The new public services for businesses in the tourism industry based on data from the tourism websites

mr.sc. Kristian Dokic, Polytechnic in Pozega, Croatia
kdjokic@vup.hr

dr.sc. Mirela Mezak Matijevic, Polytechnic in Pozega, Croatia
mmezakmatijevic@vup.hr

mr.sc. Antun Marinac, Polytechnic in Pozega, Croatia
amarinac@vup.hr

mr.sc. Zlatko Loncar, HEP - ODS d.o.o, Croatia
info@gdjenamore.com

ABSTRACT:
The importance of Web 2.0 technologies has been recognized in tourism and today the most popular global tourist portals offer registered users the ability to rate and comment on the services and destinations. On these portals, destinations and services are categorized, and the portals have a full set of data on each user, whether they are registered or not. Data collected on the basis of visited parts of portals can be used by the advanced marketing portals themselves, but this data is valuable for tourism entrepreneurs as well, and they could significantly improve their services on the basis of the knowledge gained from these data. An example would be a comparative analysis of the number of Outdoor Activities category in all cities with over 100,000 inhabitants compared to other offered categories, with an overview of visitors’ country of residence. Based on the information provided in this example, entrepreneurs in tourism would have a clearer picture of the interest and potential return on investment in Outdoor Activities in a particular area. Standard data mining methods are used to obtain this data, and the author's suggestion is that public institutions systematically collect these data from major portals and allow public access to them. In addition, it is necessary to systematically train the use of such information system by entrepreneurs in tourism.

This paper presents a Central European country (Croatia) that realizes a large proportion of its gross national income from tourism and public services related to tourism in that country that are available through public information services. In addition to this analysis, a regional portal (www.gdjenamore.com) where apartment owners in that country advertise, while the portal itself, its interface and available information are available in eight European languages (Croatian, German, English, Hungarian, Czech, Slovak, Italian and Polish), and thus targets a population of several hundred million users. Based on the data obtained from the mentioned portal, a series of information on users of the same is potentially useful for the entrepreneurs in tourism of the region and will be presented graphically.
Because of the limited geographic area, number and categories of services that are being advertised on the portal, the amount of knowledge gained is not significant, but with a greater amount of data found on global tourism portals, the amount of knowledge gained would be considerably higher.

**KEYWORDS**: log files, data mining, information systems in tourism

1. **Introduction**

Tourism in some countries of the world generates a relatively large percentage of gross national income, and public administrations of these countries are trying to help tourism professionals in a variety of ways. This assistance often comes through the implementation of various information systems. On the one hand, information systems are used to enable certain activities to be performed from a remote location, and also serve to inform or disseminate knowledge. This paper describes a public system that allows tourist registration by tourism professionals using the Internet and data from that system is used to inform tourism professionals in order to improve their services. After an overview of this system, further development opportunities are highlighted through public-private partnerships with tourist portals that in the LOG files of their web servers conceal knowledge that is currently unavailable to tourism workers.

2. **Tourism in Croatia**

Tourism is a globalization constant today. According to the definition: "Tourism is a journey for the purpose of recreation, pleasure or work and is usually of limited duration" (Source: [http://croatialink.com/wiki/Tourism:_definition,_nastanak,_razvoj_i_podjela](http://croatialink.com/wiki/Tourism:_definition,_nastanak,_razvoj_i_podjela)). According to the Tourism Development Strategy of the Republic of Croatia by 2020, the division of the type of tourism depends on the dominance of the product (the product is dominant if it contributes with at least 5% in the overall revenue structure) and products with a marked development perspective.

Dominant products in the Republic of Croatia are:
1. Sun and sea - The share of the sun and sea holds 85% of the physical stake for years, and participates with 75% to 80% in the overall revenue.
2. Nautical tourism - due to its natural assumptions (geotraffical position, indented coastline, pleasant climate, favourable winds), the Republic of Croatia is one of the most desirable tourist destinations.
3. Business tourism - lack of congress infrastructure and uncompetitive management system keep business tourism still in its infancy, with a share in the overall revenues of 10% to 15%.
4. Cultural tourism - in the last decade, much has been done in the field of cultural tourism, but in the near future it is necessary to work on commercialization and destination management in order to integrate cultural contents in the system in the right way.

Products with a distinct development perspective:
1. Health tourism - due to natural assumptions, the Republic of Croatia has comparative advantages for the development of health tourism.
2. Cyclotourism - although the Republic of Croatia is well connected to local and regional cycling trails, it is considered that cyclotourism is not sufficiently valued.

