Economic Horizons, May - August 2019, Volume 21, Number 2, 173 - 189
UDC: 33 eISSN 2217-9232

www. ekfak.kg.ac.rs

Review paper
UDC: 004:378.4]:005
doi:10.5937/ekonhor1902177B

TRANSFORMING WEB DATA INTO KNOWLEDGE - IMPLICATIONS FOR MANAGEMENT

Zita Bosnjak*, Olivera Grljevic and Sasa Bosnjak

Faculty of Economics in Subotica, University of Novi Sad, Subotica, The Republic of Serbia

Much of one's online behavior, including browsing, shopping, posting, is recorded in databases on companies' computers on a daily basis. Those data sets are referred to as web data. The patterns which are the indicators of one's interests, habits, preferences or behaviors are stored within those data. More useful than an individual indicator is when a company records data on all its users and when it gains an insight into their habits and tendencies. Detecting and interpreting such patterns can help managers to make informed decisions and serve their customers better. Utilizing data mining with respect to web data is said to turn them into web knowledge. The research study conducted in this paper demonstrates how data mining methods and models can be applied to the web-based forms of data, on the one hand, and what the implications of uncovering patterns in web content, the structure and their usage are for management.

Keywords: unstructured data, web mining, customer segmentation, on-line behavior modelling, collaborative filtering

JEL Classification: O33, L190, D85, Z13

INTRODUCTION

The World Wide Web (the web) emerged in the 1990s and has experienced an immense expansion thanks to the widespread use of microcomputers, the development of hardware (primarily microprocessors, memory elements and data storage technologies), the development of user-oriented and easy-to-use software tools, and the incredible possibilities the web, as a global information system, has created in the business domain. The second-generation web technology has enabled users to create their own content on the web, so the web we witness today is the confluence of the web pages, images, videos and other online materials that can be accessed through a web browser, as well as the confluence of interactive media (social media) and user-generated content. In order to find information about their clients, companies are no longer forced to collect information through interviews, questionnaires or some other kind of interaction with their users as these pieces of

* *Correspondence to*: Z. Bosnjak, Faculty of Economics in Subotica, University of Novi Sad, Segedinski put 9-11, 24000 Subotica, The Republic of Serbia; e-mail: bzita@ef.uns.ac.rs

information are already voluntarily provided on the web. However, a constant increase in the volume of generated data and the diversity of their sources have made the procedures for selecting adequate metrics and gaining actionable insights into data in order to improve a business more challenging. Huge amounts of data and information have made it difficult to search and find specific information, and more time is even needed to collect and transform data, whereas less time remains for strategic planning (Markov & Larose, 2007, 4). This has led to the development of the intelligent techniques and tools that provide support in carrying out an increasingly complex task of analyzing web data and extracting information and knowledge from them.

Knowledge Discovery in Data (KDD) is an area relying on the achievements of machine learning as part of artificial intelligence, with the aim of extracting potentially useful and interesting knowledge from large quantities of data. The KDD process has several phases, which include data preprocessing (i.e. data cleaning, integration, filtering and transformation), data mining (i.e. the identification and analysis of patterns), and the implementation of the extracted knowledge. Traditionally, data mining (DM) techniques have been applied to the data collected over time in relational or transactional databases or an enterprise's data warehouses, but they can in principle be applied to any type of the data repository, including web data.

The concepts, techniques and application possibilities of web mining, the discipline rapidly developing thanks to its great potentials for business, especially e-business, are described in this paper. As web mining can be carried out through different approaches and by using different technologies, its implementation is not uniform and clearly defined, but varies depending on the tasks, the scope of application, and the purpose of analysis.

Therefore, the main subject matter of this research is the key aspects that need to be taken into account when defining the application of new technologies and during web data analysis. Furthermore, the research deals with the most common types of tasks which this relatively new discipline can be successfully applied to, and describes the specificities of different approaches used in their solving.

In this paper, an assumption is made that web data mining has a very heterogeneous application that brings benefits to many spheres of business doing, thanks to the discovered knowledge that would otherwise stay hidden in an abundance of web data. In order to confirm the starting assumption, the relevant literature is examined, the applications are looked for, and the results of web mining in different industries and tasks are reported.

The advantages of web mining in different spheres of business doing compared to the other analytical approaches, as well as the possibilities for management to incorporate these new technologies in business doing, are investigated. Accordingly, we first wanted to provide a systematic overview of the different ways of web data mining depending on their source (e.g. website visits, the interconnection of web pages, the key issues relevant to web documents, etc.) and the type of the knowledge one would like to reveal (e.g. whether we want to find out the profile of the visitors of a website who made an online product purchase during a visit, derive a model for the automated classification of large quantities of documents into specified categories, or the pieces of information about which products should be recommended to the visitor of the website in order to maximize the opportunity for a potential online purchase, etc.). Furthermore, we wanted to collect some practical examples of web data mining from relevant sources. Our research study is focused on the applicability and benefits of web content mining, the structure and the usage in a business environment, on the one hand, whereas on the other, information retrieval technologies themselves and their efficiency are not within the scope of this paper.

