A CONCEPTUAL MODEL OF A WEB BASED SYSTEM DESIGNED FOR DATA MINING WEB RESOURCES

Iva MAKEDONSKA

Abstract

This article discusses the disadvantages of the current process of data mining web resources in e-commerce and the reasonable circumstances for the application of intelligent methods. The proposed web-based system aims to overcome existing drawbacks in the process of web mining, bringing together the functions of three separate types of software into one. The key module in the system is a multi-agent system that performs local expertise and web mining based on data collected from various web resources. Using such a web system for analyzing a web resource provides the tools and knowledge necessary for the proper development of the e-store.

Introduction

In the recent years e-commerce has been gaining more popularity and has seen a remarkable growth in sales and purchases of goods. There has been increasing consumer interest in online shopping, which in turn leads to an increase in the number of e-stores and online transactions. To retain customers and optimize their market position, managers of e-shops must have means of intelligent analysis and forecasts. The process of analysis includes detection of models and links in data, stored in log files in the websites or applications as well as in operational databases. Data mining facilitates management solutions for personalized content, updates the services offered and introduces more qualitative structure of websites.

There are many software products that help owners of e-stores and automate some of the processes in the analysis of traffic to the websites. The main disadvantage of these programs is that making an adequate business decision imposes pooling the obtained knowledge and results of all the products while the use of conventional technologies does not provide it. By contemporary methods from the field of artificial intelligence solutions are offered to many of the IT problems in e-commerce. The implementation of web-based intelligent system ensures the integration of various

1 Department of Informatic, University of economics – Varna, Bulgaria. e-mail: i.makedonska@ue-varna.bg
information sources related to the analysis of traffic to the site, the extraction of data and creation of suggestions and forecasts for effective management of e-business.

The aim of this article is to propose a model of a web-based system with technologies from the field of artificial intelligence that compensate for the shortcomings of the existing process of web mining from web resources.

1. Existing solutions for web mining from web resources

Analysis, and more specifically the analysis of web resources, occupies a strategic place in e-business. According to the Gartner’s report (Parenteau, 2015) "Every business is analytical, every business process is analytical and every business user is an analyst." According to this report new platforms for analysis and data mining are required, to allow non-specialists to be able to work by themselves and manage a wide range of analytical processes. The sample characteristics that such a system must contain are:

- Self-service: an opportunity for the users to create their own queries and analyze data based on their needs;
- Flexibility and expandability: the possibility of adding new data sources;
- Orientation: fast and intuitive orientation of the users in the environment when creating reports and various visualizations;
- Connectivity: access to different types of data, such as log files, social networks, files, and more.
- Mobility: access to the application from various devices.

The above listed characteristics are among the main ones that need to be covered in creating systems for the analysis of web resources. We have planned to include exactly them in the proposed model of web-based system.

A key concept in the analysis of web resources and the pursuit of knowledge is business intelligence, which in recent years has been changing at exceptional rates (Rennhackkamp, 2012), (Chen, Chiang, 2012). The use of analyzes is increasing, together with the growing volume of data and their increasing complexity. The trends that can be seen in creating software for analyzing are the transition from descriptive reports and dashboards to analysis of the data and forecast modeling. So far the traditionally used analytical tools such as ETL, data warehousing and reports cannot fully meet the growing demands of organizations. All this is a prerequisite for talking about new modern alternatives.

In the context of what has been described so far, over the last few years more and more large companies such as SAS, IBM, etc., have imposed the term Web Analytics 3.0 analysis (Philips, 2014) (Davenport, 2013) and (Davis, 2016). Its
advantages over its predecessors Web Analytics 1.0 and Web Analytics 2.0 include the ability to apply hybrid technology in the process of analyzing web resources that include traditional and not so familiar techniques as well as analysis becoming an inseparable part of business decisions. In terms of the methods applied intelligent agents and neural networks are increasingly used at different stages to automate some of the processes of analysis of web resources and data mining (Song, Fang, 2010, p. 201).

