Web page characteristics of educational adaptive web sites

Željko Eremić[1], Dragica Radosav[2]

[1]*Technical College of Applied Sciences, Zrenjanin, Serbia, zeljko.eremic@gmail.com*
[2]*Technical Faculty "Mihajlo Pupin" Zrenjanin, University of Novi Sad, radosav@tfzr.uns.ac.rs*

Abstract: Educational information about single topic may be found on many different website pages. Those web pages may have different roles, such as the display of information related to teaching, teaching content or routing to other web pages. Educational material can be placed on adaptive websites. Adaptive websites can customize their view and the structure on the basis of previously recorded user behavior. Documents on which visitors often end their navigation are called target documents, and users often visit waypost documents before visiting the target documents. Characteristics of different types of documents are being investigated in this paperwork. Also guidelines related to the design of such educational web sites are being provided.

Key words: Adaptive website, Waypost, Web design.

Introduction

Information on the syllabus are most commonly found on websites of educational institutions. One topic (that, for example, refers to specific school subject) can often be placed on many web pages. While some of the web pages contain a large amount of text and images, and are able to attract the user's attention for a long time, there are web pages whose function is to direct users to other web pages. There are also web pages from which users are allowed to start their navigation more often than from any other web pages. It's It is essential to notice that there are user visits to different types of documents (web pages) in the user navigations.

Educational, and not only educational, websites can be implemented as adaptive web sites. Adaptive web sites adapt their display and structure in order to stand out from the competition with its efficiency and ability to meet the needs and expectations of customers. One way to significantly improve the navigation of these web sites is to use a shortcut or link. It makes sense to establish a connection among the documents that were not previously linked, and often occur together in the users' paths. In this way, a minimal impact is done on the layout of the pages and it therefore allows a considerable saving of time and effort of users who have a need for accessing useful information. Links are established to documents (files or pages) which are labeled with waypost status or the target document status. "Providing a link (i.e.,shortcut) between these potential wayposts could assist users by reducing the number of clicks they have to make while browsing, pointing them in the right direction towards a specific target document." [1]

Users have a motive for finding specific content, and their satisfaction is very important in this process. "Users' activities in a web site are motivated by the need to reach specific content. The result may be either successful – the user accessed specific document or not successful – the user failed to access specific document, gave up on further search and ended the session. Target documents are those documents

that user has been successfully reached depending on the requirements. Series of documents which have been visited make user's path." [7]

It should be noted that a document can have a waypost status in relation to one or more target documents, while it has no waypost status for other target documents. "If there are some documents which are often accessed in various paths (more often than in some other documents) before target document has been accessed, then those documents can be considered as some kind of road sign and they are candidates for a status of a waypost document." [7]

This paper is built on the results of previous researches. Different strategies to improve the navigation in the adaptive web sites are discussed in papers such as [10], [1], [2] and [4]. In the high percentage of cases, useful information about the users' previous behaviors are is taken from the log files, as described in [9]. On the other hand, there are alternative sources of information such as Page Tagging or cooperation of users described in [6]. In situations where it is necessary to gather more information about the behaviors of users it is possible to use the History Enriched Digital Objects described in [8] and [2].

The aim of this paper is to find a link between the use of specific HTML elements on web pages and the type of those web pages. Also it's it is essential to notice a correlation between the type of document and its size in bytes, and the average retention in this type of document. If the designer of the educational website has a clear idea as to what type of page his pages belong to then he can take into account the results of this study and use the appropriate HTML elements accordingly. In this way, web information related to education can be presented in a way that is suited to their purpose.

METHOD

The research presented in this article is based on information about past behaviors of the educational website visitors recorded in log files. The procedure of preprocessing of data from the log files consists of several steps and is quite standardized. One example of this procedure is given in Figure 1.

