concept of the object x_i . A compare function is used, based on the edit distance between the pair wise alignment between web concepts of objects x_i and $x_j, \forall x_i$, x_j∈ x, i≠j represented by Algorithm 1[2].

Require: $x_i,x_j \in \mathbf{x},\tau$
Ensure: Aligned objects $\{x_i,x_j\}$
1: $seq(x_i,x_j) \leftarrow 0$
2: **for** $c_k \in \{x_i \rightarrow c\}$ **do**
3: **for** $c_l \in \{x_j \rightarrow c\}$ **do**
4: **if** c_k.Equals(c_l) **then**
5: $seq(x_i,x_j) \leftarrow seq(x_i,x_j)+1$
6: **else if** c_k. Synonym(c_l) **then**
7: $seq(x_i,x_j) \leftarrow seq(x_i,x_j)+0.5$
8: **end if**
9: **if** $seq(x_i,x_j) > \tau$ **then**
10: $align(x_i,x_j,k,l) \leftarrow$ Pair concept c_k with c_l for both x_i and x_j vectors
11: **end if**
12: **end for**
13: **end for**

Algorithm 1: Pair wise Concept Alignment between Two Web Objects.

Permanence time on objects:

In this phase, the time, a web user spends seeing a particular object on a webpage can be estimated. An eye-tracking tool should be used to follow what the user is seeing in a web page. Because each web object is bounded by a group of pixels, the time, a web user was visualizing a part of the page, can be known.

Vector of user's behavior:

For each session identified, the n objects that captured more attention from the user's will be selected, thus defining the Important Object Vector (IOV) according to the following equation: $v=[(o_1,t_1)\ldots(o_n,t_n)]$, where o_i is the list of objects present on the i^{th} page and t_i is the time spent by the user during each session seeing the object o_i.

Clustering Algorithms::

Once all the cleaning and transformation of data has been done, in this phase, the user's sessions represented by Important Object Vector (IOV) using clustering algorithms will be processed. Frequency of objects in the centroid of every cluster will be used to determine the web key objects. For this a measure of distance, or similarity, between the IOV will be used, as mentioned further.

Similarity measures for sessions:

The similarity defined for comparing two IOV a,b is calculated by using the following equation:

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

where$do\left(o_k^{\alpha}, o_k^{\beta}\right)$ is the similarity between two webobjects (e.g.,an edit distance)and τ_k^{α}k, $\tau_k^{\beta}$the time spent for the user's seeing a webobject

Implementation

Identification and selection of a proper website was required as the first step to implement "Session Recreation"couple of websites was identified from domains like Education, Tourism, HealthCare, etc and were analyzed for further selection.

Table1. Websites explored to check feasibility for eye-tracking experimentation.

Task Name	Category	Title	URL	Total Files	Web Pages	Media Objects
Kolhapurworld	Public info	Kolhapur World the official website of Kolhapur City...Kolhapur at its best	http://www.kolhapurworld.com/	1723	248	1370
Asteraadhar	Healthcare	Aster Aadhar Hospital Kolhapur	http://www.asteraadhar.com/	666	200	443
Kitcoek	Education	KIT College of Engineering Kolhapur	http://www.kitcoek.org/new_index.htm	1	0	0
moodle_intele	Education	Course management system for Department of Informaiton Technology	http://moodle.intele.co.in/	1	0	0
Mahalaxmikolhapur	Tourisim	Mahalaxmi Temple Kolhapur	http://www.mahalaxmikolhapur.com/	671	96	562
Mbauchile	Education	MBA	http://www.mbauchile.cl/	1240	124	163
Dktes	Education	DKTE Society's Textile&Engineering Institute - Ichalkaranji	http://dktes.com/	674	157	316