The standard applications for web analytics that are available on the market today are representatives of Web Analytics 2.0 (Dubois, 2011). The common feature of the systems of this kind is that they are three-layer. The first layer includes the data itself such as log files, social networking and other individual files. The log files are the main source used for data in a large number of applications. The log files are the main source used for data in a large number of applications. The second layer is associated with the method of processing the collected data and how the information obtained will be applied to business. The collected data are represented by a standard or modified dashboard, where the key identifiers reflect different perspectives.

The third layer is the link between the other two, which provides an overview of the development of the company in order to be able to answer questions put before the business. The disadvantage of these systems is the lack of a possibility to support the decision-making process.

Based on the above, we believe that web analytics products for small and medium-sized businesses need to step onto and borrow some of the characteristics set out in Web 3.0 analysis and Gartner report from 2015.

2. Disadvantages of the ongoing process of web mining from web resources

As mentioned in the previous paragraph, most software products for web analytics are representatives of the Web Analytics 2.0 and use as fundamental source of data log files, and as a way to visualize the obtained information dashboard are applied. Log files record every page visit from a website that is hosted on a server. Entries in the file consist of several fields such as date and time of visit, accessible resource, IP address, and browser used (Markov, Larose, 2007, p. 148). Collected data in the files are used to obtain knowledge about navigation behavior of the users in the sites. Therefore the main source of data for web mining and the process of customizing the web services are log files (Losarwar, 2012). The disadvantages of the current process of data mining can be summarized in the following:

- A lengthy process. The process goes through several stages, which leads to slowing it down. It is necessary, initially to define what information sources will be
used, then to integrate them into the system, and finally to identify ways of visualization of the offered reports by a dashboard.

- Use of different software programs. Analyzing only the log files does not give us enough information about the e-store and this requires the use of other products in the process of analyzing usability and data mining. (Croll, Power, 2009, p.353). On the software market for a website analysis there is a variety of products, each of which adds an additional type of analysis. Part of the applications use software for analysis of traffic to the site, others focus on data from social networks, still others, emphasize on the analysis of the advertising and marketing of a given company. Out of the available diversity the user trusts a number of software applications for analysis and uses them to get a complete picture of the current status of their e-business. This in turn requires additional time needed for examining the reports received by the dashboard and extra time for training to work with interfaces of different products. On the other hand, users interpret the obtained knowledge themselves, which leads to subjectivity (Troster, 2016).

- The use of various types of specialized software results in increased costs. As mentioned, the current software applications lack the means to support the process of decision making, such as forecasting tools. At this stage, making, a qualitative analysis of the e-store, which incorporates forecasting numbers and data mining, requires several types of software. The testing of various applications is associated with costs that a start-up business can hardly afford. This is why small businesses prefer to invest in quality analysts and use free software that has limited features (Kaushik, 2007, p. 84).

Thus, it can be summarized that the process of web mining from web resources requires knowledge of various software products and data exchange between them, which is time consuming and accompanied by high financial costs.

It is necessary to find a solution to the problems described above and to propose a new approach to this task. Such an alternative is the creation of a software product which integrates the advantages of products used alone to extract knowledge from web resources with appropriate intelligent methods.

The prerequisites for the creation of such software are several. Firstly to build applications for the analysis of web resources it is necessary to include intelligent methods that can handle, on the one hand, the large volume of incoming data - often incomplete and inconsistent, and on the other hand, to analyze the various data types. Secondly, the growing number of users who do not have deep knowledge of web mining, but are willing to analyze data from their e-store, requires the creation of instruments to help reading the extracted knowledge.
Taking into consideration what was mentioned above, we believe that software is necessary for analysis of web resources to assist users through intelligent methods in making decisions and providing information in a comprehensible form.

3. A conceptual model of web-based system for web mining from web resources

The proposed web system should allow for real-time analysis to be carried out and to respond appropriately to changes in the site. As the most significant phase of the system is determined the ability to communicate with the consumer and display proposals for the future development of the website, which in the current process of web analytics is missing.

The created conceptual model (Figure 1) is based on data flows and activities taking place in a web-based system. The conceptual model illustrates that users are able to communicate with each module separately and receive a result depending on their current needs, through which the characteristic “self-service” is executed.

Figure 1. A conceptual model of web-based system for data mining from web resources
The system includes three modules: Analysis, Multi-agent system (MAS) and Neural network (NN). Each of the modules performs a particular type of tasks which in the currently existing process are committed by various types of software.

