Preprocessing and Content/Navigational Pages Identification as Premises for an Extended Web Usage Mining Model Development

Daniel MICAN, Dan-Andrei SITAR-TAUT
Faculty of Economics and Business Administration
Babeş-Bolyai University, Cluj-Napoca, Romania
{daniel.mican|dan.sitar}@econ.ubbcluj.ro

From its appearance until nowadays, the internet saw a spectacular growth not only in terms of websites number and information volume, but also in terms of the number of visitors. Therefore, the need of an overall analysis regarding both the web sites and the content provided by them was required. Thus, a new branch of research was developed, namely web mining, that aims to discover useful information and knowledge, based not only on the analysis of websites and content, but also on the way in which the users interact with them. The aim *of the present paper is to design a database that captures only the relevant data from logs in a way that will allow to store and manage large sets of temporal data with common tools in real time. In our work, we rely on different web sites or website sections with known architecture and we test several hypotheses from the literature in order to extend the framework to sites with unknown or chaotic structure, which are non-transparent in determining the type of visited pages. In doing this, we will start from non-proprietary, preexisting raw server logs.*

Keywords*: Knowledge Management, Web Mining, Data Preprocessing, Decision Trees, Databases*

1 Introduction

Web Personalization System is defined as any action that adapts the information and services provided by a web system to the needs of a user or a group of users, taking into account the available data on their web navigation behavior and individual interests in combination with the content and structure of a web system. A personalization system of a web application should be able to provide the users with the information they need without their explicit request for it. The new generations of tools for web personalization try to incorporate more and more the patterns discovered by Web Usage Mining (WUM) in order to improve the scalability, accuracy and flexibility of the recommendation systems. As a result, these systems try to implicitly obtain the users' notes from the data collected after they used the application in order to find out and portray their profile [9]. (WUM) is successfully used in web applications to collect data on users' behavior, to develop a recommendation system for the content and services that can be of interest to the current user, to efficiently target marketing campaigns, etc.

Web Usage Mining

WUM refers to the application of data mining techniques for the discovery of usage patterns from the web data in order to understand and meet the needs of web-based applications. Modern web-based applications are able to capture the users' navigational behavior up to the individual mouse clicks. They have the capacity to adapt to individual users on a large scale, a phenomenon known as "mass personalization" [15]. The goal of the data mining process is to discover new useful information by applying various techniques such as: classification, clustering, forecasting, association rules, as well as identification of sequential patterns.

Web usage mining is able to manage the discovery of the users' various access models by using the log files, which register all the clicks by each user during a web application interaction [7]. WUM techniques make use of the data retrieved from the log files, which provide information on the activities carried out by a user during a navigation session.

These log files record the user's behavior and how they interact with an application from the instant they access a site to the moment they leave the site. Due to this characteristic, the application is also known as web log mining [16].

Data collection, preprocessing and filtering

Web server access logs represent the raw data source and sometime the single one that we can have in WUM processing [1], [14]. Some websites may have registered members section, but generally the owners cannot afford having a login-based access. Theoretically, they have to share the information to the whole world in order to be able to sell directly or indirectly as much products and services as possible. Fidelity strategies also have their benefits, but it is beyond the scope of this paper to go into more detail. Mainly, the login info and personal details are input by users and we cannot give full confidence to these records. Obviously, for online banking, taxes, postal services, police, or other web based applications, the personal details are entered by trusty third parties - authorities, companies' delegates - or in collaboration with users, and are proved with papers (ID cards, birth date or company registration certificates, etc.). Such rigorous and strictly supervised measures would severely restrict the number of visitors who are in fact potential buyers. Likewise, some behavioral activities implemented at the registered user layer sessions are not 100% effective due to the fact that they rely on computers' security settings (like accepting cookies or not). These are few but enough reasons treating all users as being equal. We will focus here just over the "reactive" [14] behavior of the user, the "proactive" [14] one being explored and exploited later.

In order to assure that only the data that contain information useful for the users' identification and their navigation sessions is retrieved, it is absolutely necessary for the log files to be cleaned and filtered. Therefore, it is important to identify and discard the data recorded by web robots/web crawlers, the images, sounds, java scripts and all the files encapsulated in the templates of web pages, information that is often redundant and irrelevant for the goal at hand.

Following the deletion of the irrelevant data, the attention should be focused on the identification of the real users and the sessions associated to them. According to [16], a session associated to a user is defined as "a sequence of requests by a single user during a certain period of navigation, and a user may perform one or more sessions during a period of time". A considerable amount of research [2], [3], [15] has been carried out on the discovery of ways to identify and isolate the users' sessions. However, the visitors' identification and their association to various sessions is a difficult task. Furthermore, in [16], we find more well-known methods that can help in the delimitation of a user session, the most reliable ones being those based on different navigation behaviors, which result in discrepancies in their consecutive requests. Conforming to investigations that have been carried out on this matter [2], a 30-minute threshold, which has become a standard for the delimitation of a visitor session [15], [16], was established. Moreover, it should also be taken into account the fact that the type of web application, the context as well as the content of the web pages can influence the interval between two consecutive requests.

