UDC:004.738.52
Original scientific paper

WEB LOG EXPLORER – CONTROL OF MULTIDIMENSIONAL DYNAMICS OF WEB PAGES

Mislav Šimunić
Faculty of Tourism and Hospitality Management, Opatija, Rijeka University, Croatia
mislavs@fthm.hr

Željko Hutinski, Mirko Čubrilo
Faculty of Organization and Informatics,
Pavlinska 2, Varaždin, Croatia
{zeljko.hutinski, mirko.cubrilo}@foi.hr

Abstract: *Demand markets dictate and pose increasingly more requirements to the supply market that are not easily satisfied. The supply market presenting its web pages to the demand market should find the best and quickest ways to respond promptly to the changes dictated by the demand market. The question is how to do that in the most efficient and quickest way. The data on the usage of web pages on a specific web site are recorded in a log file. The data in a log file are stochastic and unordered and require systematic monitoring, categorization, analyses, and weighing. From the data processed in this way, it is necessary to single out and sort the data by their importance that would be a basis for a continuous generation of dynamics/changes to the web site pages in line with the criterion chosen. To perform those tasks successfully, a new software solution is required. For that purpose, the authors have developed the first version of the WLE (WebLogExplorer) software solution, which is actually realization of web page multidimensionality and the web site as a whole. The WebLogExplorer enables statistical and semantic analysis of a log file and on the basis thereof, multidimensional control of the web page dynamics. The experimental part of the work was done within the web site of HTZ (Croatian National Tourist Board) being the main portal of the global tourist supply in the Republic of Croatia (on average, daily "log" consists of c. 600,000* sets, average size of log file is 127 Mb, and c. 7000-8000 daily visitors on the web site).

Keywords: *market, competition, log file, web page, dynamics, WLE-WebLogExplorer.*

1. INTRODUCTION

Global trends are increasingly more dynamic in terms of quicker and better-defined market demands the consequence of which is a demand dictating faster pace of changes on the supply side. Consequently, the supply has to find ways to respond to the demands of the market in the quickest possible way. However, there is still a question, how to do that in the best and quickest way. When a tourist city/town, tourist destination, or even the State publishes some advertising material, it is very difficult and almost impossible to publish the

new ones (their updates) in the same year or change their content because it is a time consuming and demanding process. Similarly, certain time should elapse, end of the season for instance, during which different professional and other organizations (institutes, faculties, etc.) carry out all kind of polls, surveys, and analyses, which again require a certain time to process them and to respond to the market supply in a methodological way. It is important to notice that a process as such requires quite some time from the "establishment of state/facts to the response". However, talking about the Internet and business running by the Internet, the entire issue should be approached in a completely different way. The purpose of this paper is to point out, at a scientific, theoretical, and practical level, that there is no Web site model that would allow permanent and direct retroactive effect on the web site along with a continuous statistical and semantic analysis of the log files. Monitoring and analyzing the amount of people coming to the site by means of the log file analysis make the response time shorter, give better quality and more systematic response to the market demand, the demand market requirements are continuously program monitored, and a continuous web site dynamics is ensured. With the work on the WLE software solution design, which can be implemented in already existing web site, the authors wanted to give a certain scientific and practical contribution to deepening the issue discussed.

2. WEBLOGEXPLORER (WLE) – SOFTWARE CHARACTERISTICS

Before determining and developing the module and algorithms of the WLE software, the criteria and requirements for development of WLE should be established, such as:

- recognition of a log file format to be processed,
- creation of an optional format if none from the program is suitable,
- the fastest possible processing of the log file data and its storage in a database,
- presentation of data to the user in the most intuitive way possible,
- possibility of creation of various analyses within a certain time period,
- relating those data to the dynamic change of the page,
- the simplest possible orders for a dynamic change generally given by a web programmer or design engineer,
- generation of various reports and a printout,
- selection of a software tool that could meet all or the most of above criteria,
- selection of a database

The basic characteristics of the program arise from the criteria and requirements defined before making the program:

- possibility to process the log files according to the formats: IIS server, Apache server or in the case of any other log file format, the format can be free defined according to the tags;
- recording in a centralized database;
- recording in a centralized database after data processing;
- possibility to get various analyses (statistical and semantic) within a certain time;

- a review and a printout of various reports;
- data filtering according to some special criteria;
- creation of dynamic changes that should happen on the pages selected;
- possibility to create a schedule according to which dynamic changes should take place;
- definition of aliases for optional addresses to facilitate the data representation on the screen or in the report;

2.1. ADVANTAGES, DRAWBACKS AND PROSPECTS

The specific purpose of WLE is permanent fulfilment of current requirements of demand market that is of end users. One of the main characteristics of WLE software is the possibility to perform as automatic as possible the dynamic changes in the web page content according to previously described rules. However, since WLE is still in the development stage, its advantages, drawbacks and prospects are discussed further in the paper.