The paper is structured as follows: in the second part, a review of the literature on the theoretical and methodological aspects of the transformation of web data into knowledge is presented. The third part is a description of the three key aspects of web data mining:

- web content mining by means of clustering, information retrieval systems, collaborative filtering and special approaches to non-textual data analysis;
- the web structure mining and the analysis of the structure by page ranking;
- web usage mining.

The possibilities of applying the described approaches in different business spheres are described in the fourth section, along with several examples of the best practice. The final considerations of and the future directions for web mining are given in the fifth section of the paper.

A LITERATURE REVIEW OF THE THEORETICAL AND METHODOLOGICAL ASPECTS OF TRANSFORMING WEB DATA INTO KNOWLEDGE

Knowledge in Data Discovery is a process of the extraction of the valuable information and knowledge hidden in large quantities of data and documents. The generic methodology most frequently used for the KDD, which assumes the role of the standard methodology, is the CRISP-DM (Cross Industry Standard Process for Data) methodology (Shearer, 2000; Jiawei & Kamber, 2001; Guandong, Yanchun & Lin, 2011). This methodology consists of the continuous, iterative, domain-independent steps applicable in any industry and in all business domains. According to CRISP-DM, it is essential to have a good insight into the business processes themselves, thus transforming business questions into data mining goals. At the next step, relevant data should be collected, understood, and prepared for subsequent analysis. The data preparation step, the so-called preprocessing, depends on the form and quality of the source data and on the goal of data analysis. It consists of data integration, cleansing and filtering, and their transformation into an adequate form. Data preprocessing is followed by data modelling, i.e. the extraction of hidden patterns - relations, rules or behavioral schemes. At this step, scientific methods and data mining techniques are applied. The evaluation and validation of the developed model represents the third step, which should answer the question of whether the model is compliant with the business goals or not. If the said is the case, the extracted patterns should be interpreted and a plan for the revealed knowledge deployment should be developed. There is no fixed sequence of the mentioned steps; it is, however, a common practice to return to the previous steps from the consecutive ones if the obtained results are not satisfactory. The discovered knowledge is used for defining more specific business goals to which better solutions could be found based on the experiences of the previous iterations.

In 2015, the IBM Corporation presented a new methodology, ASUM-DM - Analytics Solutions Unified Method for Data Mining/Predictive Analytics, which is an extension and improvement of the CRISP-DM methodology (IBM Analytics, 2016). This methodology covers the infrastructural and operational aspects of data mining and predictive analytics and managerial activities in the deployment phase to a greater extent.

E. Yoneki, J. M. Tirado, Q. Guo and O. Serban (2016) have described an expert-centric methodology for knowledge discovery in web data. According to their methodology, the overall task of knowledge discovery is tightly bounded to experts and experts' knowledge, implying that an expert is he who is involved from the very beginning - from the first, data-collecting step throughout the process, all the way to a pattern extraction and knowledge deployment. A broad set of software products is at the disposal of experts as a support in the iterative knowledge-discovery process and only minimal interventions are needed by the development team.

When applying the data mining methods of clustering, classification, prediction or regression to the large amounts of web data in order to leverage the business, the web mining discipline, which has three different aspects (Markov & Larose, 2007; Liu, 2007; Palau, Montaner, Lopez & de la Rosa, 2016) is implied, namely:

- Web Content Mining - This task consists of finding relevant content on web pages for a user query. Each website is so designed to provide some pieces of information to the visitors of the website. However, information is not limited to the textual form, but can also be in the form of images, graphic elements or tables, which makes a search for content more complex.
- Web Structure Mining - The task is to analyze the hyperlink structure, i.e. the way documents are linked together. The in-page structure explains the organization of HTML (Hypertext Markup Language) or XML (Extensible Markup Language) tags on the page. The information about the external connection of a web page with other web pages is in the form of a network of hyperlinks.
- Web Usage Mining - The task is to analyze web page visitors' navigating paths. The data that provide information about the web-page access scheme are most frequently stored in the Extended Common Log Format in log files on servers and contain information such as the Internet Protocol (IP) address, the website reference, the access time and the access point.

The steps in preparing and modeling web data are specific to each of the above-mentioned aspects. In the process of document clustering, each text record is represented internally by the preprocessed words that the document contains, which are referred to as characteristics. Characteristics are the dimensions by which, in the procedures of a further analysis, documents are compared with each other according to their similarity (Liu, 2007). All documents are formally represented in the Vector Space Model, which is created through the steps described in (Cheng, Healey, McHugh & Wang, 2001):

- For each characteristic, the degree of the frequency of occurrence in the observed document is calculated in order to eliminate rarely occurring words.
- The lower threshold value of the frequency is arbitrarily set at the beginning of the clustering process, which reduces the clustering dimensionality.
- Each characteristic is mapped to the n degrees of significance, one for each of the n documents from the starting set.
- Degrees of significance are recorded in the form of the $n \times j$ matrix, where j stands for the number of the characteristics in the input document set.