The Analysis Module aims to collect data from different informational web sources which the owner of a store uses to analyze the state of their business. Such sources of information can be: the website e-store, social networks and the created fan page of their company or product, applications for the analysis of the log files and applications for marketing campaigns.

Practice shows that monitoring and analysing the e-store does not just rely on one source of information but uses several web resources to be able to cover all aspects of the researched object.

For integration and visualization of data the web system relies on a standard tool called dashboard (Few, 2006, p. 12). Dashboards are interactive tools for visualization of key performance indicators which significantly improve the way of understanding the analyzed data (Lavinsky, 2013). According to the Gartner Group report (Gartner Group, 2013) for business intelligence and analysis platforms “more and more software companies build diagnostic analysis using critical advantages of visual dashboards. This software allows users to understand the different aspects of the data and to discover new knowledge.”

The main functions that implement the module are as follows:

• Gathering and summarizing data from log files and applications to analyze various activities of websites;
• Representation of user`s data in different variations by dashboards;
• Transmission of a summary of the MAS module.

The various activities in the Analysis Module are implemented in separate layers. For creating the triple layer architecture we have investigated the architectures of the three highest-rated software products for data analysis and have used the dashboards – Tableau Software, QlikView, Microsoft software. For example, communication with the consumer can be found in a separate layer. The data, which is obtained from applications, is appropriate to be in a separate layer and its transformation should also be separated. This yields a three-layer architecture, on the base of which the module is created (Figure 2).
Figure 2. Analysis Module architecture

Data layer. There are diverse data that are received by the applications to analyze websites in this layer. Except the log files and data from social networks, data for sales analysis and others are included in it.

Server layer. The layer is composed of two types of servers. One server is used for obtaining the converted requests from users and for displaying data such as images. The other server centrally manages data sources and executes processes on data analysis, and an analysis of the incoming log files. As a result of the executed query the returned data is visualized by various graphs and charts (Hanrahan, 2006).

The role of the connectors is to connect the server layer with the application data. A number of standard connectors for databases such as Microsoft Excel, SQL Server, Oracle and others are also included. There is also a general ODBC connector for systems that do not have their own connector. One of the main advantages offered are the two modes of operation: "live link" and "memory storage." The users choose what mode to work in and which apps to visualize data from. The request is submitted to a router, played by Gateway. In turn it transmits the request to the layer of servers that converts it and submits it to the connectors. The latter in turn connect with the selected applications. In reverse order a response is returned to be submitted in an appropriate form to the customers.

The explored architecture greatly resembles the architecture of the applications of Web 2.0 analysis, where there is a division into separate layers of the information sources and their way of processing. The advantage of this division is the ability to add new sources at any time, which also executes the characteristics of flexibility and expandability. The envisaged module functions overlap with the existing applications used. By adding the other two modules, the process of web analysis could fully meet the needs of the users and help making business decisions.
The model of the web system contains the idea of a multi-agent system which performs functions of local expertise, conversion of input and output data, communication with the other modules and the user. Every one of the other modules works with a different type of data and the relationship between them would not be possible without adding an intermediate environment.

MAS is the key to the whole web system because it is exactly where the automated data mining and preparation of proposals for making business decisions is carried out. According to Jones and Jacobs (2000) the cooperation between the agents, the exchange of knowledge and results creates prerequisites for solving a problem.

The composition of the MAS includes 4 agents (fig. 3), each of which performs a different task, and the result is used to achieve the purpose of the entire module. Given the participation of the four agents and the streams of data to be exchanged between them, we believe it is appropriate to choose an organization of agents in the MAS type “team” (Horling, Lesser, 2005) (Argente, 2006). In the team organization several agents “agree” to work together to achieve a common goal.

**Figure 3. A conceptual model of MAC**

The role of the **Web Agent** is to connect with external systems that do not share a common language for communication among the agents in MAS, namely Analysis modules and NMmodule. Thus the results of the two modules of the Web system are transferred for further processing by the agents before the end results are presented to the user.

