UDC:004.738.52
Original scientific paper

LOG FILE ANALYSIS AND CREATION OF MORE INTELLIGENT WEB SITES

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

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

Abstract: *To enable successful performance of any company or business system, both in the world and in the Republic of Croatia, among many problems relating to its operations and particularly to maximum utilization and efficiency of the Internet as a media for running business (especially in terms of marketing), they should make the best possible use of the present-day global trends and advantages of sophisticated technologies and approaches to running a business. Bearing in mind the fact of daily increasing competition and more demanding market, this paper addresses certain scientific and practical contribution to continuous analysis of demand market and adaptation thereto by analyzing the log files and by retroactive effect on the web site. A log file is a carrier of numerous data and indicators that should be used in the best possible way to improve the entire business operations of a company. However, this is not always simple and easy. The web sites differ in size, purpose, and technology used for designing them. For this very reason, the analytic analysis frameworks should be such that can cover any web site and at the same time leave some space for analyzing and investigating the specific characteristic of each web site and provide for its dynamics by analyzing the log file records. Those considerations were a basis for this paper.*

Keywords: *log file analysis, dinamic web site, intelligent web site*

1. LOG FILE CHARACTERISTICS AND IMPORTANCE

The basic characteristic of a log file is a stochastic and unordered record created by the web server containing the data on the web site visitors, who and when visited the web site, and which pages were visited (see Fig. 1). However, the use and analysis of a log file record could be at a much higher level.

Generally, any serious *web host (web service provider in a broader sense)* gives its server logs to Webmasters of certain web sites to analyze their traffic (*or they give already processed files i.e. complete statistics*). Many programs use the log files to analyze the traffic, but there are also aspects of considering the analysis issue in view of raising it to a higher level. What does that mean? The answer to that question should be given throughout

this paper but, to put it briefly, a log file should be used as much as possible for deepening **its analysis in terms of statistics and semantics and for providing for a program controllable web site, which is not an easy task [2], [3]** . In fact, almost all data file analyses come down to the use of current analytic programs (WebTrends, Webalizer) as illustrated in Figure 2, the result of which is daily detailed graphical and tabulated listing of all the events on the server.