Preprocessing of log files vs. online data extraction tool

The analysis of log files raises a series of problems. These files contain significant data for the web mining process, but they also include a large amount of noise. Among the problems identified in log files processing, we mention the following:

- the need for a large storage space due to the considerable volume of data saved on a disc;
- the existence of a large amount of data that is irrelevant for the web mining process, as a result of accessing the files

containing images, java scripts, etc.
- the storage of requests performed by the search engines and various automated scripts;
- the storage of data containing error messages, such as 400 (bad request), 500 (internal server error), etc.;
- in general, prevention from real time data usage due to the batch processing of log files.

The development of a tool for data collection solves most of the problems that have appeared in preprocessing the log files. The benefits of such a tool are:
- the storage space is reduced as only the data relevant to the web mining are stored;
- the data resulted from accessing files containing images or java scripts are not taken into account;
- the search engines are immediately identified and the data generated by them are not stored in the database;
- real-time storage in the database of only those content and navigational pages of the site in an interactive process.

The noise in the log files can compromise the precision of the web mining process. Therefore, it is crucial that the log files be cleaned in order to improve the results of the process of web usage mining.

Due to the complex process of preprocessing and cleaning the log files, appeared the problem of developing a tool which to collect only the data from the web sites that were strictly useful in the process of web usage mining. The focus was thus shifted from the process of preprocessing and filtering the information that contains noise to the collection of the information that is effectively needed for the specific goal of the user (figure 1).

The traffic on a web site is divided into two important categories: human users and search engines, respectively web crawlers. Search engines are programmed to navigate and index the data contained by web sites. The process of identifying the web crawlers is important as, according to our research, they can generate over 90% of the traffic on web sites. The identification of web crawlers is an essential stage in the data collection process. This involves the identification of the search engines in order to dismiss the data generated by them and retrieve only the information generated by human users, namely the relevant one (no image views, scripts, etc.).

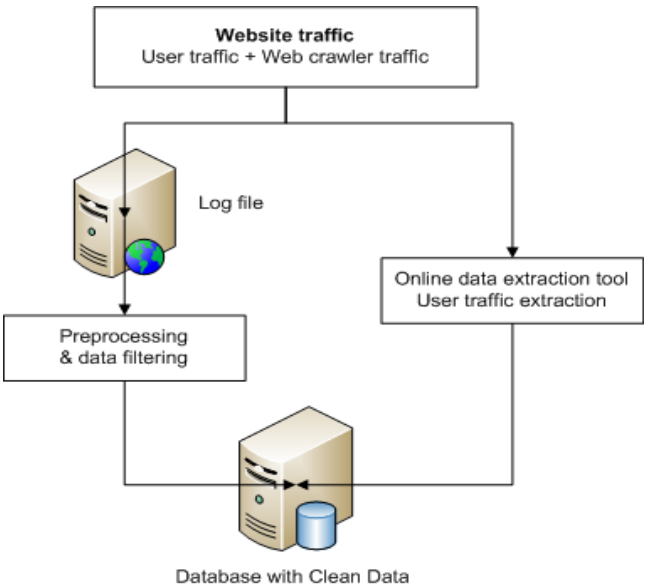

Fig. 1. Classical web log mining vs. online data extraction tool

Here are some strategies to filter the robots' visits:
- to store the IP ranges for the bots to whom our SEO instruments are directly

addressed. According to the new coming robots, the list can also be updated manually or automatically;

- in PHP we can use $_SERVER[‘HTTP_USER_AGENT’]. This PHP system variable returns an identification string for the visitor, not for the IP. The robots contain specific codes that are able to capture a wider range of IP addresses. Such string examples can be ‘Google’ or ‘googlebot’ for Google search engine, ‘Yahoo’ for Yahoo, ‘msnbot’ for MSN, ‘Lycos’ for Lycos spider and so on. This technique is more elegant than filtering isolated IPs or ranges. The problem is that not all the robots spidering the web can be identified with this method;
- if a non-desirable search agent can establish rejection policies in robots.txt file for the whole site or just for certain sections (E.g. the shopping cart, site administration pages, etc.);
- machine learning algorithms can be used here, too. Regarding the visits that we have identified with the previous strategies, a large amount of data is available. Likewise, there are already identified visits coming from known human sources. According to previous researches, we can identify some patterns that could be applied to the new experiences. From this point of view, in our case, the training dataset covers more than 95% from the analyzed data. The classifiers able to work with incomplete or missing data can manage this situation. The neural networks, like the naive Bayesian classifier used in spam detection [5], [10], are suitable for this domain. Moreover, some decision trees - like C4.5 - accept “imperfect” data.

However, we have identified all robot visits with the first two strategies, not being enforced to test the last one. As stated by [8], [13], the preprocessing means for our system cleaning the bounces (0 seconds visits) or the unrealistic visit periods of more than 30 minutes from the database.

Content pages, navigational pages and visualization time

The length of time spent on a web page is important since it is a measure that reflects the visitor’s interest providing implicit ratings for that page.

The length of the visit on a web page comprises the difference in time between two consecutive visits of the same user, during the same navigation session.