The goal of data preprocessing is to obtain a list of user sessions. A list of target documents is obtained from the list of user sessions. Two approaches are described in [1]. Next, extracting of user paths from user sessions is performed. User paths are extracted from the user sessions as extracting arrays of user visits which are ended on a document that was previously declared as target document. Similar user paths are grouped in the clustering process and waypost documents, that are often accessed before accessing the target documents, and are documents which are being searched for in such groups. An example of clustering procedure is illustrated in [5]. Based on the list of user's paths, target and waypost documents, it is possible to form lists of shortcuts that link documents (web pages and files).

Target and waypost documents are calculated on the basis of their position in the user sessions, and on the basis of calculated duration of user visits. The question that is left open is related to what these documents contain? The computer does not have the ability to perform reasoning conclusions at the level of a human being, but on the other hand it has the ability to analyze large amounts of available information. In this article the focus will be on the specific use of certain HTML elements, which exist in various types of documents.

The analysis of the source code of website pages is performed by computer. This analysis counts those HTML elements that are relevant to this research. Documents are divided into those that are target documents and other (non-target documents). A type of accessed document is being determined for each user visit which belongs to any user paths. The survey is conducted on two levels:
- Documents – where all the web pages are being classified into two group of document types: target and non-target documents
- User visits – where all the user visits are being grouped on the basis of document type. Groups of user visits are as follows:
  - Normal (N)
  - Start (S)
  - End (E)
  - Waypost (W)
  - Target (T)

It should be noted that a single document (web page) may have a different type in different user paths. This rule is not applied only when a documents is a target document, because target documents have a "Target" type in each user path.

A list of visited documents and a list of unique documents that exist on the website are is established.

User visits are grouped based on the types of documents listed in Table 1. For each unique document a number of following HTML elements which are

Figure 1: "Overview of data preprocessing in kdwud" [10]

There are several characteristics that can be recorded on the document (web page) or during a visit to the document. The characteristics of the documents that are taken into account in this experiment are certain HTML tags which are contained in the document (web page), and its size is expressed in bytes. A characteristic that is observed at the level of the user visits to certain documents is duration of user visits.

The research should provide answers regarding the type of document and intensity of use of certain HTML elements. As a result, expected answer to the question which groups of documents use which HTML elements below average and above average is expected. It is also expected to compare the intensity of use of certain HTML elements between the different types of documents. Average size of the document in bytes and the length of the retention of certain document type are being compared as well.

Results

An experiment is conducted based on the log files and web pages of Technical College of Applied Sciences in Zrenjanin [12]. Log files are related to activities in the period from 1st June 2011 until 31st January 2012. These inputs are also used in the first experiment in [3].

described on the basis of [13] are being recorded:
- font – Defines font, color, and size for text
- p - Defines a paragraph
- td – Defines a cell in a table
- em - Defines emphasized text
- u - Defines underlined text
- ul – Defines an unordered list
- tr – Defines a row in a table
- b - Defines bold text
- table - Defines a table
- strong - Defines strong text
- ol – Defines an ordered list
- li – Defines a list item
- a - Defines an anchor
- img - Defines an image
- i – Defines italic text

Attribute "length" which represents the size in bytes of the document is also observed at the level of a single document. Duration of user visits expressed in seconds is observed at the level of user visits. The experiment was conducted on over a total of 351 web pages whose HTML code is analyzed, and over 202,220 tracked user visits to the documents where each user visit belongs to a certain user path. User visits are classified based on the types of visited documents (see Table 1).

Table 1: The Frequency Of User Visits, Grouped And Sorted By The Type Of Visited Documents

| Document type	Frequency
Normal (N)	110163
Start (S)	52504
End (E)	25468
Waypost (W)	7570
Target (T)	7085

Documents (web pages) can be divided into those which are target documents and those which are non-target documents. There are 53 target web pages, while the remaining 298 web pages are non-target documents. The average value of the frequency of HTML elements is calculated from the analysis of the source code of these documents. The average length in bytes of these document types is calculated as well (see Table 2).