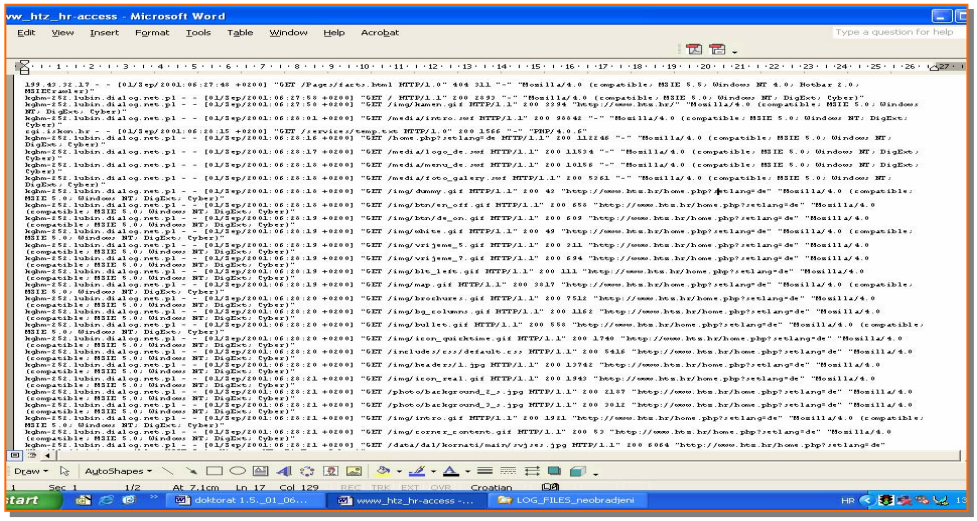


Figure 1: Stochastic and unordered record of a data file

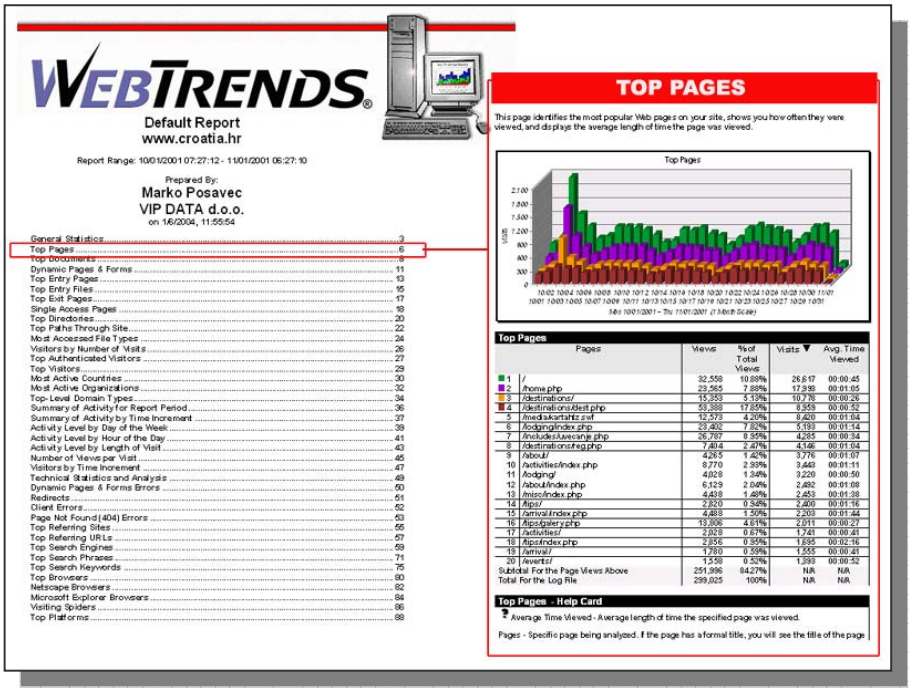


Figure 2: An example of a processed log file (WebTrends Statistics)
Source: "Web trends default report" completed by the authors

However, in most cases this is not the end of the story. Why is it so? We shall try to answer through the following facts:

- Web analyses are very extensive.
- Web analyses are narrow-profiled and adapted to the profession i.e. to the web site type.
- From day to day the world of business is getting more and more dynamic, managers do not have time to spend a few hours every day just for analyzing statistical reports to decide what is next to be done with the web site.
- Initially, the idea of «Web analyses» was good, but with time and development of technology, the analyses have become insufficient.

2. LOG FILE AS A DATA SOURCE AND A BASIS FOR ANALYSIS

Digital environment allows for application of new methods for monitoring and analyzing the usage of digital sources. With adequate software support, any activity or transaction with the server computer system could be automatically analyzed, so-called **log file analysis**. Originally, a log file analysis was intended for server administrators to monitor the system load, but with time, also analysts of different profiles and needs started using it. Basic elements for analyzing the log file contain the answers to the following questions:

- **Who?** – User identification; in most cases IP address of PC from which the user accesses is identified. When the access is controlled by the user's account (username and password), the user and its activities can be personally identified.
- **When?** – Date and time of access;
- **What?** – Identification of web pages used and the method of using the material (*statistical analysis*). There is software to perform standard monitoring but for the needs of this paper, the software is improved and raised to a higher level.
- **How?** – Identification of the way information i.e. web page is used (*semantic analysis*). To perform monitoring, additional and improved software is required.