Analyzing the time spent by the visitors on a web site, we distinguish between two categories of web pages [1], [6], namely content pages and navigational pages. Navigational pages contain links to those pages that are of interest to the user, that is, to the content pages. Content pages entail information needed and requested by the visitors.

According to [1], the distinction between navigational and content pages is made in terms of the distance in time between the requests for two consecutive pages. Therefore, if the lap of time between two successive pages A and B is bigger than a certain threshold, then A can be considered a content page; if the lap of time is smaller, then it is a navigational page. Thus, we argue that users’ sessions are composed as a sequence of navigational and content pages.

In this study we intend to discover a time threshold for the identification and classification of web pages into the two categories mentioned above. Based on the length of time spent by a user on a web page, we aim to identify whether they are navigational or content pages. We deem that the distinction between content and navigational pages is essential in the recommendation process as well as with regard to the development of some ample web personalization systems.

Regarding specific web sites or web sections we will be interested mainly in content pages, which depend on the interest area of a site or its sections. Still, the site design, usability, content quality, the brand, reputation, online history etc. have already been mentioned above. If the data stored by the server does not provide enough

information to distinguish between a navigational and a content page, then some rules - heuristics - [13] can be tested and adopted if the results are satisfactory. One rule states that if the length of the visit on a page takes more than a given threshold then this page is a content page, otherwise it is a navigational one [1]. It is obvious that the length of a visit on a navigational page takes less than on a content one and also the rule can be applied easily. On the other hand, it is also stated in the literature that "there is no widespread threshold for page-stay time" [8], [14]. Since the problem is still debatable, we will test the rule mentioned above using real data taken from three different sections of two successful Romanian sites. In order to identify the content and navigational page status for each site or each section type that are based on proved heuristics, we can extend our system by issuing some specialized tools for additional preprocessing of the logs.

2 Methods

The tool designated to check the hypothesis refers to C4.5 in its cloned implementation in Weka [17]. This was developed by J. Ross Quinlan and is probably the most popular machine learning algorithm which also constitutes an extension of the capabilities of its predecessor, namely of the ID3 algorithm (Iterative Dichotomiser 3) [11]. One of the capabilities examined in this paper refers to the continuous numeric values handling, as for instance: the time spent on a page. The C4.5 algorithm is based on Information Gain Theory and its output rules are generated from decision trees [12], which are very suggestive regarding their resulted visual interpretations. The strengthen and accuracy of the conclusions obtained by using this algorithm, can be quantified by few indicators, such as kappa statistic, mean absolute, root mean squared, relative absolute, and root relative squared errors; TP and FP rates, precision, recall, F-measure, and ROC curve [17]. These indicators provide additional interpretations that enrich the simple conclusions derived from the number or rate of correctly/incorrectly classified instances.

Preprocessing is a very important and significant resource consuming step in a KDD (Knowledge Discovery in Databases) process [4], being responsible for the quality of data provided to the data mining tools. Furthermore, we will develop an application that collects data from different sites, after a necessarily cleaning process. Under such circumstances, almost the whole needed preprocessing will be made in real time and the data will also be prepared for future web usage mining. Once we are aware of the sites' structure, we can determine for sure the type of a page. Classification is suitable under any of these circumstances.

3 Case Study

We have developed a log system that collects data for the following sites: www.bizcar.ro - Classifieds and Press Review sections - and www.superfarmland.ro with its e-Commerce-oriented section. We consider that for our current goals and for future research in WUM area, the database depicted in figure 2 is able to accomplish the requirements. The test version ran for exactly one week. We consider that this period was long enough in order to capture the user's behavior for the site sections from bizcar.ro. In order to track all seasonal events, this period is not representative for the latter examined website, but for establishing similarities in comparisons, we opted for equal periods. Likewise, compared to the other two sites, the traffic here is limited, due to the less interest shown in the area. Using C4.5 we test if a threshold time can be taken into consideration so as to distinguish between content and navigational pages. If this hypothesis is confirmed, theoretically, we can extend our web usage mining model to any site, after assigning the page type, by applying the heuristic discovered.