3. Gastronomy and oenology - although the benefits are consumed by all tourists, only a fraction of tourists travel exclusively for these reasons.

4. Rural and mountain tourism - rural and mountain tourism account for 3% in the overall travels.

5. Golf tourism - the Republic of Croatia today is not on the map of tourist golf offers due to scarce capacities. Namely, it has only four golf courses with 18 fields.

6. Adventure and sports tourism - despite the natural resources, this type of tourism has not been sufficiently developed.

7. Ecotourism - ecotourism in the Republic of Croatia is considered to be quite neglected, even in areas of protected natural sites.

3. **eVisitor system**

The importance of tourism is growing rapidly. According to the World Tourism Organization, it is considered that tourism will continue to grow at an annual rate of 4%, especially with the growth of electronic commerce. In order to facilitate the registration and check-out of tourists to the tourist offices, as well as the competent police station, since January 1, 2016 the eVisitor system is applied in the Republic of Croatia. Within 24 hours, legal and natural persons providing the overnight service are obliged to register to the tourist office of their municipality, city or their branches all persons using the overnight stay services. Likewise, they are obliged to check them out. eVisitor is the central electronic system for registering and checking out tourists in the Republic of Croatia (except for charter and cruising which will report data using electronic data exchange at the ministry responsible for maritime affairs). The eVisitor service is available as a web application over the Internet. The purpose of the system is the interconnection and accessibility of sojourn tax billing and both system providers and users have all the relevant data in real time. In addition, the eVisitor service is also important for calculating and controlling the sojourn tax collecting as well as for processing the statistical data. Figure 1 shows the appearance of eVisitor login interface.

![Login interface to the eVisitor system](image)

Since 2017, after the eVisitor system was in operation for one year, access to their databases was made available to the Central Bureau of Statistics and they had a much more accurate insight into a number of data related to tourists using private accommodation in the Republic of Croatia. Until then, the data was collected in other ways, including surveys, and were not entirely reliable. Monthly base
reports related to foreign tourist nights can be accessed on the website of the Central Bureau of Statistics of the Republic of Croatia, at https://www.dzs.hr/. Using eVisitor system, i.e. data stored in system databases, both the accuracy of information and knowledge that is interesting to tourism professionals have increased, but still there are number of improvements to be made. In addition to using data mining methods that would enrich reports and increase the amount of knowledge, another option is the use of other data sources that are also relevant to tourism. Below are suggested some of these new methods.

4. Features and applications of LOG files

After the eVisitor services were presented and the information available to the public sector in the Republic of Croatia which is related to tourism, the possibilities for further development will be presented below. The options described below are global in character, meaning they are applicable in all countries around the world.

The information available through the aforementioned statistical reports is useful to tourism professionals in a way that they can estimate the periods with most tourists, where the tourists come from etc. This information is a direct result of data processing from the eVisitor database. Figure 2 shows a client computer, a web server with the eVisitor services and SQL server pages where the database is located. Each client’s entry via the web site is recorded to the database on the SQL server.

![Figure 2 - Schematic display of data entry in the eVisitor database](image)

The basic function of a web server is to distribute web pages to interested customers on the Internet. One of the specifics of these systems is that all events and requests for web pages by clients are recorded in the so-called LOG files. These are most commonly used files for storing on web servers and they contain a number of information on clients that have requested a file or web site from a web server. LOG files differ depending on the web server, and Grace states that there are four types of LOG files:

a) Transfer Log  
b) Agent Log  
c) Error Log  
d) Referrer Log (Grace, 2011)

Katkar et al state that LOG files readily store the following information for each access to the web server: date of access, time of access, remote host IP address, remote login name, request, status code, bytes, referrer, and user agent et al., 2014). An example of a LOG file is presented in Figure 3.
Figure 3 shows the elements mentioned by the author Katkar et al., i.e., the date and time of access, IP address of the client who accessed the website, the address of the accessed web page, the user agent and some other information.

More authors have recognized the LOG files as a valuable source of information, so Azam et al. presented the use of the WEKA mining tool for analysing LOG web server files. The research analysed the LOG files from one server in India and came to the conclusion that there is a difference in the websites accessed by clients from India and clients from other countries. Clients from India generally accessed an introductory web site on a web server while clients from other countries directly accessed the web pages they were interested in. The authors argue that the reason for this difference is more frequent use of web search engines by clients outside of India (Azam et al., 2014).