Table 2: Average Frequency Of Occurrence Of Observed Elements In Target And Non-Target Documents. The Average Length Of These Types Of Documents Are Given In Bytes

Document type	Target	Non Target
font	2,66	1,4
p	29,28	16,55
td	22,57	15,54
em	0,38	0,29
u	0,6	0,62
ul	0,55	0,6
tr	3,15	3,9
b	2,13	2,93
table	0,2	0,29
strong	6,04	8,85
lenght	3149,21	4628,37
ol	0,23	0,34
li	2,74	4,8
a	1,45	5,43
img	0,038	1,12
i	0,04	1,16

In Figure 2 there has been given a ratio of the average frequency of HTML attributes, and the size in bytes between the target and other documents. Target documents have frequent usage of HTML elements font, p, td, and em, while on average they use less the remaining elements, u, ul, tr, b, table, strong, length, oi, li. a img, i. The size in bytes of the target document is only 68% of the length of non-target documents

Figure 2: "Percentage Ratio Of The Average Frequency Of Html Attributes, And The Size In Bytes Between The Target And Non-Target Documents. Target Documents Are Given In Dark Gray Colour
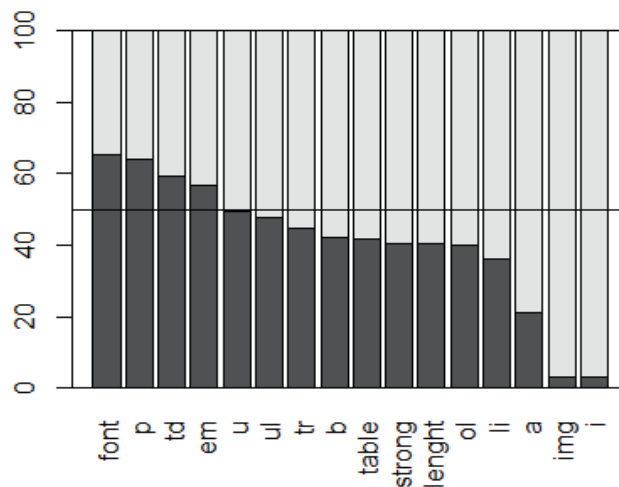
The second part of the experiment is related to the level of user visits. Target documents remain the same when the user paths are considered. Actually user path in this study is considered to be a series of visits from the same IP address with the time difference between the two visits no more than 30 minutes, and which is ended by a visit to a target document. End document is a document which terminates user session, as defined in [1]. User path starts with start document and waypost document status is determined by the methodology given in [3]. Normal documents are all those documents that are not associated with any of these groups.

In Table 3 there have been given the average values of HTML elements appearing in each type of document. The average size in bytes of document types has been also given. In Table 4 there have has been given the same data, but expressed as a percentage and relative to other types of documents.

In Figure 3 there has been given the percentage of ratio of the frequency of HTML elements and the attribute "length", which refers to the size of web pages in bytes.

Average time of the users retention in certain documents can be seen at the level of the document type. In fact, users, on average, retain longest in the target documents (140.285522 seconds), and about the same in normal documents (63.328522 seconds) and

TABLE 3: THE AVERAGE VALUES OF THE FREQUENCY OF HTML ELEMENTS, AND THE SIZE OF DOCUMENTS GIVEN IN BYTES, GROUPED BY TYPES OF DOCUMENTS