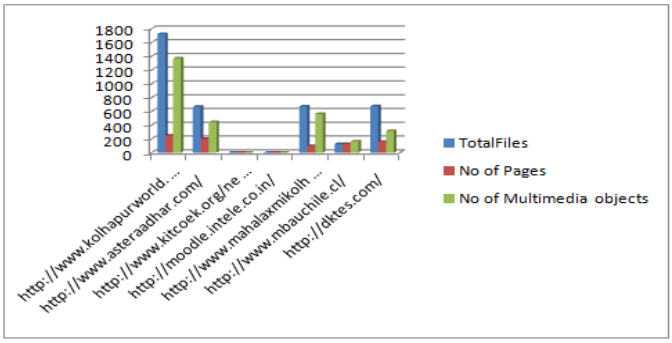


Figure 3. Comparative analysis of different website

Apache HTTP server 2.2.25 is installed and configured on a local system. The website is hosted on this server and used to generate web usage logs required for recreating the user sessions. The web usage logs were cleaned to identify the pages visited by web users,different multimedia contents, the host that requested the web pages (ip address),the timestamp for the request (includes hours, minutes, seconds).Microsoft Access was used to store the web usage log data. As a prototype, A total of 552 log entries, were stored in the database and were able to recreate the user sessions and identify the time spent by user on a single web page.

The test website contained a total of 248 web pages and a total of 1370 media objects of various formats and was characterized as below:

•All pages address different information, and if two pages share similar information it is presented with a different focus.

•The users are interested in a certain set of pages and not interested in the remainder.

•The Website is maintained by a website administrator who can choose if a page stays in the site based primarily on its success to attract users attention.

Figure 4: The test website composed of different data formats

An xml document representing the site map of the website is prepared.

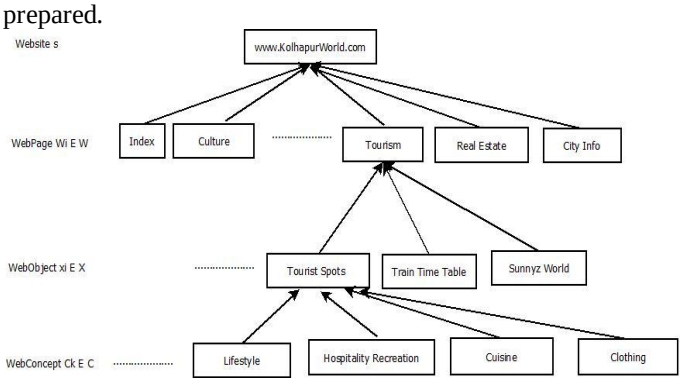


Figure 5: Graphical representation of the proposed xml Document

Each webpage included as a part of site are enriched with meta data using regular HTML meta tags. Sample meta data is as below.

<meta name="description" content="Kolhapur is no more away from you. Wherever you go you will all the way get complete reference, from latest news every week to indepeth view of city's development,business reference directories, vastu, chat, email, tourist info , Career Help , History and just about everything about kolhapursanglipune region. Kolhapur Municipal Corporation information. Visit us and find the geniune info of city. ">

<metaname="keywords"content="Kolhapur,ekolhapur,Kolhapur online business,kolhapurdirectory,Sugarmills,metal companies, gold,silver,chat,Kolhapur pictures, Kolhapur Tourism, Kolhapur Shopping, email, West Maharashtra, PremanshuManghirmalani, SunnyZ World Of Computers,Rajarampuri,Shopping complexes, business, computers, maharashtra, Kolhapur Municipal Corporation, KMC , MahanagarPalika , Corporation office" >

Figure 5 represents the content model view of the schema document used to validate the xml file representing the interrelationship between the web site ,web pages and the web concepts associated with every webpage.

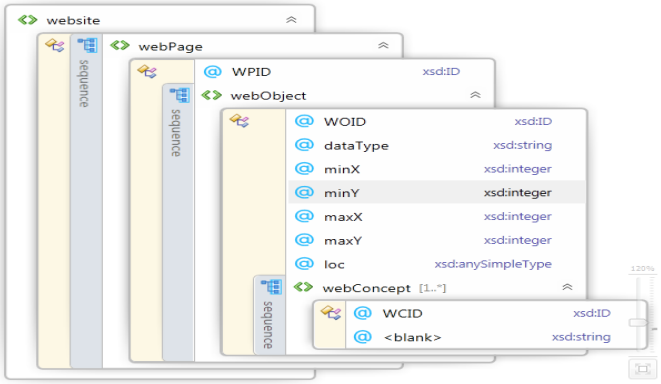


Figure 6. Content model view of the XML schema for proposed XML Document

After defining the objects, it is necessary to estimate how long a web user is seeing a particular object on a web page. The methodology is to apply an eye-tracking tool. An eye-tracking tool Camera Mouse 2013 version 2.1 is used to follow what the user is seeing in a web page.