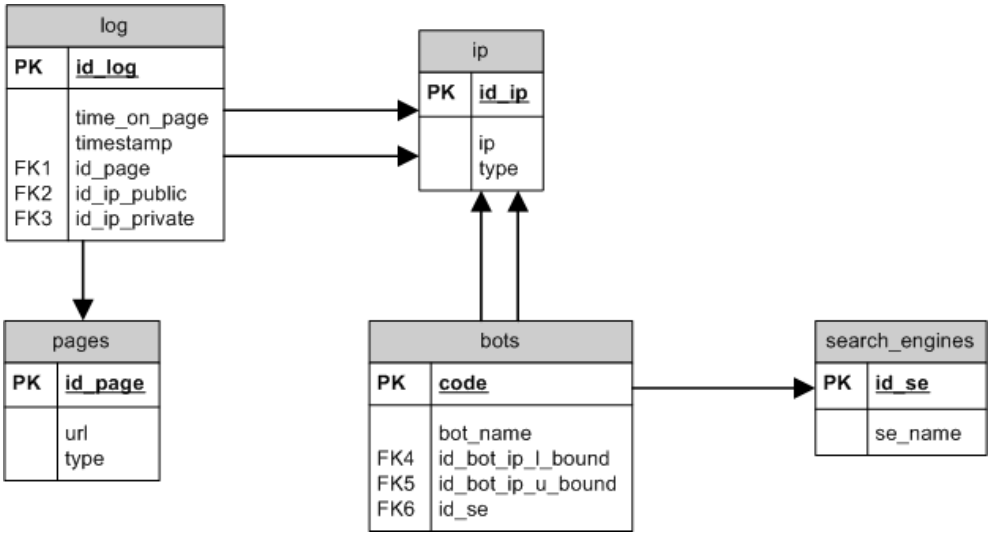

Fig. 2. Proposed DB schema

SEO techniques are widely and effectively used in our studied sites. They generate around 89.09% of the visits using the search engines' agents (web robots). According to the aim of this paper, we will ignore these automatic visits in order to get as clear human being-website interaction as possible. Furthermore, we collect these records so as to study the effectiveness of SEO process that we performed, and to improve it if we notice any lacks. The study of the previous logs did not allow us to determine exactly which entry came from a robot and which did not. However, we cannot issue a simple rule like implying the length of a visit, because we have already had 0 second visits for the known bots, but also 50 seconds. The strategies mentioned above succeeded in robots detection process.

During the period between 10.24 - 10.30.2009 the raw log encountered 479,436 records, 52,311 being stored for our analysis. The data were collected from bizcar.ro, from sections that contain classifieds (24,489), and press reviews pages (24,662) as well as from superfarmland.ro, e-commerce platform (3,160). Thus, 49,151 records are from bizcar.ro and 3,160 from superfarmland.ro.

The rest - 423,390 automated visits performed by web agents, 2,395 bounces (0 seconds page landing), and 2,173 unrealistic page staying time (longer than 30 minutes) – were filtered by our application.

3.1 Hypothesis tested on whole sections

The instances from all 3 sections - 52,311 - have been tested without taking into account the section they belong to. Actually, we want to test is whether the result - if positive - can be applied to any web site and any web site section.

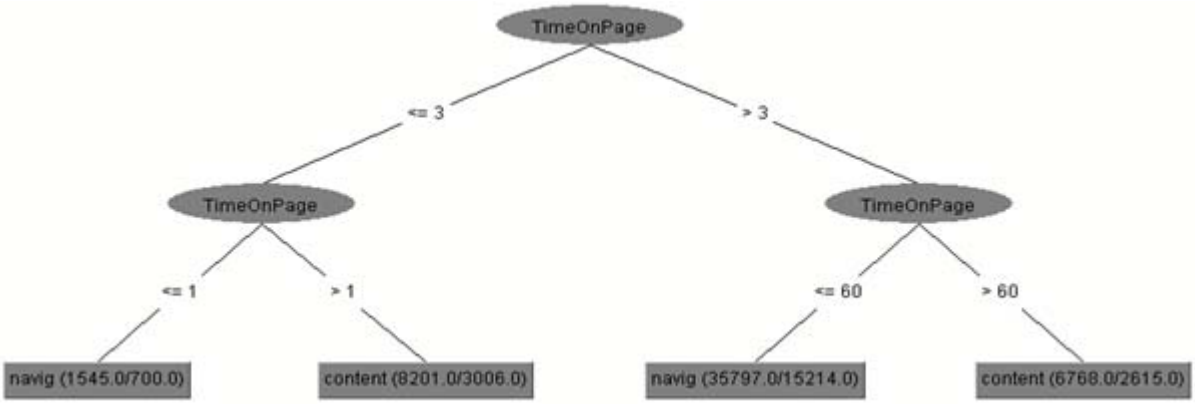

Fig. 3. Navigational/content page status: All data, individual time on page factor

According to figure 3, the pages visited for 1 second or between 4 and 60 seconds, fall under the category of navigational pages. The rest of the pages, namely those visited for 2 and 3 seconds or for more than 1 minute, refer to content pages. The result does not confirm the existence of a clear threshold. Still, it is surprising that one second landing pages are considered to be content pages, whereas those that are visited between 4 and 60 seconds are regarded as navigational ones. Nevertheless, these rules are able to identify correctly just 58.75% instances, what we regard as unacceptable. Further investigations will be carried out in order to establish if the average duration of a visit on a page, instead of considering individual page times, qualifies better a page type.

At this point we have 28,499 instances of distinct pages visited during a week. Therefore, the mean will probably uniform the user behavior regarding an accessed page, if the number of visitors is significant. In our case the most visited page has 701 visits. Even if such a rule is not easily implemented as a single row condition, we consider this approach to be more appropriate.