2.1.1. Advantages

The advantages of the WLE software are supported by the facts noted below:

- Program application for log file processing with no changes in the Web site code meaning that we need to install the software, set its basic parameters and start the processing, which is in contrast to FLORID [2] model *(**F-LO**gic **R**easoning **I**n **D**atabases – a deductive object-oriented system of database using F-Logics as a language for data definition and as a Query Language)* whose object model should be installed in every web site page if we wish to get the data we need;
- Simple definition of optional formats if none of the formats suggested by the software is applicable;
- Log file processing and storing into a database every "n" lines thus reducing the possibility of error or the software "breakdown";
- Possibility to analyze the data within a certain period;
- Dynamic change of the content according to the analysis data within a certain time;

2.1.2. Drawbacks

The drawbacks in application of WLE software (which we shall try to remedy during our further work on the program development and its improvement) could be summed up as follows:

- Slightly lower accuracy of the data processed from the log file in contrast to the FLORID model, which has its object model installed in the web pages for which we wish to get certain analyses.
- Data is processed a bit later than in the case of the FLORID model, which immediately stores the data in the database from which it performs further analyses.

- Reduced accuracy of processing and analyzing if most users in a day log surfed on the web site from an institution in which case only one TCP/IP address (proxy) is listed thus making impossible to identify the number of people actually «surfing». This is also the main problem of all other similar software.
- Depending on the web servers it is not possible to include recording of the beginning or end of session for some extensions of pages (this is possible on ISS server with a certain extension or if the pages are *.asp or *.html) thus reducing the accuracy of the users' navigation maps and paths.
- It would be ideal if the dynamic change of pages could be made fully automatic.
- Some servers have the possibility to record the log files directly in the database with some program extension that should be purchased. In this case, the program could be adapted to that database thus making the data processing much faster (there are already some program extensions for IIS server).

2.1.3. Prospects

Briefly, the prospects of WLE software could be described as:

- Improvement in accuracy of data processing;
- Introduction of new analyses oriented to end user's interests;
- Improvement in dynamic changes of the page content according to the specific nature of the Web site for which the software is designed and/or adapted;

3. SOFTWARE DEVELOPMENT PROCESS FOR LOG FILE ANALYSIS

Before making the program, we selected MS Access database (because it was sufficient for experimental investigation process; for practical use for more complex models, we recommend MySQL, MSSQL, Oracle, PostreSQL) and the program language C# from Visual Studio .NET 2003 [4] that could meet our requirements and criteria. We have designed DBMS, created the tables, attributes and interrelations. Once we have chosen the software tool and the database into which we shall store the data from the log file, we made the basic classes for connecting the program and the database. The class for the database consists mostly of functions for storing and reading the data. Also the class for as accurate processing of log as possible has been made taking into account the criteria set. To create an optional log file format, a specification of tags has been introduced as a support to the WLE software:

- %S server name
- %d day
- %M month
- %Y year
- %h hour
- %m minute
- %s second
- %E method (GET, PUT, POST)

- %e URL path
- %q URL query
- %c server status
- %f host name
- %B browser and OS
- %j any other symbol (space and the like)

Before arranging the tags, one should know the log file format. In addition, a graphical interface has been designed as well as other parts of the program to help the user with the program application. After that, it was necessary to create the functions to get necessary statistical analyses within a certain time. To get a semantic analysis, we developed a class that enables sorting the data processed by servers and defining the navigation paths by which a user visited certain web pages within a web site.

Above was followed by creation of a window and the logics for dynamic change of pages. To define the rules following which the dynamics will take place, we used the data stored and a "substitute code" to be entered to change the page content.