Camera Mouse 2013 version 2.1 is a tool developed by Prof. James Gips and Prof. MargritBetke from Boston College for assisting people with disability to access computer and is available free under end user license agreement.

Figure 7: Eye-tracking tool in use

The Eye-tracking tool is enabled and configured to track the web users eye movements. Web camera available on the users laptop is used as the hardware. No exclusive external hardware is required making this implementation cost effective. Any user with a webcam can use this implementation.

Apache Tomcat 8.0 is used as the web server to accept the *X,Y* coordinates send from the web pages at the client and stored at the server in a file "positions.txt". This file contains the *X,Y* coordinates, permanence time (in milliseconds) i.e. the time user spends at a particular location on the web page.

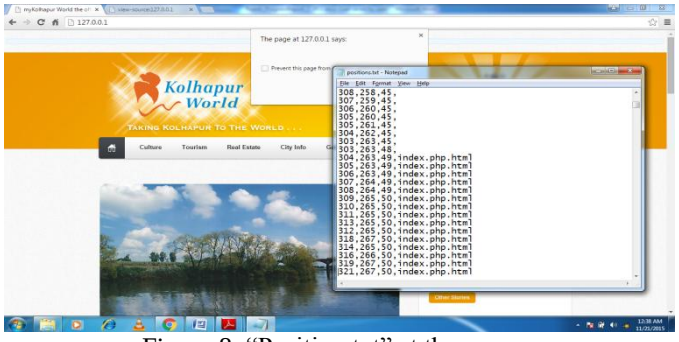


Figure 8: "Position.txt" at the server

C# is used as a programming language to recreate the users eye movements on the web pages based on the positions, and the time obtained above which are mapped to the web objects.

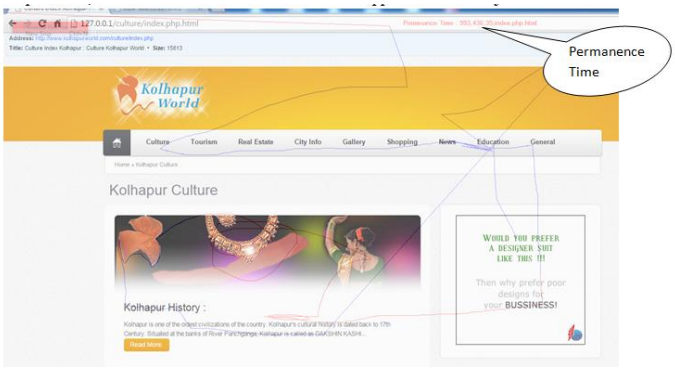


Figure 9. Recreating of the user eye movements on the sample Web page

With this the permanence time on web objects that make up the pages of the site was obtained.Once all web objects are mapped, session details are computed for the recreated user session. Session details include "Total Session Time, Total pages visited, Time spent on each page in seconds".

The above details are used to compute the user behavior vectors which include 1.Object visitor vector, 2. Important Visitor Vector

Object Visitor Vector

$oov = [(wo_1, t_1) \dots (wo_n, t_n)]$where$wo_i$ is the list of objects present on the i[th] page and t_i is the percent time spent by the user's during each session seeing the object wo_i.

Important Object Vector

Important Object Vector (IOV) is calculated by selecting i web objects from Object Visitors Vector according to the following equation: $v = [(o_1, t_1) \dots (o_n, t_n)]$ where oi is the ith most important object and ti is the time in percent spent by the user during viewing the object o_i.

The test website contains web pages which use AJAX to sent the *X,Y* positions of the fixations where the user sees on the area of the web page.The test website is browsed using Mozilla Firefox 47.0 web browser, which is running in responsive mode to render the web page in size 1000*700

pixel. Firefox web developer screenshot capture tool is used to further capture the webpage. The captured webpage are further processed graphically to draw gridlines to help define the web object bound boxes, a part of web object metadata, as shown in Figure 10.