Kumar et al. also use WEKA data mining tool for analysing LOG web server files in their work, but they use the so-called Apriori algorithm that enables them to predict client behaviour on that web server (Kumar et al., 2015).

Schegg with his associates also analyses the LOG web site files of 15 hotels and based on the information obtained finds the ways in which the websites can be upgraded (Schegg, 2005).

If we look at the information available in the LOG files, it is obvious that some of them are interesting to the tourism industry. The LOG files contain the client's IP address, and by using the databases that connect IP addresses and countries we can find out from which country the client is accessing a particular web site. The fact is that similar information is provided by the public sector institutions in the Republic of Croatia, as described above, but given that the so-called geolocation databases that link IP addresses and countries also include data on individual regions, tourism professionals can from the LOG files see from which region is an individual client accessing a particular web-page.

An important factor defining the amount of available information from the LOG files is certainly the amount of information available on the web server itself, i.e., more importantly, the way of classifying available information on the web server. One of the examples of a web portal whose LOG files are extremely "rich" in information is a specialized web search engine at www.tripadvisor.com. One of the specifics on this portal that contributes to the "richness" of information in the LOG files is that there is a categorization of attractions across the entire range of categories. Some of them are the following: Nature & Parks, Sights & Landmarks, Fun & Games, Museums, Outdoor Activities, Zoos & Aquariums, Nightlife, Boat Tours & Water Sports etc. In this way, the tourism professionals can get from the LOG files information which tourists are more interested in, for example, Nightlife, which are more interested in Nature & Parks etc. Figure 4 shows a home page of the mentioned portal and TripAdvisor specialized web search engine.
From the example above, it is clear that the amount of information on the web page itself increases the "wealth" of available information in the LOG files themselves.

5. An example of LOG file analysis

For the purpose of this paper, the LOG files from the regional portal and the specialized web search engine for private accommodation providers in the Republic of Croatia are analysed. The portal is located at [www.gdjenamore.com](http://www.gdjenamore.com), and the domain name itself in the Croatian language literally means "Where to go to a seaside". On this portal, approximately 1000 private providers offer accommodation and the portal is translated into eight languages spoken in the wider area. These are Croatian, English, Italian, Czech, German, Polish, French and Hungarian. The basic page layout of the portal is shown in Figure 5.
After uploading the LOG file, there is a range of data display options and Figure 7 shows countries from which the users viewed the portal web site on the day the data was logged in the LOG file. That date is May 1, 2017.

Considering the way the map structure of the analysed portal [www.gdienamore.com](http://www.gdienamore.com) is made, apart from basic reports, it is also possible to create more complex reports related to certain areas. Furthermore, since map structure is defined by the names of municipalities and cities, by viewing the individual map it is possible to obtain a range of information related to that geographic area. On Figure 8 we present a report generated using the filters in a way that only the portal users from the area of Germany who have accessed the portal between May 1, 2017 and May 10, 2017 are shown, and then the result was sorted by folder name, i.e. file name. In this way, we can obtain a clear view of how much interest has been expressed by German citizens for apartments in a given area.

The capabilities of the Web Log Storming tool allow a variety of queries and reports from LOG files, but that is still a program that does not include advanced data mining capabilities, thus some reports need to be done step by step. In addition, the lack of graphical data display and the creation of three-dimensional graphs further limits the amount of knowledge we can obtain from LOG files.
6. Conclusion

Knowledge stored in LOG files is relatively large and is still often inaccessible to tourism professionals. This paper presents a brief overview of the use of LOG files and the information stored in them. The program used does not allow a deeper analysis of data nor mining methods, but provides guidance as to how and what can be analysed. LOG files of large tourist portals are generally stored unused on hard disks of web providers, but through different forms of public-private partnership, this knowledge could become accessible to tourism professionals and thus enhance the quality of tourist services.

7. References


Ankit Kumar, Omkar singh, Vinay Rishiwal, Rakesh Kumar Dwivedi, Rakesh Kumar, (2015), Association Rule Mining On Web Logs For. Extracting Interesting Patterns Through. Weka Tool, International Journal Of Advanced Technology In Engineering And Science, Volume No 03, Special Issue No. 01, February 2015

Roland Schegg, Thomas Steiner, Thouraya Gherissi-Labben, Jamie Murphy, Using Log File Analysis and Website Assessment to Improve Hospitality Websites, Information and Communication Technologies in Tourism 2005, pp 566-576