4. ANALYSIS OF STRUCTURED DATA AND EFFECT ON THE WEB SITE

Talking about the analysis of structured data and the retroactive effect on the Web site [1], [3] means talking also about the operability of WLE whose tasks are:

- processing of the current state of visit of a web site and
- changing the web site (retroactive effect)

Processing of the current state is carried out according to the data collected within the period the users' interactions with the web site took place. This data is stored in the web server log files, whose record can take place in certain time intervals. By its format it differs from one server to another. The usual intervals are hour, day, week, month, or according to the file size.

The basic log file format according to W3C standard [6] is:

- date and time
- TCP/IP user address
- server name
- method
- web address (link)
- web query (link query)
- server status
- users' host
- (browser and operating system)

Change of a web site covers the changes of the content of specific web site pages in line with the current state of users' interactions with the web site. Decisions on modifications are based on the criteria such as web page visiting rates, server's frequency (users'), frequency of navigation parts and the like. This data are obtained from the program mentioned and the database. A qualified person without any specialized programs for dynamic changes can make the changes in the content, or the process could be partially automated by application of such a program. Full automation is restricted by recognition of different web languages and the logic used to "write" the web pages. Since we wish to perform the dynamics on already existing web sites, the web programmers should properly understand the code and find the places for possible changes. To avoid that, the changes in the page and in the code are performed by specialized **WLE software** [5].

5. WLE – DESCRIPTION OF SOFTWARE COMPONENTS

The software is divided into several wholes:

- Main software (Figure 1)
- Report overview and printing
- Parameter setting
- Dynamic change

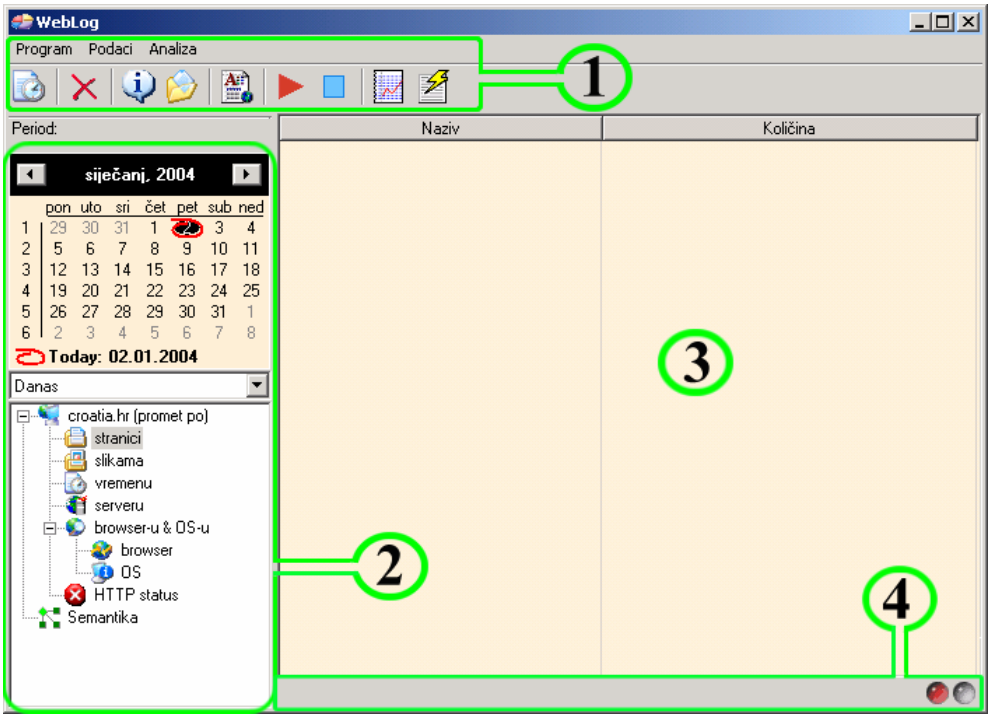

Figure 1: WLE – Main software (interface)

5.1. WLE Software – main part

The main part of the software consists of:

1. Menu and toolbar with icons
2. Left panel (containing a calendar, filter and control with the areas that can be analyzed)
3. Right panel (check list for data presentation)
4. Status bar (showing if data processing is initialized)

Because of the restricted size of this paper we have not given a detailed description of WLE software, which would considerable exceed its framework and the size.