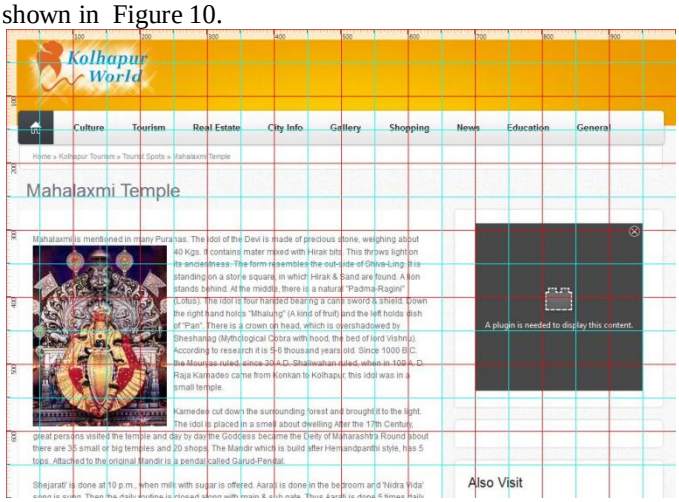
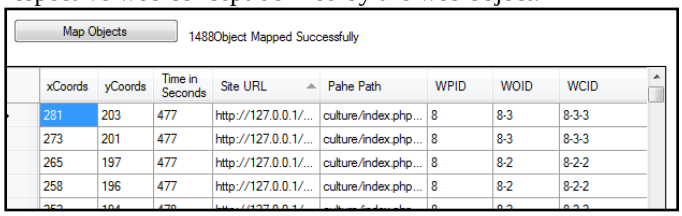


Figure 10: Graphically processed webpage with grid lines.

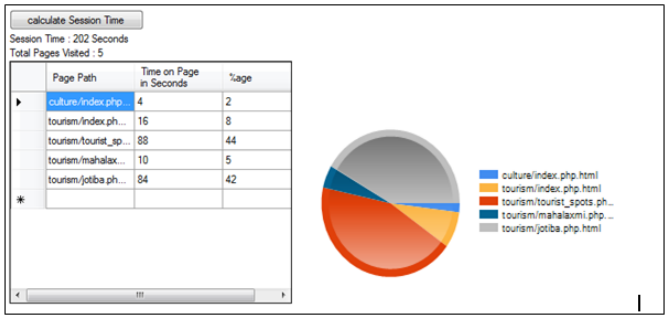
After loading the position file which contains the details of user eye movements, all web objects from all web pages which contributed to the recreated user session are mapped as shown in Figure 11. Each user eye movement is mapped to a webpage, a corresponding web object and a respective web concept defined by the web object.



xCoords	yCoords	Time in Seconds	Site URL	Pahe Path	WPID	WOID	WCID
281	203	477	http://127.0.0.1/...	culture/index.php...	8	8-3	8-3-3
273	201	477	http://127.0.0.1/...	culture/index.php...	8	8-3	8-3-3
265	197	477	http://127.0.0.1/...	culture/index.php...	8	8-2	8-2-2
258	196	477	http://127.0.0.1/...	culture/index.php...	8	8-2	8-2-2

Figure 11: Web object eye movement mapping.

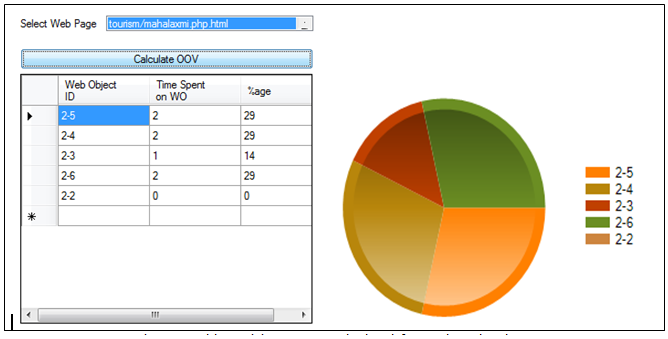
As a sample, 1488 web objects mapping was performed for a user session with details as shown in Figure 12.