**The control agent** is based on rules, and based on the information received from various agents, it checks the rules applicable to a given situation and displays the result. It receives the user's query, converted by a communication agent, considers
what data it needs further and sends a request for information resources to other agents. After receiving the necessary data, the rules are triggered and a solution is output and then transmitted to the agent communicating with the user.

The purpose of the communication agent is to provide graphical interface for the user interaction with the web system. Once the application is processed a response is returned by the web system.

The main goals of the Information Agent are to collect data from other agents, convert them to an appropriate form and take care of their transmission and store them into an information block. It directly contacts the control agent by offering information resources.

Even if one of the agents stops functioning, the work of the MAS will not be compromised (Riekki and Huhtinen, 2003, p. 1188).

The main functions that MAS implements are:

- Communication with the user and understanding of their goals (communication agent);
- Collecting information about the usability of the site from the Analysis module and from the Neural Networks module (Web agent);
- Transformation of data (Information agent);
- Analyzing the information received from Analysis module and Neural Network module, data mining and outputting proposals related to the web site (Control agent);
- Answer the purposes put by the consumer (Control agent).

The agent structure approach that is proposed to build the MAS, enables clear differentiation of the tasks and the functionalities of each of the agents in the MAS. This, in turn, allows for expansion of the MAS, by adding new intelligent units. Thus, the MAS can easily adapt to the changes made and requirements.

The observation of data from different sources and their analysis is often insufficient for the development of an e-store. Important for the business is the ability to predict values and move to different scenarios in order to better understand the state of the e-store. In order to predict certain values we believe it is appropriate to use a Neural Network module. The benefit of applying NN is evidenced by a number of examples from the businesses, related to user profiling with various forecasts, pattern recognition, data mining and others. The advantage of combining NN software is the ability to automatically process the data from NN which come from different sources. The connection of the NN module with the user is expressed in two ways. On the one hand, users can import their own files with sample data on the basis of which to build a NN. On the other hand, to submit applications and find predictive values with the already “trained” NN (Heaton, 2010, p. 112). The reverse
communication is expressed in displaying query results and the identified relationships between selected values. The Neural Network module is composed of two parts – “Creating NN” and “Applications”.

The input data for the sub-module “Creating NN” represent files containing tabular data from which to create, train and test NN. The files can be obtained from the MAS or directly from the user. The possibility of obtaining data from various sources ensures that the system will work even if there is an error in the other modules.

Functions that a NN implements are as follows:
• Processing and clearing the noise of received files;
• Creating and training NN;
• Displaying the relationships between variables.

The advantage of using NN is the possibility of files being processed to contain data from different applications. This can capture connections and discover new relationships that by other means could not be implemented. This tool is used especially when connecting data from essentially different software products (e.g. Google Analytics, Piwik, Sales Force and others) used for the analysis of the e-store. Sub-module “Applications” assists the user in the query to NN and putting the results into a type that is comprehensible to the user. It is most clearly visible in this sub-module that a self-service feature is implemented, which was mentioned in the Gartner report of 2015, and namely, that the user himself sets the request depending on their current needs.

Combining MAS with a NN and a dashboard in one web system supports the user into performing a qualitative analysis of information resources for an e-store and data mining, taking advantage of each of the types of software. All of the characteristics of a web system in the Gartner report that we started with, are implemented in the individual modules of the system.

A web-based system is built on a modular principle and brings the advantages of the modular architecture (Homer, 2013). The chosen principle ensures minimum correlation between the components of the system, and thus the risk of failure is limited. The ability of the user to communicate with each one of the modules separately, gives freedom to choose according to specific needs. Based on the described manner of operation of the system and data mining, it is assumed that the system adequately meets the needs of users and provides the appropriate recommendations for future action.
Conclusion

The article proposes a model of a web system, which eliminates the obstacles in the process of web mining from web resources. The proposed system consists of separate modules, which makes it flexible and adaptable to changes. A key role in it have the intelligent technologies that enable operating under conditions of enormous volumes of information (the big data problem), processing different types of data, solving problems for which there is no algorithm and so on. The model was successfully approbated.

The web-based system is designed for companies and individuals that use different web based systems and need forecasts and analyzes of dependencies. The proposed system meets the needs of users, providing cooperation over the steps to achieving the objectives set by means of communication and process automation.

References