6. WLE ANALYSIS AND GENERATION OF DYNAMICS

Since the log files obtained from HTZ (Croatian National Tourist Board) for the analysis were stored as *.zip files, to be able to use them and to subject them to a complete analysis process by WLE software modules, we have to decompress them and transfer them to a folder defined by WLE software for off line analysis (Figure 2). All decompressed *.zip files have the same default name (www.croatia.hr-access) so it was necessary to change the names according to the time they refer to (in this particular case, we determined a day), and only after that transfer them to the folder intended for processing. When all previously noted conditions are met, it is possible to perform WLE stochastic analysis of the web site usage, *on HTZ web site in this particular case*.

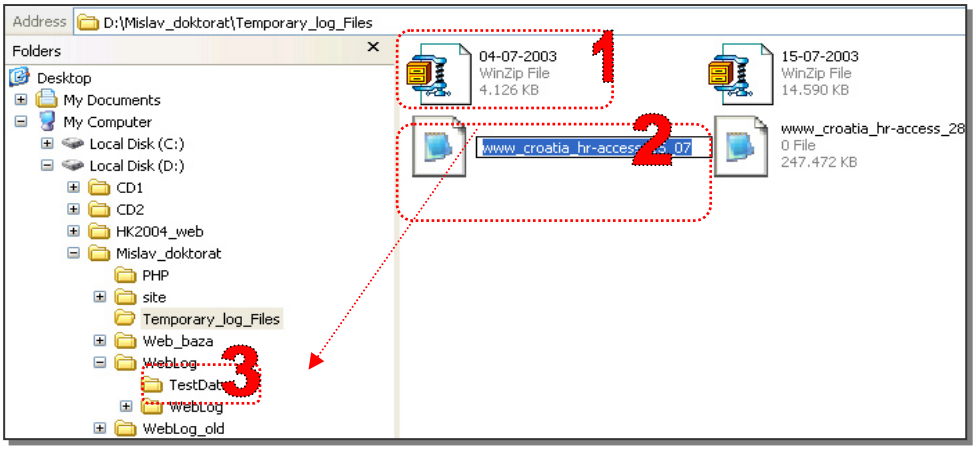

Figure 2: Data preparation (log files) for analysis

At the beginning of the analysis process, the log file should be parsed, providing it is previously decompressed, renamed, and placed in WLE software defined folder for processing. The initialization of the analytic process and the log file processing itself are illustrated on Figures 3 and 4.