Page Path	Time on Page in Seconds	%age
culture/index.php...	4	2
tourism/index.ph...	16	8
tourism/tourist_sp...	88	44
tourism/mahalax...	10	5
tourism/jotiba.ph...	84	42

Figure 12: Computation of Web Session Details

Object Visitor Vector as defined above was calculated for every web page selected as shown in Figure 13. Per web page OVV represents mapped web objects on the respective web page, permanence time of user on the corresponding web page, and percent time spent by user on the corresponding web object.Important Object Vector (IOV) as defined in the methodology can be calculated by setting a threshold user permanence time which changes or can be pre-configured in the system. This will identify those web objects in which the user is more interested relatively.



Web Object ID	Time Spent on WO	%age
2-5	2	29
2-4	2	29
2-3	1	14
2-6	2	29
2-2	0	0

Figure 13: Object Visitor Vector calculated for a selected web page.

Clustering algorithms (*K-Mean)*are used to process the Important Object Vectors to identify the similar web objects.

With this the web objects that make up the pages of the site and characterize by an identifier, its format and a list of concepts that describe its content were identified.

Conclusion

With the methodology described above, every web object identified is associated with a group of concepts that define its content. For any two given objects every concept between them is compared. Every concept in an object is sorted. After sorting every concept for a given object every concept is concatenated to form a string of the form string1,string2,…,stringN. These strings are then passed to a compare function to calculate the Levenshtein distance also known as edit distance between them.[6]

Finally the distance between two objects is calculated using the equation:

$$do(O1,O2) = 1 - \frac{L(O1,O2)}{max|O1|,|O2|}$$

where$L(\cdot,\cdot)$ is the Levenshtein distance between two objects Based on the edit distance, the two objects are compared to indicate whether they are identical, very similar, similar, not similar or totally different.

Table 2. Similarity Measure between identified web objects

Similarity					
objectx1_id	**object_x1**	**objectx2_id**	**object_x2**	**edit_distance**	**similarity**
1-1-1	culture lifestyle	1-2-3	culture cuisine	0.5625	similar
1-1-1	culture lifestyle	1-2-3	culture misal	0.6875	similar
1-1-1	culture lifestyle	1-2-3	About the company	0.1764706	Totally Different
1-1-1	culture lifestyle	1-2-3	culture lifestyle	1	identical
1-1-2	culture cuisine	1-2-3	culture cuisine	1	identical
1-1-2	culture cuisine	1-2-3	culture misal	0.6428571	similar
1-1-2	culture	1-2-3	About	0.2941177	not

	cuisine		the company		similar
1-1-2	culture cuisine	1-2-3	culture lifestyle	0.5625	Similar
1-2-1	culture missal	1-2-3	culture cuisine	0.6428571	Similar
1-2-1	culture missal	1-2-3	culture missal	1	Identical
1-2-1	culture misal	1-2-3	About the company	0.2352941	not similar
1-2-1	culture misal	1-2-3	culture lifestyle	0.6875	Similar
1-2-3	About the company	1-2-3	culture cuisine	0.2941177	not similar
1-2-3	About the company	1-2-3	culture missal	0.2352941	not similar
1-2-3	About the company	1-2-3	About the company	1	Identical
1-2-3	About the company	1-2-3	culture lifestyle	0.1764706	Totally Different

The web administrators can apply this methodology to live sites to know the preferences of the site users. This information can be further utilized to enhance a website by enriching it with the information that users are more interested in while structuring the website it in an appealing format.

References

[1]J. D. Vel´asquez, 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.

[2]Vel´asquez, J.D., Bassi, A., Yasuda, H., Aoki, T.: Towards the identification of keywords in the web site text context: A methodological approach. Journal of web information systems 1, 11–15 (2005)

[3] Andreas Bulling, Jamie A. Ward, Hans Gellersen "Eye Movement Analysis for Activity Recognition Using Electrooculography. IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 4, April 2011

[4] Novak, T., Hoffman, D., Yung, Y., 2000. Measuring the customer experience in online environments: A structural modeling approach. Market. Sci. 19 (1), 22–42.

[5] Nielsen, J., Pernice, K., 2009. Eye tracking Web Usability. New Riders Pub

[6] Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals.
Sov. Phys. Dokl., 705–710 (1966)