When collecting the data, the deficiencies and imperfections should not be overlooked but pointed at and this will be done further in this paper:

- The user's activities are recorded by the time the data is received from the database, so we do not know in what way it is further used e.g. do the users read it directly on the screen?
- Data search in the cache memory could be kept for a few days. If within that time the user looks again for the same content, it will be provided from the cache memory and the server will not record that access.
- Do the requests registered come from one or more persons? It is not always possible to see if multiple requests for the same document come from one or more users. Individual user cannot be identified by means of an IP address if there is no fixed connection to the Internet. Even if there is a fixed connection, the access by the same person cannot be systematically monitored if the access is made from more PCs. Similarly, more people can use the same computer

meaning that for different enquiries, the same IP address is recorded and again the data we get on the usage is inaccurate. It should be mentioned that the usage monitoring is possible when a log in is required for using the web site i.e. username and password should be entered.

Consequently, it is suggested that there are realistic theoretic and practical bases for more in-depth improvement of the model design issue aimed at constantly meeting the users' interests to the maximum possible extent [5].

3. STOCHASTIC DATA WITHIN A LOG FILE

Log file as a text (txt) ASCII file is a computer i.e. data representation of a web site use process by a certain number of users within a time interval monitored. The representation changes quickly meaning that we can characterize it as a dynamic change. Dependent on the size and "popularity" of a Web site, the change dynamics can be faster or slower which, of course depends on the number of users surfing on the Web site, its size, user's activities on the Web site, and the time users spend on searching for information about the Web site. Any activity of any user is registered in a log file. No future state of the records in the log file can be predicted upfront because no future interaction of users with the Web site can be predicted either. Therefore, we find the log file almost fully stochastic. Similarly, the log file records are not sorted by any structure already defined. As regards their format, they differ from one server to another. The record in the log file is very much changeable and it is always monitored and stored according to predefined time intervals (hour, day, week, month, etc.), which is generally determined by the web site size and the surfing intensity. The consequence of it is also the very size of the log file and the dynamics of changes taking place in it. To process previously noted records-data (Figure of unordered log file), we need a program that will recognize the log file format, read all necessary data, structure them and store them in an adequate database – the log can be entered directly in the database. All data collected in this way is used for further analyses, which gave rise to making this paper.

4. LOG FILE AS A BASIS FOR PROVIDING FOR WEB SITE DYNAMICS

As previously noted, a log file is a source of stochastic data. Nowadays, there are programs (Webalizer, WebTrends, etc.) intended for analyzing stochastic data collected in a log file. However, the analyses such as these although often intrinsically good are just not efficient enough and in most cases, they are just a set of tables and graphs that remain unused in practice. This was why the author's thoughts were focused on improving the usability of the log file analysis for realization of a program controllable web site. The idea of developing a program controllable web site [7] is illustrated on Figure 3.

As showed on the figure above the novelty is **WLE GUI** (WebLogExplorer Graphical User Interface), an integrated software solution providing both for the data analysis from a log file and the effect on the web site that is the web site dynamics. Basic components of the WLE software solution are a module for data analysis and a module for ensuring the dynamics within the web site. By implementation of the WLE software solution, although currently in an experimental stage, the dynamics within the HTZ (Croatian National Tourist Board) web site has been successfully performed, as showed further in the paper.

Figure 3: Program controllable web site

5. WLE GUI – A SOFTWARE SOLUTION FOR WEB SITE DYNAMICS

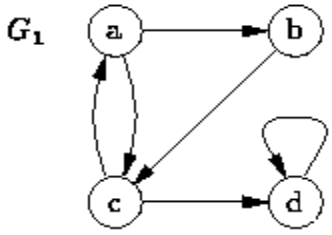
The paper further discusses and describes the role of WLE as a software solution for ensuring the web site dynamics. The WLE software has been developed by a theory of graphs [4]. In fact, the theory of graphs is a mathematical discipline studying the regularity of (mathematical) graphs. The graphs are mathematical structures for which we can say they consist of vertices and edges connecting them. More precisely, among many regularities and definitions the one below can be singled out:

Oriented graph is an ordered pair $G = (V, E)$ of below characteristics:

The first component V is a final non-empty set. The set elements are called the graph vertices.

The second component E is a finite set of ordered pairs of vertices, i.e. $E \subset V \times V$. The elements of that set are called the graph edges.