This is the so-called Vector Space Model (VSM). The degrees of significance in matrix cells are calculated as a ratio between the term frequency in a specific document and the term frequency in the other documents from the input set. This is the well-known TF-IDF measure (the term frequency-inverse document frequency).

Information retrieval techniques are used in order to find the information of interest to the user in web content mining. These techniques observe the topic/subject matter of documents when selecting those relevant to the user. The efficient filtering techniques using the topic of the documents so as to select the relevant ones are: vector-space queries, intelligent agents and information visualization (Cheng, Healey, McHugh & Wang, 2001; Kumar, Bharani & Mohamed, 2018). When querying in the vector space, documents are searched and ranked according to the (cosine) similarity with the vector presentation of the given query, which is analogous to the already described method of calculating the similarity of the document with the cluster centers.

Collaborative filtering improves information filtering techniques and the performance of information retrieval systems by concerning attributes beyond the mere analysis of the content of documents, such as the user's preferences and tastes and his/her comprehension of the quality of a product, a service or some other entity of his/her interest. J. L. Herlocker, J. A. Konstan, A. Borchers and J. Riedl (1999, 232) stated an assumption that the users who demonstrate a similar behavior share similar interests, i.e. that there is a high degree of correlation in their preferences. This assumption is taken as the foundation in collaborative filtering (Resnick & Varian, 1997, 57) and lies behind the recommender systems operation.

By definition, image mining deals with the extraction of image patterns from a large collection of images (Fayyad, Djorgovski, & Weir, 1996). Therefore, its focus is neither on understanding of and/or extracting specific features from a single image nor on the problem of retrieving relevant images. The image mining process consists of image storing, image processing (in order to improve their quality), the extraction of features (in order to generate the important features from images), image indexing and retrieval, patterns and knowledge discovery (Madhumathi & Selvadoss Thanamani, 2014, 1818). So-generated features enable mining by using data mining techniques so as to discover the significant patterns evaluated and interpreted in order to obtain the final knowledge that which can be applied to applications.

A video is an example of multimedia data as it contains several kinds of data, such as the text, the image, meta-data, visual and audio data: a video consists of a sequence of images with some temporal information; an audio consists of speech, music and various special sounds, whereas the textual information represents its linguistic form (Vijayakumar & Nedunchezhian, 2012). In order to be able to analyze such heterogeneous data types, data should be transformed into structured-format features (Rui & Huang, 2000). A video data model is a representation of video data based on their characteristics and content, as well as the applications they are intended for (Kokkoras, Jiang, Vlahavas, Elmagarmid, Houstis & Aref, 2002). It is based on the idea of video segmentation or video annotation. M. Petkovic and T. D. Jonker (2001) proposed a content-based retrieval data model with four layers:

- the raw video data layer with a sequence of frames, as well as some video attributes;
- the feature layer, consisting of the domain-independent features that can be automatically extracted from raw data, characterizing colors, textures, shapes, and motion;
- the object layer, with entities characterized by a prominent spatial dimension and assigned to regions across frames;
- the event layer, with the entities that have a prominent temporal extent describing the movements and interactions of different objects in a spatial-temporal manner.

There are two main approaches to audio data mining (Leavitt, 2002), namely:

- the text-based indexing approach, which converts speech to a text and then identifies the words in a dictionary of several hundreds of thousands of entries,
- the phoneme-based indexing approach, which analyzes and identifies sounds in a piece of audio content so as to create a phonetic-based index.

A dictionary of several dozens of phonemes is used for the purpose of converting the user's search term into a correct phoneme string and the system looks for search terms in the index in order to find the most adequate one.

A video text can be obtained from three sources (Ma, Lu, Zhang & Li, 2002), namely: as a scene text (e.g. from billboards, a text on vehicles, and writings on human clothes), a superimposed text (a mechanically added text to the video frames in order to provide additional information for the purpose of a better understanding of the video), and automatic speech recognition.

For the web structure mining, the manner of the calculation of page ranks is crucial. The calculation of the ranking of a website is a difference between the number of all the inbound links and the number of all the outbound links from that page. The page that has more inbound links than is usually the case for the network being viewed is called the reference node, whereas the page with a significantly larger number of outbound links than the average is the so-called index node in the network.

The basic idea behind a slightly different approach to the calculation of a rank is that hyperlinks are the indicators of human judgment about the mutual relevance of the connected pages, and that the appearance of a link on page *p* to page *q* carries the latent information that the author who created page *p* and included a link to page *q* on it "transferred" a certain degree of importance of page *p* to page *q*. Consequently, the greater the number of inbound links from other websites, the greater the significance or authority of a page. Therefore, complex ranking approaches take into account a wider organization of the web in addition to authoritative pages with many inbound links. They also reveal pages related to many authorities, i.e. pages with many outbound links, the so-called hubs. Hubs linking authorities for a common topic allow the detection and rejection of "false" authorities, i.e. unrelated pages with a large number of input links. Any improvement of the performance of the existing methods for searching relevant websites is closely related to the issues of the efficiency of search algorithms and available data storage capacities.