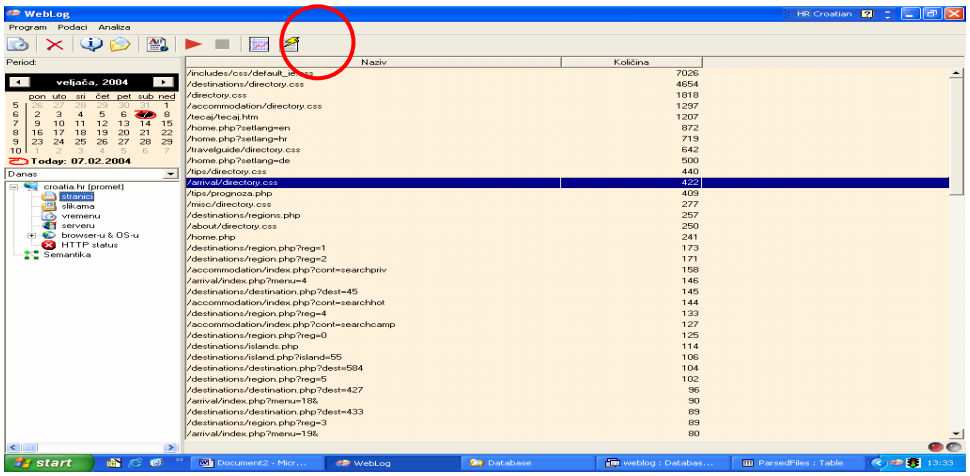

Figure 3: Initialization of analytic process – LOG FILE PARSING

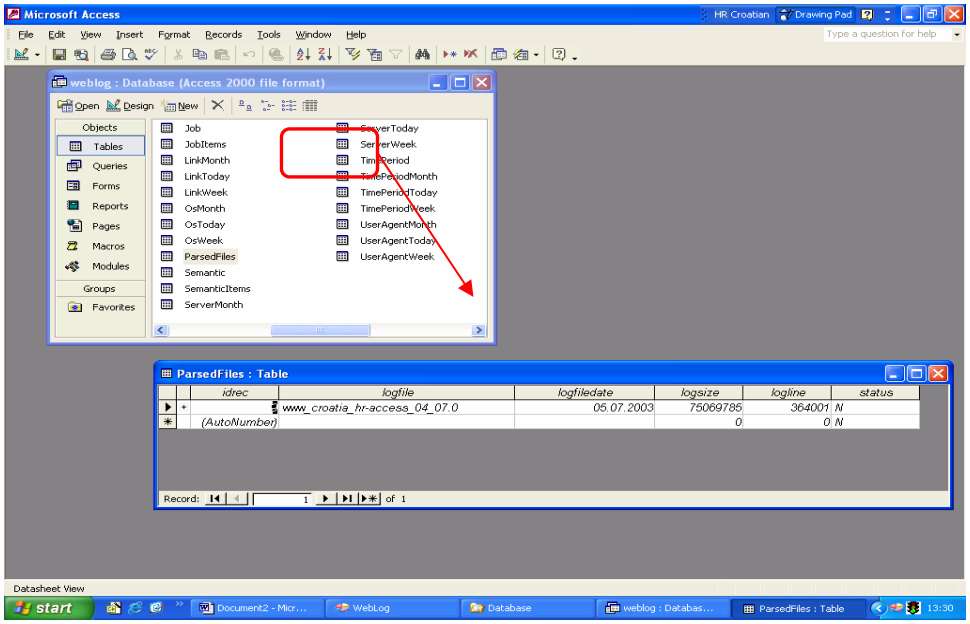

Figure 4: Log file processing by WLE (MS ACCESS Database)

When the log file parsing is completed, we get a statistical and semantic data representation of log file records (Figure 5), which is used for dynamic changes to the web site.