	N	E	S	T	W
font	0,222724	0,152309	0,093230	0,893685	0,013606
p	69,543594	61,424026	105,945223	36,822043	61,781373
td	123,299755	88,114050	227,067499	30,875315	82,930118
em	0,054918	0,078398	0,033426	0,479678	0,074372
u	0,822662	1,116367	0,719297	0,363719	1,403830
ul	0,817806	1,114195	0,718478	0,345809	1,401453
tr	18,396376	14,496525	32,117095	5,149712	16,554557
b	10,451222	9,732553	13,031235	3,218828	9,860105
table	0,620725	0,595736	0,972173	0,444546	0,957199
strong	38,836605	35,722672	56,359705	8,141216	37,656406
lenght	14203,039523	14184,497357	18485,144008	4172,196785	15650,966050
ol	0,881666	1,087737	0,833612	0,274626	1,226816
li	11,400751	15,640147	7,224782	2,250975	15,530515
a	26,631654	27,265057	27,973087	1,502640	21,512549
img	0,209534	0,151042	0,170672	0,115958	0,038969
i	0,224185	0,162118	0,174101	0,115958	0,040554

TABLE 4: THE AVERAGE VALUES OF THE FREQUENCY OF HTML ELEMENTS AND THE SIZE OF DOCUMENTS GIVEN IN BYTES, GROUPED BY THE TYPES OF DOCUMENTS, EXPRESSED AS PERCENTAGE COMPARED TO THE VALUES FROM OTHER GROUPS

	N[%]	E[%]	S[%]	T[%]	W[%]
font	16,19158535	11,07256	6,777633	64,9691	0,989129
p	20,72733948	18,30732	31,57678	10,97474	18,41382
td	22,32531523	15,9544	41,11406	5,59045	15,01577
em	7,619118969	10,87665	4,637399	66,54874	10,31809
u	18,58755613	25,22365	16,25209	8,218013	31,7187
ul	18,59604738	25,33562	16,33743	7,863333	31,86757
tr	21,2149362	16,71758	37,03784	5,938714	19,09093
b	22,57578707	21,02338	28,1489	6,953022	21,29891
table	17,2885648	16,59257	27,07717	12,38159	26,66011
strong	21,97677192	20,21467	31,8927	4,606933	21,30892
lenght	21,2952393	21,26744	27,71559	6,255557	23,46618
ol	20,48262998	25,27002	19,36625	6,380038	28,50106
li	21,90465111	30,04995	13,88122	4,324875	29,83931
a	25,39129265	25,9952	26,67025	1,432655	20,51061
img	30,53652494	22,01217	24,87296	16,89919	5,679163
i	31,2707486	22,61325	24,28471	16,17456	5,65673

in the start documents (59.94747 seconds), while the lowest average retention of the documents is recorded in waypost documents (38.325 099 seconds).

CONCLUSIONS

Study was conducted over an educational web site [12]. From the results of the research it is pos-

sible to make a number of conclusions. Target documents have frequent usage of the following HTML elements: font, p, td and em. The size in bytes of the target document is only 68% of the length of other documents.

The results can be observed at the level of user visits. Start documents extensively use HTML elements related to the table: tr and tb, and very rarely elements font and em. End documents also less commonly use HTML elements font and em. Waypost documents very rarely use HTML element font, i, im and relatively frequently HTML elements ul, ol, li. Target documents often use HTML attributes

font and em while less than average HTML elements td, tr, u, b, string, ol, li, a. The remaining (Normal) documents less frequently use HTML element em. The size in bytes of the target documents is noticeably smaller than in other document types.

Users, on average, stay the longest in target documents around 140 seconds, and then more than double less in normal documents - around 63 seconds and in the start documents almost 60 seconds, while the least noticeable user retaining is in waypost documents around 38 seconds. Web designer of educational website could accept the suggestion that in the potential waypost pages does not there is no use of