Web usage mining explores data about all the activities the user carries out with respect to accessing the web that are automatically generated until the user logs out: ranging from where the visit was made, the path the visitor moved along during the search, the time spent on each page, where the user went after the visit, etc. These data are stored in log files on servers. Each mouse click after the user logs in corresponds to a single web page request, and the sequence of such clicks corresponds to the sequence of links to the pages the user has visited. In B. Liu (2007), the basic concepts of web usage mining are explained: the user's visit to a website is called a session, while the sequence of the pages viewed during a single session is called clickstream data or web clicks. The method of generating association rules from transactional data, as known as market-basket analysis, which is widely used in commercial applications, can also be used in order to analyze the web log data (Jiawei & Kamber, 2001; Li & Feng, 2010). The association rule is the rule of form $X \Rightarrow Y$, which should be interpreted in a sense that the purchase of the item X entails the purchase of the item Y in the same transaction. In the web environment, the same rule indicates the relationship between the HTML pages X and Y, which frequently appear side by side in users' sessions (Markov & Larose, 2007). The association rule does not carry information about the chronology of the visited sites. The techniques for generating sequential association rules that include a temporal component are applied for these types of analyses. In sequential rules, a visit to a website listed in the antecedent of the rule was made prior to a visit to the websites listed in consequence of the rule, i.e. the sequence of clicks is analyzed in the time dimension (Markov & Larose, 2007).

The existing algorithms generate too many rules, out of which it is difficult to distinguish important ones and neglect those deprived of a potential business value. Therefore, it is necessary to determine the criteria in order to ensure the implementation of only useful association rules, whereas all the rules that cannot be sufficiently used to achieve desired business improvements should be rejected. These criteria are called interestingness measures (Tan, Kumar & Srivastava, 2004; Hilderman & Hamilton, 2013).

The question of the accuracy of the log file analysis due to web caching and the need to delegate a web usage mining task to companies specialized in data mining have resulted in a different approach to web data collection, the so-called page tagging. The method uses JavaScript embedded on a web page, so that every time a user searches the page through a web browser, or clicks on a link, he/she generates a request to an analytic server with third parties. Both web-usage data collection methods can be applied in order to report traffic on web pages.

THE MAIN ASPECTS OF WEB DATA MINING

The incredible opportunities offered to a business by the web as a global information system could be viewed from the following three aspects: the opportunities provided by web content mining, the

business opportunities obtained by the web structure mining, and the advantages gained by insights into web usage mining. These are further described in the sequel.

Web Content Mining

Web content mining is defined as the "research method for making repeatable and valid conclusions from information about their context" (Krippendorff, 1980, 36). The content of the web as the repository of the data made publicly available by the user carries information about the web-user's attitudes, preferences, opinions and behavior, which can be significant in many spheres of business.

The first issue of the web as a global information system is the automated storage, access, finding, organization and presentation of data in web documents. The concept of the document was previously associated with the text files generated in the word processor, only to now be used to describe any type of the file an application can generate. A web page is also a document usually written in HTML, which can be accessed by a browser by specifying a uniform resource locator (URL). In addition to the textual content written in some of the natural languages in a free form (not following a predefined or prescribed structure or having just a partially defined one), a web page may contain images, audio, and video data and hyperlinks to other documents. Most broadly viewed, e-services, archived e-mails, and other content can also be categorized into web documents. The analysis of their content is the subject matter of the specialized approaches that go beyond the scope of this paper.

The essence of web content mining lies in the ability to find the documents that are relevant for the user. Content analysis is a widespread method for the objective and systematic quantitative testing of the delivered content and may be useful for discovering or understanding users' preferences and behaviors in the user-generated complex social and communication trends and patterns (Kim & Kuljis, 2010, 373). Because of the semi- or unstructured document format, interactivity, decentralization, and the hyperlink network structure, classical database management systems cannot be utilized to locate the required documents in a variety of web documents and newer approaches are necessary instead.

Clustering

The clustering task can be described as the segmentation of a heterogeneous input set into subgroups of elements with a high degree of mutual cohesion. In the context of the web, the starting population can be made either of the documents/web pages grouped into subgroups according to the meaning of the containing terms (clustering by similarity), or web users, which are clustered based on the activities they perform when visiting the web. When web content mining is considered, documents are clustered, while web usage mining clusters website visitors based on their web usage. Although the terms web user and web visitor are often used as synonyms, yet there is a distinction between them in the context of web mining, which results from the data types automatically generated and stored on web users/visitors. The term user is more general and denotes an individual accessing the web. A unique identification code is given to each user, which distinguishes that particular user from other web users and remains unchanged even if the user switches the device or browser on a return visit to the website. The term visitor denotes an individual accessing the web from a specific device (e.g. from a laptop computer, a mobile phone, etc.) or a specific browser (Google, 2019). In this manner, one user can be represented by several visitors. It is very important for web data mining that access from another device or browser creates a new client ID as it enables data analysts to detect differences in the online behavior of the same user, e.g. in the case when such a user accesses the web from his/her own laptop computer, and in the case of access from a mobile application. Having the terminological distinction in mind, the term visitor will be used in the sequel of the paper only if we want to emphasize the fact that the specific device/browser is taken into account during web data mining. Otherwise, the term user is used.