An example of the graph is given on Figure 5-3. In the graph $G_1 = (V_1, E_1)$ is $V_1 = \{a, b, c, d\}$ and $E_1 = \{(a,b),(a,c),(b,c),(c,a),(c,d),(d,d)\}$.



Source: 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

Figure 4: Oriented Graph

For presenting the navigation paths, we used an oriented graph like the one showed above: the graph vertices are the web sites, the oriented edges are the links from one web site to another, and the vertices labels resemble the symbol of the navigation path to which a vertice belongs. The oriented graphs we used are somewhat more complex than the usual ones: any vertice can belong to none or to a larger number of navigation paths i.e. every vertice can have no slabels or several labels.

6. WLE CHARACTERISTICS (WEBLOGEXPLORER) – LOG FILE ANALYSIS PROGRAM

Before the characteristics of the WLE software were developed and defined, it was necessary to set the criteria and requirements 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 that 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 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;

However, 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 [6]** 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 design 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 of data processed by servers and defining the navigation paths by which the 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. With that, we have fulfilled the concept of permanent dynamics within a web site based on the user's activities.

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

When talking about the analysis of structured data and the retroactive effect on the Web site, we actually start talking about the web site dynamics issue involving the following tasks:

- Processing of the current state of visiting a web site and
- changing the web site (dynamics i.e. 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 log files of web server, 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 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)

Some servers have the possibility to records some other additional parameters such as the beginning and end of user's session, amount of pages received and sent in bytes, etc. Order of the format parameters depends on the web server (Figure 5 - Parameters on ISS Server).

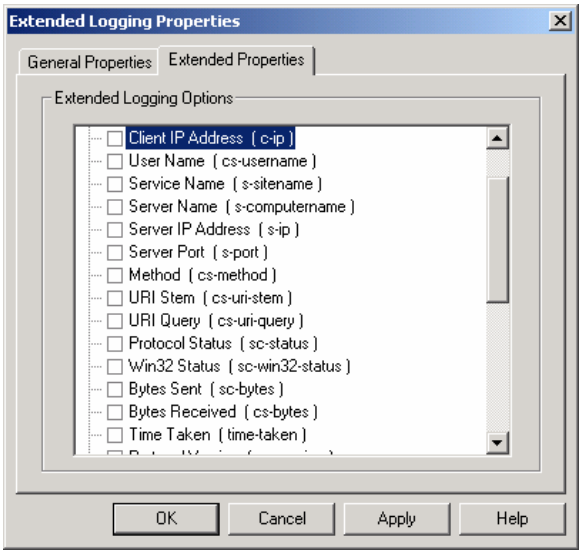


Figure 5: Parameters on IIS Server

To process the data mentioned above, a program is required that will recognize the log file format, read all necessary data, and store it in an appropriate database. All the data collected in this way is used for further analyses on which the web site dynamics is based on.

Change of a web site covers the changes in 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 is obtained from the program mentioned and the database. A qualified person without any specialized programs for dynamic changes could 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**.

8. COMPARISON WITH SIMILAR PREVIOUS RESEARCH (FLORID)

The Web provides access to large data sources which are not explicitly organized as databases. Instead, the information is presented as *semistructured* data. In contrast to integrating classical distributed databases, integration of such data raises several new problems such as schema discovery, wrapping and reorganizing the data sources and coping which changes in autonomous sources. FLORID project system has been used for extraction and integration of semistructured data from the Web.

In 1997-1999, the FLORID system has been extended with Web access capabilities (Versions 2.x). A methodology for wrapping and integrating HTML pages by mapping the information in an integrated F-Logic data model representing both the structure of the data sources and containing an application/level model of the information has been developed. HTML pages are wrapped using generic rules for the usual structuring means (i.e., lists, tables, comma-lists, emphasized keywords). In 2000, FLORID has been extended to FloXML with special functionality for handling XML data (XML/DTD parsing, metadata provided by DTDs and XMLSchema, XML export functionality).