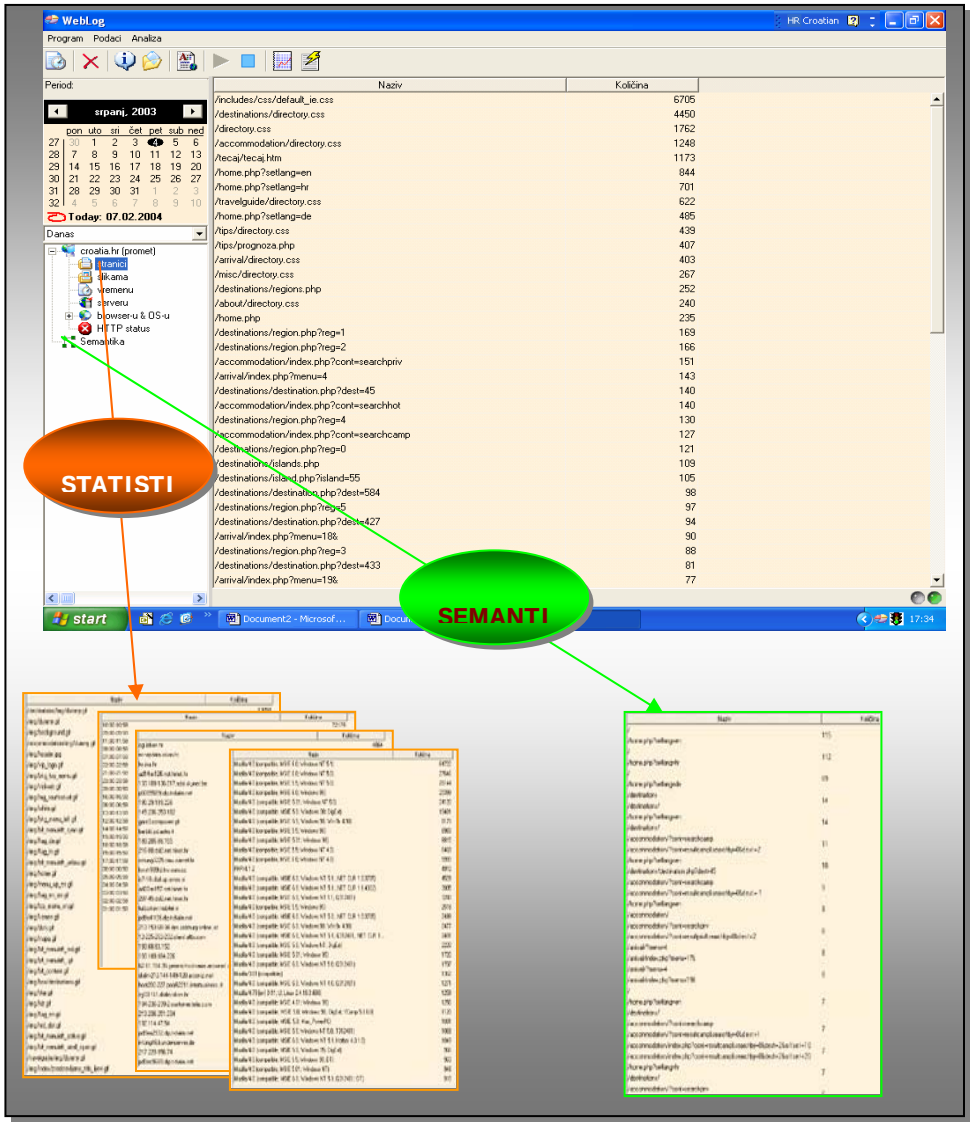

Figure 5: Area of basic events – Statistical and semantic data representation obtained by analyzingthe log file records

For easy understanding, below is a description of a simple example of a dynamic change to the web site based on the data obtained by a statistical and semantic analysis of log files by WLE. Figure 6 illustrates the status of HTZ web site home web (home.php), which will undergo, after the dynamic change, some corrections/changes required by statistical-semantic WLE analysis and generated dynamics.

At this moment, it is important to note the status (order) of the items in Dolazak menu. Below is showed the process of giving commands for generation of changes to the web page selected and the web page look after the commands for dynamic changes.