Clustering algorithms use the VSM matrix to calculate the centers of clusters. At the next step, the distances of each document from the input set to all of the cluster centers are calculated and the results are recorded in a matrix, the so-called Document Vector Model. Based on the obtained similarities, discrete clustering is performed, i.e. each document is assigned to exactly one single partition of the starting set, namely to the one to whose center the degree of similarity of the document is the greatest. Clustering algorithms can find the documents relevant to the user very quickly and accurately, so they can be considered as an effective technology for finding information on the web (Fan, Liu, Tong, Zhao, Nie, 2016, 42).

Information Retrieval Systems

It is customary for web users to demonstrate a need for specific information in the form of a user-defined query. After the user specifies his/her query, the documents relevant to a particular topic or a specific subject matter of the query are found. Information retrieval systems enable search for documents by the subject matter or the topics describing unstructured or semi-structured content, thus overcoming the deficiencies of database/data warehouse management systems. At the same time, the task of the information retrieval system is to exclude from the search results as many documents irrelevant to the user query as possible. The tools that filter the documents content compare the formal representation of the content of a web document with the formal representation of user-specified content, which is indicated through the user query, thus performing a selection of the right information for each user.

Collaborative Filtering

It is common in the real world that friends give us advice on or a recommendation about the interesting products we should buy, a book worth reading, a movie we might happen to like, and so on. Formally in this scenario, they are said to be collaborating with us in the selection process. By analogy with this kind of the collaborative selection of the entities of interest, Recommender systems provide users with the pieces of advice based on the information about the behavior and preferences of other users.

Recommender systems often encourage users to explicitly evaluate entities, or provide their own preferences, which are then stored in the system. The users of a collaborative filtering system share their own analytical judgments and rankings of entities (the evaluation of purchased products or services), so that other users can more easily decide which product to buy or which action to undertake. Based on the estimated previous experience and the prior users' satisfaction degree, a recommender system creates a personalized recommendation for the new user of the entities that he/she could find interesting. Based on user evaluations, users are classified into the categories of those with the same taste or information needs.

The performance of a recommendation system is directly proportional to the degree of user collaboration (Palau, Montaner, Lopez & de la Rosa, 2016, 145). That is why a collaborative filtering method is especially useful in a world increasingly being networked over the Internet, in which a network of documents on the web builds on the common efforts of the users themselves.

Non-textual Web Content Analysis

Although the content of web documents is merely textual, documents may also contain images, video and audio data, archived mails, and other data as well. Images can be automatically classified or clustered based on the values of the basic colors (the so-called RGB components), or the texture values. Entropy is used to compare the images with some threshold constraints. Image mining is highly utilized in order to classify the medical images with the purpose of diagnosing the right disease verified earlier (Babu & Mehtre, 1995).

The objective of video data mining is not only to discover and describe interesting patterns from a huge amount of video data and automatically extract the content and structure of a video, the features of moving objects and the spatial or temporal

correlations of those features, but also to discover the patterns of the video structure, the object activities and video events out of the vast amounts of such video data with a little assumption of their content.

Audio is what plays a significant role in the detection and recognition of events in a video. Audio data mining can be used to separate different speeches, detect various audio events, analyze the spoken text, emotions, etc. The potential applications of video mining include annotation, search, traffic information mining, event detection/anomaly detection in a surveillance video, a pattern or trend analysis and detection.

Compared to the mining of the other types of data, video data mining is still in its infancy.

Web Structure Analysis

Millions of online web users with different intentions and aims continually create the content of the web and link their documents by hyperlinks. Therefore, the structure of the web is very complex and it is impossible to plan its development or influence its evolution. Web structure mining is the process of finding information from the organizational structure of web pages, where it consists of the hyperlinks that link pages formatted in HTML together. Roughly speaking, the task is to discover those websites that are relevant to the query, whereby the quality of the search is subject to people's subjective evaluations due to the inherent personal aspect of the relevance criterion.

Web users can set up very specific queries for which there are very few documents containing the response to such a query. In such cases, it is difficult to find those documents in a plethora of web documents. This phenomenon is known as the scarcity problem. In contrast to this scenario, web users can post a query whose topic is very general or broad, so that thousands of relevant websites that contain the information requested and respond to the query can be found. This problem is known as the abundance problem. In this scenario, it is necessary to find a smaller subset of the most relevant and the most significant documents, i.e. to filter out the most authoritative documents, the so-called authorities, from a large resulting collection. Since the most authoritative pages are not always those on which a particular term appears more often, but they can be the pages on which the term does not appear at all (e.g. as the leading car manufacturer, Honda does not make a mention of the term "car manufacturer" on its website), it is clear that the text-based approaches are not adequate for ranking pages by authority, and the approaches that rely on the hyperlinks that link websites are used instead. The results show that the text contained in the document citing another document as a reference often has a greater discriminatory and descriptive value than the text in the original document (Glover, Tsioutsioulklis, Lawrence, Pennock & Flake, 2002, 566).