FLORID (*F-LO*gic *R*easoning *I*n *D*atabases) is a deductive object-oriented database system employing F-Logic as data definition and query language. The development was supported by the Deutsche Forschungsgemeinschaft (project La 598/3-2). With the increasing interest in *semistructured data*, Florid has been extended for handling semistructured data in the context of Information Integration from the Web. The Experiences with F-Logic and FLORID have been continued with the LoPiX project for XML. Languages supporting deduction and object-orientation seem particularly promising for querying and reasoning about structure and contents of the Web, and for the integration of information from heterogeneous sources. FLORID, an implementation of the deductive object-oriented language F-logic, has been extended to provide a declarative semantics for querying the Web. This extension allows extraction and restructuring of data from the Web and a seamless integration with local data. Since the functionality of wrappers and mediators is integrated into a single declarative language, the development of advanced applications based on the Web as an information source is significantly simplified.

The folloeing goals served as a main motivation for the research in the FLORID:

Querying the Web

- (express declaratively how to query navigate on the Web, extract data from Web pages for populating a database (Web data warehousing)

Management of Semistructured Data

- structure is irregular, partial, unknown, implicit in the data
- example: HTML pages
- querying/navigation using general path expressions
- discover structure

Information Integration

- heterogeneous sources with different structure, wrappers, mediators [3].

Some ideas and knowledge from FLORID project and F-logic has been implemented in our research and integrated in the WLE program algoritms [7].

9. EXAMPLE: WLE IMPLEMENTATION ON HTZ WEB SITE

The role of WLE software is multidimensional in terms of recognition of the log file format, its processing allowing different analyses in certain time intervals, and finally giving the order for dynamic change of the web page within a web site.

The first task of the WLE software is to recognize the log file format and its processing. As noted on several occasions earlier in the paper, the log file content is very unordered and stochastic. Similarly, it is very unlikely that its content is repeated twice or more times implying that each time the WLE software is started different results will be obtained i.e. a review of the status and activities on the market demand side reviewed by the log file.

Consequently, a research analysis performed in the context of such consideration would mean to subject a certain log file to the entire WLE process. For the needs of this paper, we have collected the data of the log file of HTZ Web site from July 2003 on (31 off log files). An average daily "log" is c. 600,000 sets, which is a large amount of data and reflects a representative picture of the activities i.e. demand market interests. Because of the daily log size, among the available data (log file), a random selected log file will be processed and parsed to show the stochastic characteristics of the data within the log files. Having obtained the data in this way, the file selected will be subject to the entire WLE process. When the log file selected is decompressed from *.zip file, and moved to the WLE software module defined directory for off line analyses, a parsing process of the same log file should be run to obtain a statistical and semantic data representation of log file records, which is then used to define and run dynamic changes to the web site. Because of the restricted size of this paper, only one simple example of defining dynamics within a web site based on a log file analysis will be given to justify the essence of the issue discussed in the paper. The figure below illustrates the status of HTZ home web page (http://localhost/home.php). Attention should be given to the order of items in the menu "Dolazak" ("Arrival")