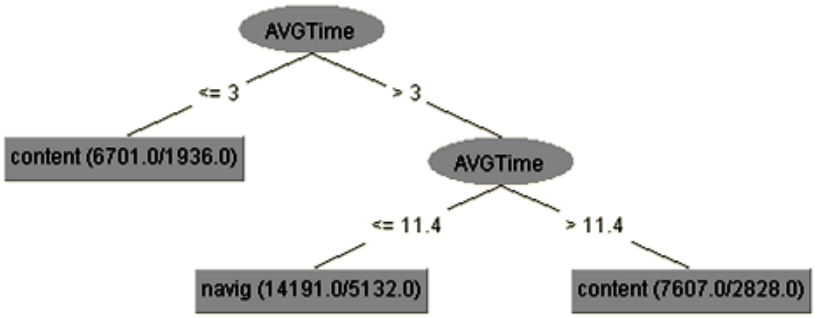

Fig. 4. Navigational/content page status: All data, average time on page factor

Unfortunately, figure 4 denotes a remarkable negative fact as from 6,701 1 to 3 seconds page visits (in average) only 1936 could not be considered content pages. The rules generated by this decision tree classify correctly 18,575 (65.18%) instances. However, neither the rules nor the accuracy are satisfactory for our goal.

In the following paragraphs we will test if the site section has any influence on the results obtained before. Taking the whole data into consideration, it might be possible that there are interest domains in which the heuristic related to the duration of a visit can determine the type of a page.

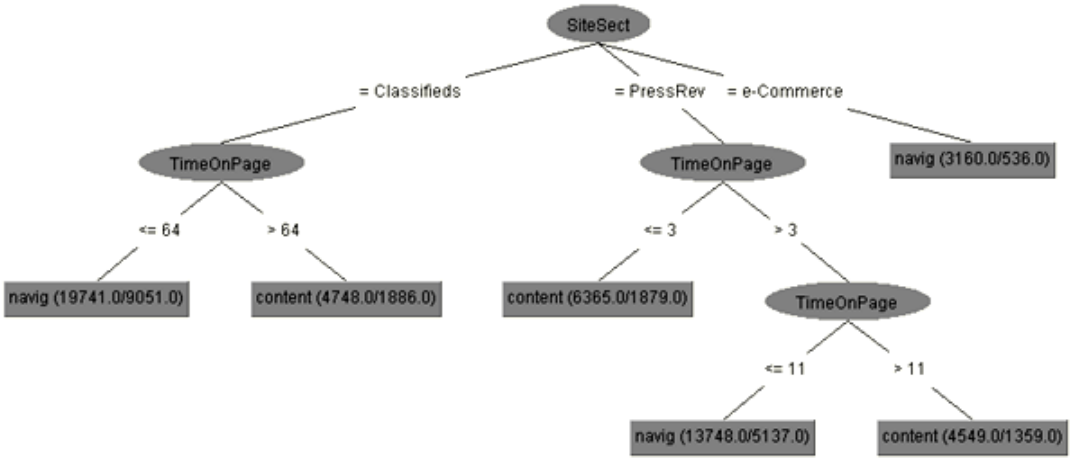

Fig. 5. Navigational/content page status: All data, individual time on page and site section factors

The rules depicted in figure 5 indicate that all e-commerce related pages are navigational. Only classifieds section indicates a possible threshold. This is a little bit higher but it can be explained by the specificity of the section as follows: the users can read all topics related to the announcements that are to be found on a page and then go into more detail if one item is interesting. For press review section the users try to find interesting news in 3 or less seconds. The threshold of 11 seconds sounds plausible. Nevertheless, all the hypotheses presented in this paper are reliable with a certain factor of 61.96%, which cannot be considered significant. Furthermore, we want to test whether these hypotheses are marked considering the average time instead of the individual one.

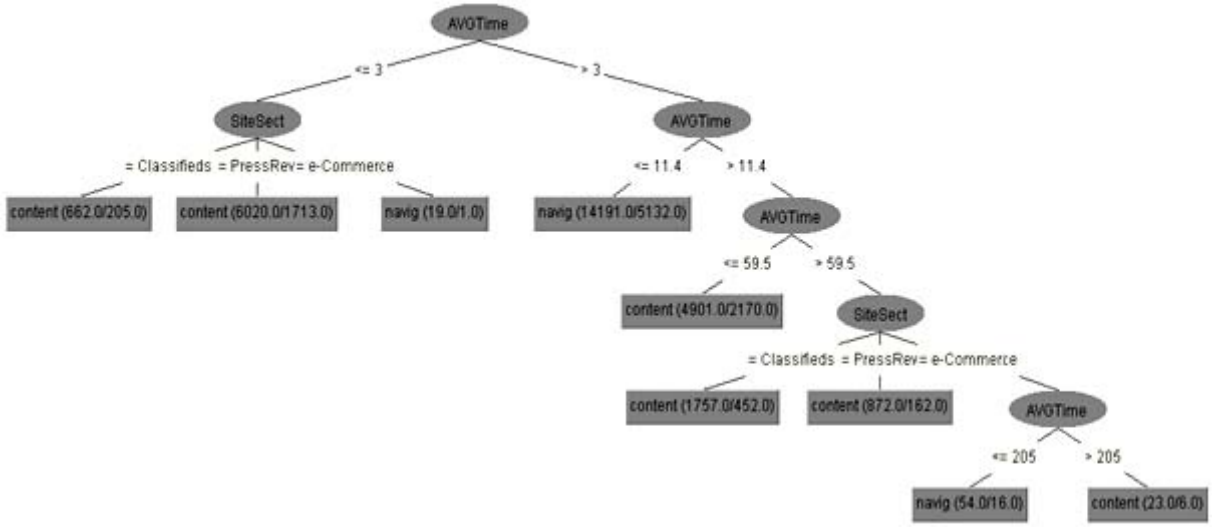

Fig. 6. Navigational/content page status: All data, average time on page and site section factors

The structure of figure 6 was expected to be simpler than figure 5. The explanation is that the content pages from classifieds sections are those visited averagely for more than 59.5 seconds. Likewise, the press review content pages visited for 3 or fewer seconds are identified correctly in about 68% of all instances. Overall, the rules presented in figure 6 can be applied just for 65.18% of the instances. In the following sections we will perform inner sites tests.