Page Ranking by Relevance

There are many approaches to finding relevant pages in the context of the hyperlink structure of the web. They rely on the page ranking procedure, where a rank implies an associated numerical indicator. This observation is oversimplified since the number of inbound links in popular websites (such as www.yahoo.com) is very large, so they could be considered as relevant for all queries. Although links and content are still among the most important ranking parameters for web pages, contemporary ranking systems include many more indicators, none of which gives a holistic rank, but strives to highlight the most important websites and make their content more visible. In its browser, Google uses over 200 factors for determining a page rank, but their specification falls under the business secret (Search Engine Land, 2017).

Web Usage Mining

The data collected in log files on web servers have a certain degree of analogy with the data from the purchases transaction database. Each required web page can be considered as analogous to one item in a transaction, whereas a set of all the pages that a particular user requested during a single visit to the website can be considered as an analogue to one transaction, or a set of the items found in the

shopping cart at the time of purchase. One example of an association rule on a particular website could read as follows: "If a user has requested the websites *A, B* and *C*, there is an indication with a confidence level of 23% that he/she will also require pages *D* and *E*."

Web usage mining is focused on analyzing the data on the visited web pages during user sessions and allows the extraction of the knowledge of web visitors' behavior. The interesting patterns or behavioral schemes we were not aware of before bring a potential benefit to both the website owners and the visitors. The difficulty encountered by the analyses of the data collected in log files on web servers is the tendency of algorithms to generate too many association rules, among which, even by utilizing measures of interestingness, it is difficult to choose the rules providing the information useful in the given domain of application. Furthermore, both methods for collecting data - from log files or by tagging - could be used for web usage analysis. Which approach will be utilized in an enterprise depends on the cost of the analysis done in-house vs. the cost of engaging a third party.

WEB MINING APPLICATIONS

Web content mining is widely used in social and humanity domains given the fact that all social aspects relevant to web users are embedded in the content they produced - through using symbols, transmitted messages, images included, or through the phenomenon of content linking and organizing. I. Kim and J. Kuljis (2010) described the specific use of the analysis of the content of web documents in examining a cultural impact on design and the utilization of blogs. Since users create and maintain blogs independently, the tested assumption implies that blogs reflect the system of values and the preferences of authors, which are a result of cultural heritage. The results of the survey in South Korea and the United Kingdom have denied the assumption and have shown that bloggers from the countries that are traditionally characterized by a low degree of risk and uncertainty tolerance (such as South Korea) are less likely to reveal personal information about employment than the bloggers from the countries traditionally tolerant to risk and uncertainty (such as the United Kingdom), but also that they provide information about their age and contact link more frequently.

The structure of web links may indicate the similarity between linked web pages. Therefore, if we find the page *p* that we consider relevant for a particular topic, i.e. sufficiently authoritative for the given subject, by inspecting the links surrounding the page *p* it is possible to generate a response to the question which other topics are related to the starting topic in the opinion of the author who created the page and included the links in it. If page *p* is highly referential, there will be an enormously large number of independent opinions about the linkage of page *p* with those other pages, but if we use the notation of authorities and hubs, it is sufficient to find the nearest (local) authorities around page *p*. Local authorities are a kind of a summary of the general topics related to the page-p topic. Based on an insight into the hyperlink structure of the web, web pages can be categorized, or an insight can be gained into the hierarchical or network structure of the existing links on a website within a particular domain.

By analogy with the extent of authoritativeness of a website, an analysis of the web link structure can be used to measure the status, or an impact in social networking. In this sense, mechanisms for measuring the "significance" of individuals are developed through the citation of their papers, or for measuring the impact factors of the journals linked by the quoted references in an implicitly created network of scientific papers, which is used in bibliometric evaluations. The well-known measure of authority in this context is the Journal Impact Factor (Garfield, 1972), which is calculated for the year under review as the average value of the number of the citations of the scientific papers published in that particular journal in the past two years (Egghe & Rousseau, 1990). The impact factor is basically the ranking mechanism that only takes into account the input links in the network when calculating the authority.

Web usage data capture the interaction of a user with the website, based on which a user model is created as a picture of his/her behavior, interests, and personal preferences. Web usage mining has basically two applications: the analysis of traffic on web pages and application in e-commerce. Traffic analysis includes an analysis of the navigation data generated during the user session, describing the path that the user followed. Data on how many pages the user visited during the session are also recorded. Analysis usually reveals details about the visitors, such as their search style, preferences, forwarding a product/service to their friends, the number of clicks on the link and hits on a particular page, etc. Thanks to the discovery of associative rules, the visitor's behavior can be predicted by comparing website search schemas with the search schemes extracted by mining log data from other users. Based on recognized similarities and shared common interests, web content could be personalized for the user (Siddiqui & Aljahdali, 2013, 42), and recommender systems may suggest the most appropriate path for visitors to their preferred websites, or the usual route to certain e-purchases.