FIGURE 3: "PERCENTAGE RATIO OF GROUP ELEMENTS WHEN HTML ATTRIBUTES ARE OBSERVED, AND ATTRIBUTE "LENGTH" (INDICATING THE WEB PAGE SIZE IN BYTES)
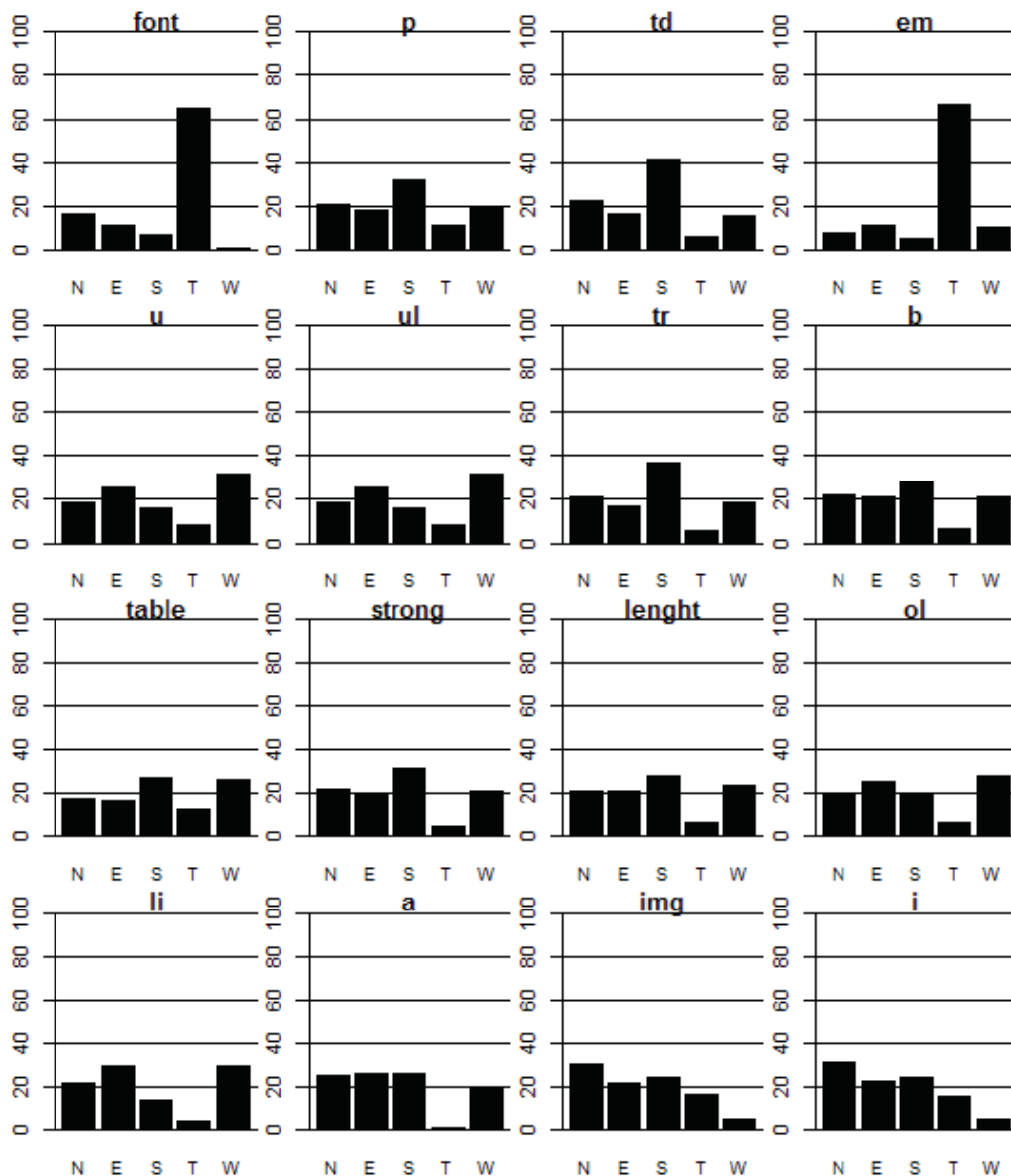
a specific font, italic letters and pictures which probably would not be loaded, because the user retaining time in these documents is the shortest on average.
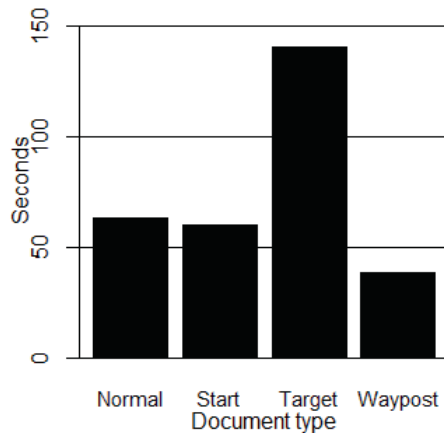
FIGURE 4: Average Duration Of User Visits Depending On The Type Of Document