3.2 Hypotheses tested on bizcar.ro website

Regarding the section as being irrelevant for bizcar.ro, the tree illustrated in figure 7 generated some non-certifiable rules. They can be applied correctly for 59.41% from the 49,151 page visits.

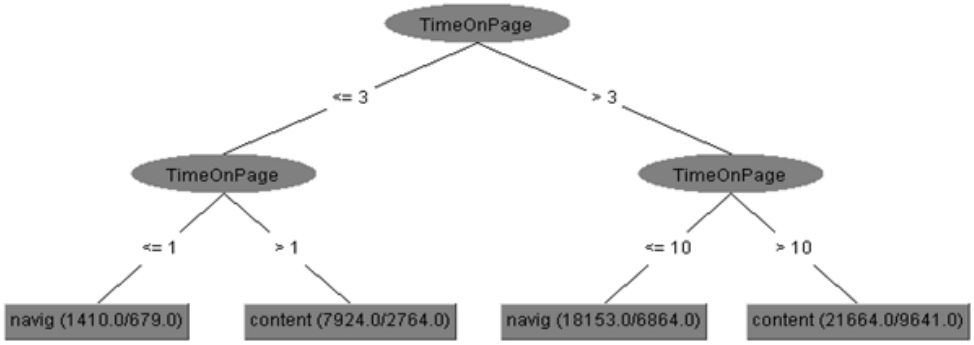

Fig. 7. Navigational/content page status: bizcar.ro data, individual time on page factor

We hope in better results considering the average time.

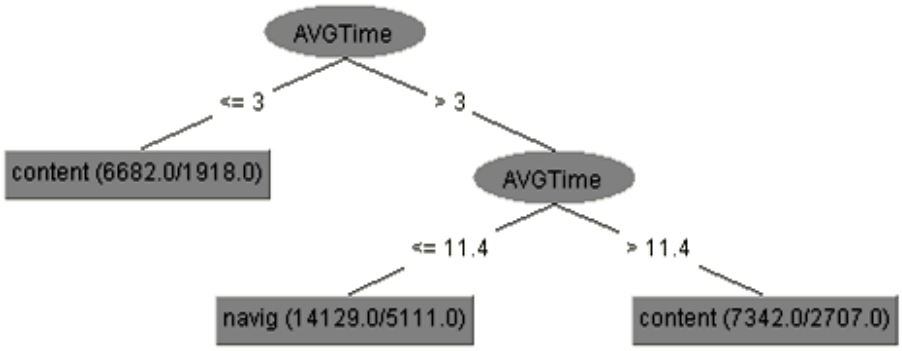

Fig. 8. Navigational/content page status: bizcar.ro data, average time on page factor

The simplified structure of the tree shown in figure 8 is more suitable. This is related to the content pages that are visited on an average for more than 11.4 seconds. Still, according to a precise rule, more than one of 3 pages that are visited for 3 or fewer seconds are regarded as content pages. All the rules depicted here can be applied in 65.34% of all instances from 28,153 distinct pages.

Furthermore, we will consider the two sections we have on this site.