Large companies throughout the world have long recognized that e-commerce is not merely sales and shopping online, but also the ability to increase efficiency in a competitive market by exploiting the knowledge hidden in the large amounts of data available online. The integration of web mining techniques with e-commerce applications enables an e-store owners to improve their performance and services, and to collect information about clients and the behavior demonstrated by the customers accessing products or services online through the website. E-commerce websites generate the data that contain information on the reasons, dynamics and ways of navigating through websites, and therefore their analysis may indicate better access to preferred web content, improve the purchase process, and generate added value for consumers. In order to meet visitors' needs to the greatest possible extent, especially when consumers and loyal customers are concerned, customized web services are offered during their visit. Based on the observed similarities in the prior behavior of their visitors and the anticipated interests of the new visitor, certain websites prepare customized product catalogs (Shukla, Silakari & Chande, 2013, 8). Many of the world's leading e-commerce companies (Yahoo!, Amazon, eBay, IBM, etc.) have customized their web presentations for personalized access and use recommender systems as a form of customer support.

Web mining is very important in e-commerce for customer relationship management (Li & Feng, 2010, 279). The emphasis is on attracting new customers, the retention of the existing ones, the cross-selling of products, and the prevention of consumer churn. Customer profiling allows organizations to predict who their potential customers or consumers of their products/services are and what kind of behavior schemes they can expect from them. Web usage mining results in an analytical insight into visitors' behavior on the company's website; by understanding which part of the population they account for (concerning their respective age, gender, location, and other characteristics); how they came to the company website; which pages were visited most frequently (which content users find the most interesting); the overall visit performance, etc. Web usage mining can be very important for marketing activities as it reveals frequent navigation routes through the company's website. Every time a visitor clicks a link, an image, or another object on a page, that piece of information is recorded and remembered. One can discover the habits of each individual, but it is more useful when one saves thousands of navigation paths and finds out users' global habits and tendencies (Jokar, Honarvar, Hamirzadeh & Esfandiari, 2016, 321). Aggregated information in a form of the overall visit statistics is very useful to decision-makers because they are a clear indicator of the frequency of visits to each page, the time spent on each page, the activity of the visitors on a particular page, the ratio between the number of the visits and the number of the conversions, the popularity of products or services, available consumer choices, etc. Information about the visited pages and their order also points to the most frequent place to leave the company's website, as well as where most products are added to the shopping cart or removed from it. The analysis of the clickstream may determine the effectiveness of the website. To do this, it is necessary to quantify the visitors' behavior during

the session, i.e. record all sales transactions during the visit. The web-usage-related association rules can be used in order to better organize the website content or give recommendations for effective cross-selling (Liu, 2007). Clickstreams may reveal the pages the majority of the visitors came from and indicate the best place for placing ads, or may show whether an online marketing campaign is successful or not (whether there is a connection with online purchases). Website navigation information may indicate an (in)adequacy of online forms or the process of the selection of an item for the virtual shopping cart, and the (in)efficiency of the payment methods.

In addition to e-commerce websites, there are the websites that are not intended for sales, but rather serve as product catalogs, while transactions are made offline. Monitoring clickstream data on these websites provides us with advanced information about demand significant for supply, the inventory planning and production (Huang & Mieghem, 2014, 334). The authors point out the fact that the businesses that use visits and clicks data may reduce the storage and ordering costs by 3-5%.

The advantages of web usage mining are also evident regarding the improvement of the performance of web servers and server applications through different cache and pre-fetching strategies in order to shorten the user query response time. Website owners can simultaneously improve the usability of web pages through a better design and implementation, e.g. in order to optimize the website for searching frequently used keywords, or to eliminate mediators through which visitors came to the website.

Examples of the Best Practice

AMSOIL Inc., the first manufacturer of synthetic lubricant for cars and machines with a tradition of over 40 years, chose Mozenda's software solution (www.mozenda.com) in order to search the web so as to support its business. The Mozenda Cloud-Hosted Web Harvesting Solution allows AMSOIL to extract and organize unstructured web data in order to develop and maintain the planning resources (Mozenda, 2018). The Strategic Planning Team at AMSOIL uses the information Mozenda collects from the web for the retail distribution network planning. This primarily involves the mapping of the retail and services locations in order to identify potential locations for new retailers. Collecting and keeping track of competitors' prices is one of the most common options for every organization selling online. The Mozenda tool continuously collects publicly available information from e-commerce websites for the purpose of better understanding the market and tracking product categories and competitive prices compared to the AMSOIL prices. From e-commerce operations through retail to retail planning, quick access to unstructured web data helps AMSOIL to smoothly address a variety of strategic and tactical requirements and to successfully deal with bigger brands, such as Mobil, Pennzoil, Shell, Castrol, and Valvoline.