These conclusions are consistent with the assumptions in [1] and [3], where waypost documents are considered as signpost documents to the target documents. Considering that the waypost documents have noticeably more lists, those lists probably contain information about the possible target documents. At the same time it is logical that users do not retain too much in waypost documents because they have a tendency to find the desired information in the target pages. Also it is logical that users retain longer in the target documents as they need more time to read the information they are looking for and which they need and those that information are is often highlighted by using a different font. This paper is based on a set of input data used in [3], which refers to the educational organization website. Data are is transferred to appropriate form suitable for further processing. The results can be applied to the organization of educational content on web sites, but it can also be applied to other areas. This work is associated with web design because it suggests which HTML elements are to be used in which types of web pages. Research findings provide interesting recommendations for designers of educational web sites, when it comes to the use of HTML elements on web pages.

Authorship statement
Author(s) confirms that the above named article is an original work, did not previously published or is currently under consideration for any other publication.

Conflicts of interest
We declare that we have no conflicts of interest.

References

[1] Bathumalai, G. (2008). Self adapting websites: mining user access logs. The Robert Gordon.

[2] Eremić, Ž. (2012). History-enriched digital objects as a factor of improvement of adaptive educational web site navigation. Journal of Information Technology and Applications, Banja Luka: 39-43.

[3] Eremić, Ž. (2012). Unapređenje navigacije adaptivnih veb sajtova korišćenjem log fajlova. Ph.D. Thesis, Technical faculty "Mihajlo Pupin", Zrenjanin.

[4] Eremić, Ž. & Radosav, D. (2011). Adaptive Web Sites in the Function of Information Access Improvement in Education. Information technology and development of education – ITRO 2011, Zrenjanin: 343-347.

[5] Eremić, Ž. & Radosav, D. (2012). Distances between clusters in the agglomerative hierarchical clustering of strings. Metalurgia International. No. 8, Vol.17: 67-74.

[6] Eremić, Ž. & Radosav, D. (2012). Collaborative user support as a contribution to navigation improvement of adaptive websites for distance learning. Information technology and development of education – ITRO 2012, Zrenjanin: 265-269. Eremić, Ž., Radosav, D. & Markoski, B. (2010). Mining User Access Logs to Optimize Navigational Structure of Adaptive Web Sites. Proceedings of the CINTI 2010 : 11th IEEE International Symposium on Computational Intelligence and Informatics, Budapest: 271-275.

[7] Hill, W. & Hollan, J. (1993). History-Enriched Digital Objects. Proceedings of Computers, Freedom: and Privacy (CFP'93).

[8] Pamnani, R. & Chawan, P. (2010). Web Usage Mining: A Research Area In Web Mining. Paper presented at the International Conference on Recent Trends in Computer Engineering ISCET 2010, Punjab, India: 73-77.

[9] Perkowitz, M. & Etzioni, O. (1997). Adaptive Sites: Automatically Learning from User Access Patterns. In Proc, 6th Int, World Wide Web Conf, Santa Clara, California.

[10] Raju, G. & Satyanarayana, P. (2008). Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology. IJCSNS International Journal of Computer Science and Network Security, Vol. 8: 179-186.

[11] Technical College of Applied Sciences in Zrenjanin. (2013). www.vts-zr.edu.rs

[12] W3Schools. (2013). http://www.w3schools.com/tags/default.asp.