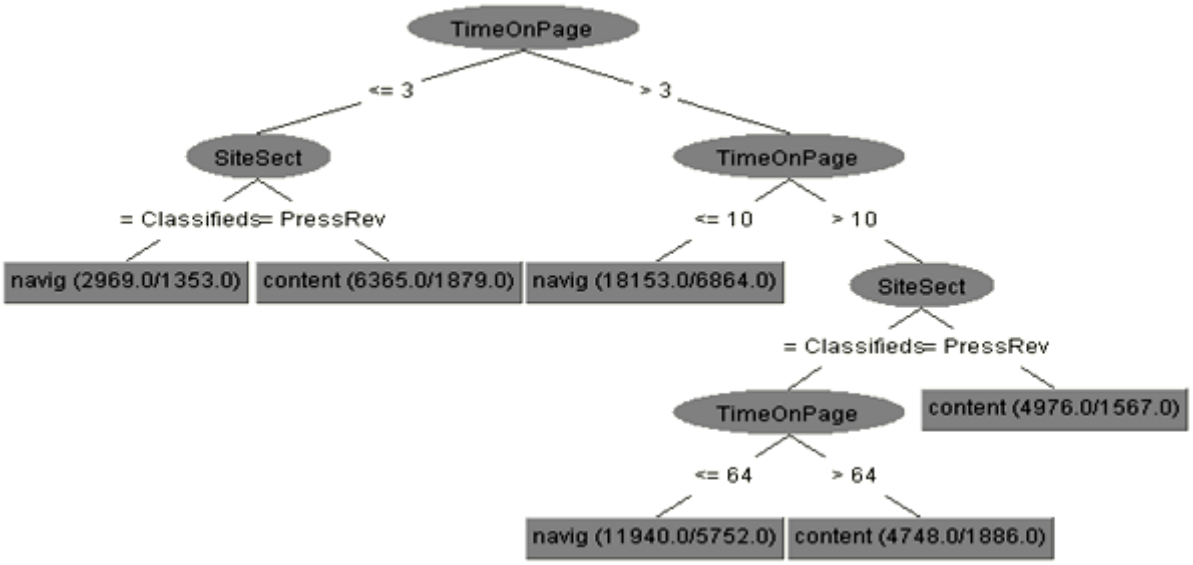

Fig. 9. Navigational/content page status: bizcar.ro data, individual time on page and site section factors

Since in figure 6 we had more than a knot labeled “SiteSect”, we are not able to deduct the rules for this situation. Moreover, 60.62% of the instances follow the rules illustrated in figure 9. For press reviews the content pages require a staying time of more than 10 seconds or less than 3 seconds, whereas for classifieds it must be of at least 64 seconds. For classifieds there is a page landing time of 3 seconds or fewer for navigational purposes, which is acceptable. Besides, between 4 and 64 seconds we also have navigational pages for the same category. Therefore, 64 seconds can be considered a threshold for classifieds section. The other section together with the poor rate of the successful rules mapping, constitute the main problem. But what if the average time is able to correct the problem?

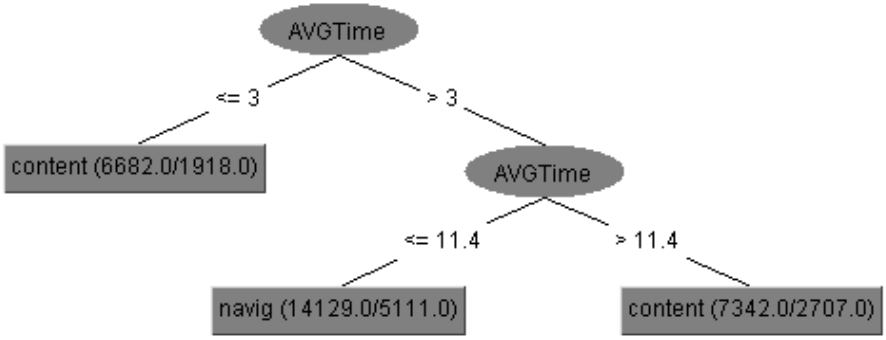

Fig. 10. Navigational/content page status: bizcar.ro data, average time on page and site section factors

From figure 10 we observe, that even if we considered the site section, the average staying time is completely independent of this aspect on bizcar.ro. Unfortunately, the branch from the left side alters the possible success. Otherwise, a threshold of 11.4 seconds could be issued but not generalized for this site. Still, the correctly classified instances rating of 65.34% is poor.

3.3 Hypotheses tested on superfarmland.ro website

The number of the user-visits on this site is only 3,160. Since we considered just the traffic for one section of the site - e-commerce - including or omitting the site section variable, this did not affect the results of our analysis. The conclusion can be easily drawn on the basis of figure 5 (the first level right hand branch), and consist in a single leaf. Even if the indicator reflecting the correctly identified instances is relatively good - 83.038% - the kappa statistics is 0 due to the fact that the model is not able to identify at least one content page.

Furthermore, we will test our last hypothesis regarding the possible heuristic that we wanted to issue, by considering the average page landing time on this site.

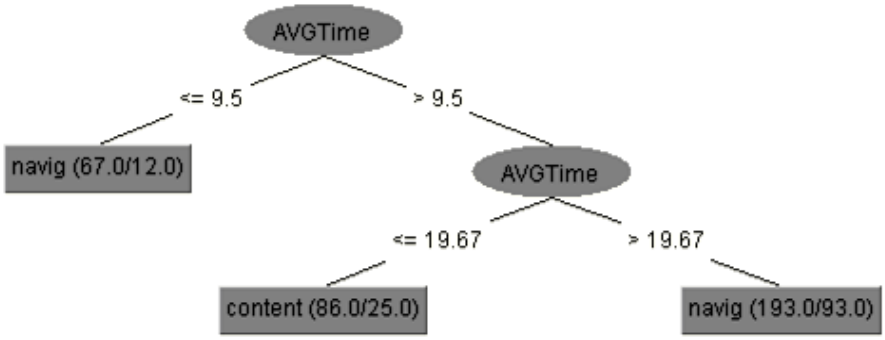

Fig. 11. Navigational/content page status: superfarmland.ro data, average time on page factor

There were visited only 346 different pages. From the rules illustrated in figure 11 we can consider only the one indicating that navigational pages are visited averagely for at most 9.5 seconds. The other rules seem to be unexplainable so that no further tests are necessary.