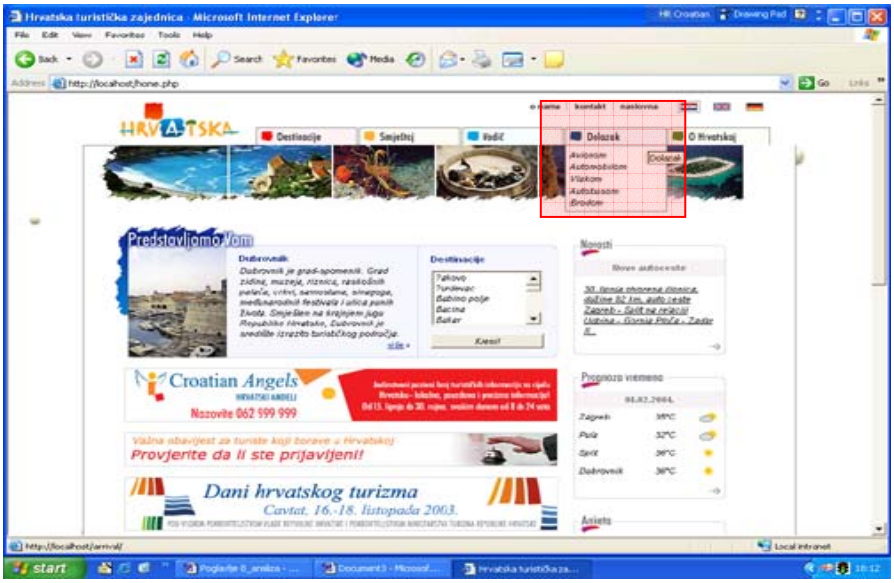


Figure 6: Web page http://localhost/home.php before dynamic changes

We have chosen for the analysis a log file of 4 July 2003 having the characteristics given in the table below:

Table 1: Characteristics of the log file for analysis (4 July 2003) – L4

	Characteristics of parsed log file **(L4)**				
Item	**Analytic process**	**logfile***	**logfiledate***	**logsize***	**logline***
1	**L4**	www_croatia_hr-access_04_07_2003.0	05.07.2003	76 770 373 bytes	364 459

After statistical analysis of the log file selected, below values of the parameters frequency relating to the use of items from the menu "Dolazak" (arrival) have been obtained:

Table 2: L4 –Internal relation of the menu items "DOLAZAK" obtained by the analysis

L4: Internal relation of the menu items DOLAZAK (100 d.n.r)*		
Item	**Web Page Name**	**Quantity**
1	/arrival/index.php?menu=3	51
2	/arrival/index.php?menu=4	155
4	/arrival/index.php?menu=5	0
5	/arrival/index.php?menu=6	0
3	/arrival/index.php?menu=7	66
	TOTAL	**272**

*100 top ranged events

When on the basis of the data obtained the parameters are defined within the WLE software in order to define the dynamic changes to the web site, the layout of the items in the Menu "Dolazak" on the web page http://localhost/home.php looks as follows:

Figure 7: **Web page http://localhost/home.php, after dynamic change**

10. CONCLUSIONS

A log file is an important source of data whose basic characteristics reveal a stochastic and unpredictable record created by the web server and contains the data on the web page use processes within a web site. Worldwide, there are software packages for analyzing log files (WebTrends, Webalizer, etc.) and the methods for monitoring log file records (Clickstream). However, their main feature is generating reports. The authors wanted to make a scientific, theoretic, and practical contribution to development of a new more sophisticated software solution (WLE) providing a program controllable web site i.e. web page dynamics based on the users' activities on the web site thus increasing the log file usability to a much higher level.

Based on the facts presented, it is suggested that there is every reason for defining the controllable dynamics web sites including also further development of the WLE software. The authors are fully aware of many shortcomings of the WLE software but for the purpose of this paper, the intention was to explain the standpoint and define the frameworks within which the WLE exists and functions. For the business systems, the acceptance of the idea of program controllable dynamics of web sites based on the users' activities would mean quicker and better global communication and exchange of information by the Internet thus considerably contributing to raising the company's performance and competition to a higher level. Such and similar segments of business operations are vital for the prosperity of companies, business systems, and the country's economy as a whole.

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

[2] 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, Volume 29, Issue 4

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

[4] 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

[5] M. Randić, M. Šimunić, P. Knežević: "*Modelling Structure and Dynamic Behaviour of Web Pages Augmented with Dynamic HTML*", DAAAM Symposium, Vienna, 1998. pp. 413-414.

[6] M. Stiefel, R.J. Oberg: "*Application Development Using C# and .NET*", Prentice Hall PTR, NJ 07458, 2002

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

Received: 06 June 2005
Accepted: 05 December 2005