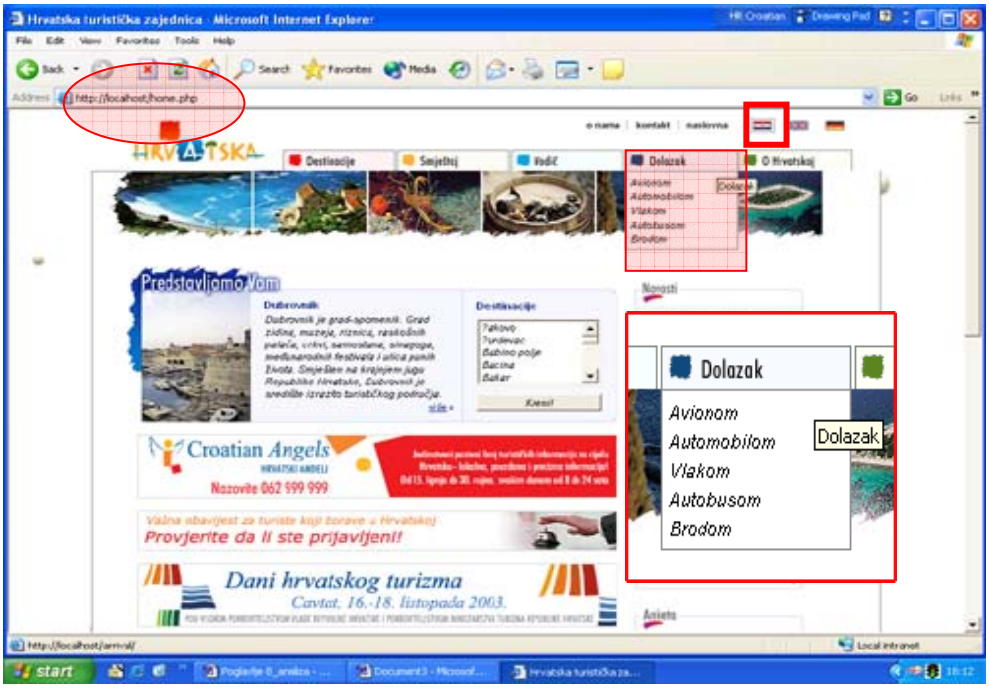

Figure 6: Web page before the dynamic change

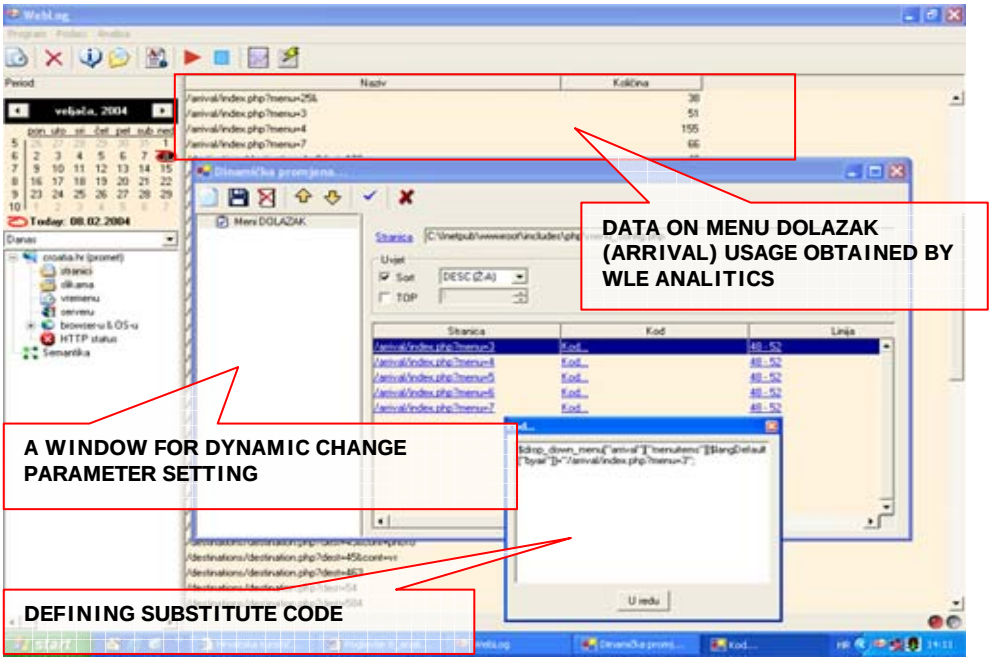

Figure 7: Parameter setting for performing dynamic changes on HTZ web site on the Web page menu_config.php (based on statistical analysis)

After adjustment of parameters for performance of the dynamic changes, the time to perform the change is defined in the main window of WLE software. At the time set, WLE starts the process of dynamic changes and in this particular example the items of the Menu DOLAZAK are rearranged (Figure 8) according to the selected parameter of WLE analytics (Figure 5).

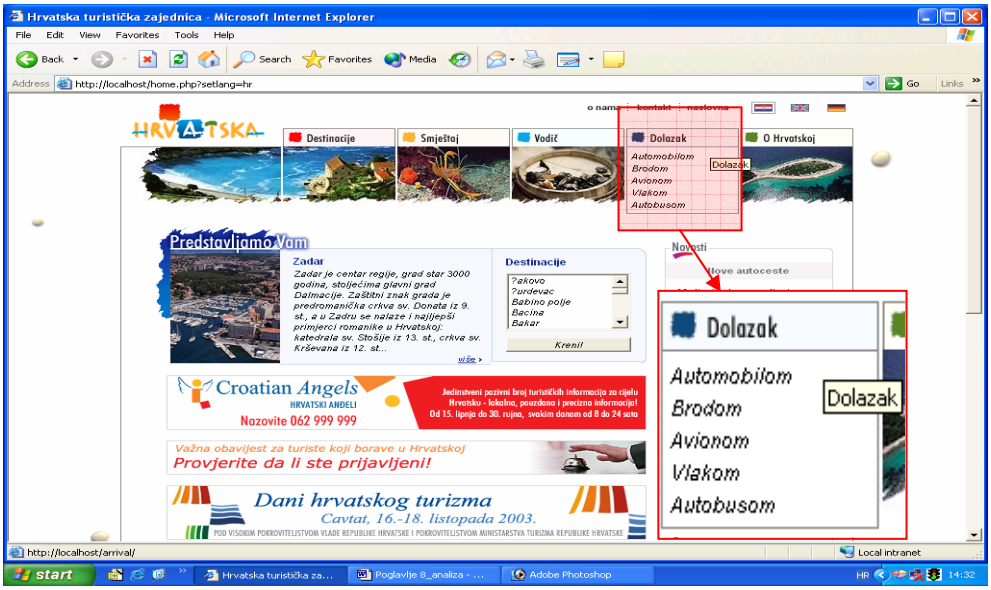

Figure 8: Web site home.php after dynamic change

Above figure (Figure 8) shows a place on the web page subject to WLE dynamic change according to currently relevant parameters. As already noted in the chapter describing in detail the WLE software, the dynamics can be controlled according to the current interest and adapted to the current demand market changes permanently recorded in the log file. In this example, the web page statistics was a basis for defining the changes. A basis for generating the changes can be any sub-analysis described under the WLE Software analysis description.