Even if the threshold theory was infirmed, the results that we obtained aim to redesign the sites. 0 seconds visits, content pages visited in less than 3 seconds, or impressive page landing time on navigational pages are abnormal behavior. These critics are not addressed to the user. This is the normal user reaction for a service that we offered. This service need to be improved.

4 Conclusions

In this paper, we applied the web usage mining in order to discover behavioral

patterns for human beings - website interaction, and tried to enrich the website capabilities so as to anticipate the user moves, by providing real time and adaptive solutions. On accomplishing this goal, we started with known websites, but in the future we intend to extend our framework to other non-proprietary websites. Along with this, we developed some software tools and a database able to manage the necessary and already preprocessed data from custom access log servers. In order to apply the system for our websites, we distinguished between content and navigational pages. The hypothesis, according to which there is a threshold time for making the mentioned distinction, has been rejected, using matching learning tests on various real and clean datasets. This negative result does not affect the aims related to our websites, but indicates that we need not only the additional preprocessing and heuristics, but also the adoption of new strategies for any other website and site section. Nevertheless, the "navigational" or "contextual" aspect cannot be considered absolutely a website page attribute.

Likewise, some rules generated during the testing process made us think about improving the usability of our site in future works, in order to reduce the bounce rate, page loading speed, usage intuitiveness, and to improve the information quality.

Acknowledgement

This paper was supported by Research Project No. 947, ID_2246/2009 Code, and part of PN II Program financed by the Romanian Ministry of Education, Research and Innovation – The National University Research Council.

References

[1] M. Baglioni, U. Ferrara, A. Romei, S. Ruggieri and F. Turini, "Preprocessing and mining web log data for web personalization," *Ser. Lecture Notes in Computer Science*, Vol. 2829, October 2003. [Online]. Available at: http://www.springerlink.com/content/7at3ta2agrflf0a9/

[2] L. D. Catledge and J. E. Pitkow, "Characterizing browsing behaviors on the world-wide web," *Tech. Rep*., 1995. [Online]. Available: http://hdl.handle.net/1853/3558

[3] M. S. Chen, J. Soo Park and P. S. Yu, "Efficient Data Mining for Path Traversal Patterns," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 10, No. 2, Mar./Apr. 1998, pp. 209-221.

[4] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, Vol. 17, 1996, pp. 37-54.

[5] J. Goodman, G. V. Cormack and D. Heckerman, "Spam and the ongoing battle for the inbox," *Communications of the ACM*, Vol. 50, No. 2, 2007, pp. 24-33.

[6] Y. Li and B. Q. Feng, "The Construction of Transactions for Web Usage Mining," Vol. 1 *2009 International Conference on Computational Intelligence and Natural Computing*, 2009, pp. 121-124.

[7] B. Liu, *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag Berlin Heidelberg, 2007.

[8] R. Cooley, B. Mobasher and J. Srivastava, "Data preparation for mining world wide web browsing patterns," *Journal of Knowledge and Information Systems*, Vol. 1, No. 1, 1999, pp. 5-32.

[9] B. Mobasher, R. Cooley and J. Srivastava, "Automatic personalization based on web usage mining," *Commun. ACM*, Vol. 43, No. 8, August 2000, pp. 142-151.

[10] L. Pei-yu, Z. Li-wei and Z. Zhen-fang, "Research on E-mail Filtering Based On Improved Bayesian," *Journal of Computers*, Vol. 4, Issue 3, March 2009, pp. 271-275.

[11] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc, 1993.

[12] A. Sharan and H. Imran, "Machine Learning Approach for Automatic Document Summarization," *Proceedings of World Academy of Science,*

Engineering and Technology, Vol. 39, March 2009.

[13] M. Spiliopoulou and L.C. Faulstich, "WUM: a tool for Web utilization analysis," *Proceedings EDBT Workshop WebDB'98*, LNCS 1590, Springer, Berlin, Germany, 1999, pp. 184-203.

[14] M. Spiliopoulou, B. Mobasher, B. Berendt and M. Nakagawa, "A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis," *INFORMS J. on Computing*, Vol. 15, No. 2, Apr. 2003, pp. 171-190.

[15] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *SIGKDD Explorations*, Vol. 1, No. 2, 2000, pp. 12-23.

[16] A. Vakali and G. Pallis, *Web Data Management Practices: Emerging Techniques and Technologies*. Idea Group Publishing, 2007.

[17] I. H. Witten and E. Frank, *Data Mining: Practical Machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.

Daniel MICAN is a PhD student in Business Information Systems at Babeş-Bolyai University of Cluj-Napoca. His main research areas are: Web Applications and Search Engine Optimization. He is Teaching Assistant at Faculty of Economics and Business Administration, Babeş-Bolyai University of Cluj-Napoca.

Dan-Andrei SITAR-TĂUT has a Bachelor's degree in Business Information Systems from Faculty of Economics and Business Administration, Babeş-Bolyai University of Cluj-Napoca and a Master's degree in Informatics Strategies Applied in Economy and Business from the same educational institution. He also holds a PhD diploma in Cybernetics and Economic Statistics. He is the author of 2 books, about 35 papers in the field of Databases, ERP, Data mining, and Web related fields.