CONCLUSION

To enable successful performance of any business system and continuous generation of benefits for the supply as a whole and for the supply segments, any company or a business system should use the world trends and advantages of present-day technologies in the best possible way. This paper discusses and explains the application of a new approach to generation of permanent dynamics by implementation of WLE software as an integral part of a web site. Below are basic and concrete conclusions:

- The purpose of the WLE software at this stage was experimental, so for its commercial uses it requires improvements and algorithms that are more sophisticated. This particularly refers to the part involving the mathematical model and analytic algorithm that would enable better connection from the statistical and semantic analysis to the web site.

- WLE is particularly acceptable for transient stages of web sites until a certain business system decides on designing a new web site within which its own object-oriented model is created and incorporated in the web site during its design stage. However, the parer does not address this issue in detail.
- The preliminary analyses and pilot investigation of the WLE software included also *.css files. However, because of their large and dominant frequency of events (which is understandable) which is irrelevant, *.css files can be excluded when implementing the WLE software because they are not relevant indicators in the analysis (they have been excluded from the current version of WLE program WLE 1.0). In most cases, the use of htm, html, asp, and php files is sufficient.
- With regard to the web site size (number of files), technology, the language in which the web site pages were made, it is necessary, after the implementation of WLE program, to define the parameters on which basis the analyses will be made because after data parsing they are a basis for dynamic changes to the web site.
- Special attention should be paid to defining the items that represent the basis of dynamic changes. The dynamic changes themselves, writing and setting of algorithms for their execution, and regularities of their execution should result in cooperation among the experts from the business-problem solving area, informatics, psychology, mathematics, etc. In effect, the cooperation would facilitate development of more sophisticated WLE or similar versions that would base on algorithms of higher sophistication and the results of such an approach would for sure have consequences in better performance of any business system.

REFERENCES

[1] E.G.Abels, M.D.White, K.Hahn: "*A user based design process for Web Sites*", Internet Research: Electronic Networking Applications and Policy, vol. 8, n.1, 1998, p. 39-48

[2] H.Davulcu, J.Freire, M.Kifer, I.V.Ramakrishnan.: "*A layered Architecture for Quering Dynamic Web Content*", Bell Laboratories

[3] H.Davulcu, J.Freire, M.Kifer, I.V.Ramakrishnan:"*A layered Architecture for Querying Dynamic Web Content*", Bell Laboratories

[4] Jane Greenberg, Stuart Sutton, D Grant Campbell*: « Metadata: A fundamental component of the Semantic Web»,* Bulletin of the American Society for Information Science and Technology. Silver Spring: Apr/May 2003. Vol. 29, Iss. 4;

[5] W.May: "*Information Extraction from the web with Florid*", Dagstuhl Seminar, Declarative Data on the Web, September 1999

[6] W. May:"*An Integrated Architecture for Exploring, Wrapping, Mediating, and Restructuring Information from the Web*", ADC 2000, Canberra, Australia, 2000

[7] B. R. Preiss: "*Data Structures and Algorithms with Object-Oriented Design Patterns in C++*", University of Waterloo, Waterloo, Department of Electrical and Computer Engineering, Canada

[8] M.Stiefel, R.J.Oberg: «*Application development using C# and .NET*», Prentice Hall PTR, NJ 07458, 2002

[9] M. Randić, M. Šimunić, P.Knežević: "*Modelling structure and dynamic behaviour of Web pages augmented with dynamic HTML*", DAAAM Symposium, Vienna, 1998. p. 413-414

[10] M. Šimunić: "*A Model of Double Dynamic Web Pages*", Faculty of Organization and Informatics, Varaždin, 2004 – Ph Theses

[11] World Wide Web Consortium (W3C) - http://www.w3.org

Received: 06 June 2005
Accepted: 08 December 2005