Tesco.com is an e-commerce subsidiary of Tesco PLC, a UK-based retailer of food and consumer goods, operating in the United Kingdom, Europe, Asia and North America. Introduced in 2000, Tesco.com has been trading with food products, consumer goods, clothing, banking and insurance services. Tesco's developers and data analysts have introduced the Splunk Enterprise Software Package in order to better understand which products and web pages users use and which web pages have resulted in the highest number of conversions. This has enabled the company to improve customer satisfaction based on new insights into online data, reduce potential revenue losses, accelerate the development cycle, and improve team collaboration. The Splunk digital intelligence solutions (Splunk, 2013) outperform classical marketing analytics and provide a complete insight into the user interaction through various digital channels, including the web, mobile communication, social media and offline interactions. The Splunk Enterprise solutions are used along with other web analytics tools in order to extract information from the historical and real-time data generated both on the client and on the server sides.

With its over-60-years-long retail experience in over 40 countries throughout the world, IKEA, a furniture and furnishing goods manufacturer, has also decided

to intelligently analyze web data in order to improve its business (Harapiak, 2013). Buyer demographic and psychographic data are primarily used to reveal their lifestyles and living habits, and consequently to improve the visual experience of a home through exhibition showrooms arranged as a real living space. Analyzing customer behavior data, IKEA has discovered a purchase pattern they named the domino effect: buying just one piece of furniture pulls out some other purchases in order to best fit it into the existing space. They have also found out that, in addition to transactions data, client demographic data are not sufficient for an analysis of consumers' habits because they do not provide information on their value system. The incorporation of psychographic data has revealed interesting behavior patterns. One of them concerns the very specific habits of the inhabitants of Pittsburgh, for which IKEA specifically created a catalog in which they primarily emphasized the price value the inhabitants in this region were particularly sensitive to, unlike other furniture characteristics prospective customers did not show sensitivity to (Dudovskiy, 2017). In IKEA, they emphasize the fact that by relying on classical analyses, this specificity would never be noticed.

AstraZeneca Co. is a multinational British-Swedish multinational biopharmaceutical company headquartered in London. AstraZeneca has seen a marked improvement in the standard of its medical research since they implemented data mining software. The Linked Life Data, a platform for gathering information from a broad range of biomedical data sources, has enabled the identification of the causal relationships hidden inside the text and making informed assertions on exactly how they are causal. This has led to the inclusion/exclusion of certain criteria for clinical studies (AstraZeneca, 2019). The Lexiquest Mine application from the SPSS software firm enables the extraction of information from the unstructured textual content of the repository that grows by thousands of documents each day (Thomas, 2002).

LIAT (Leeward Islands Air Transport) is a Caribbean airline providing interisland scheduled services operating on 15 destinations. The company has improved its customer service by mining travelers' textual messages. Text mining and the predictive classification models built in the RapidMiner route customers' messages to relevant departments (RapidMiner, 2019a). This way, the Customer Service Department is freed from manually triaging these messages, which enabled that department to focus on responding to customer's issues and needs instead. According to LIAT, a negative social media sentiment has dropped from almost 90% to as low as 40%. This early success has opened the way for many data mining applications in other business operations.

Austria's leading mobile phone service provider, Mobilkom Austria, receives more than 800,000 emailson a monthly level. Even after spam filtering, more than 80,000 customer requests still remain. Given the fact that customers expect a timely reply, Mobilkom started automating the email classification process by analyzing the textual content of incoming emails by using the RapidMiner's Data Science Platform, by which doing email requests are automatically and quickly forwarded to the support person in charge of this topic (RapidMiner, 2019b). A competent answer is guaranteed in the shortest possible time.

PayPal is a faster and safer way to pay and be paid online, or to send and receive money around the world. With 143 million active accounts in 193 markets and 26 currencies around the world, PayPal enables global commerce, processing more than 8 million payments per day. The one never-ending task for the company is to manage customer satisfaction and reduce customer churn. The background knowledge of what makes customers satisfied and what drives the improvement of product experience is gained by applying text analytics to customers' feedback in over 60 countries worldwide. They have succeeded in identifying, classifying and counting customers as "top promoters" and "top detractors" (RapidMiner, 2019c).

Nowadays, even more software tool providers (FeaturedCustomers, 2019), nevertheless these tools are entirely dedicated to data mining or have specialized platforms or modules for intelligent data

analysis, report on their websites on their positive experiences and highly diversified applications of the best practices in web data mining.

CONCLUSION

The incorporation of achievements in the field of artificial intelligence in all business domains is nowadays considered as one of the key strategic determinations, which requires not only a great degree of innovation, but also a transformation of business doing. Due to the complexity of new technologies and the changes they require, all those who want to remain competitive in the future must begin to shift towards new technologies and exploit their capabilities straightway. Having this in mind, the multifaceted aspects, tasks and possibilities of web data mining are presented in the paper, as their adequate application enables companies to reveal valuable information and knowledge, which would otherwise remain hidden in large data collections, and the potential benefits lost.

Taking the discovered patterns as the starting point, companies can undertake actionable steps for business leveraging. Therefore, it is important that a holistic and systematized overview of the diversified set of methods for each web data mining aspect should be provided. Different web data mining aspects are described in this paper, and they are matched with different methodological issues: web content mining is associated with the clustering techniques, information retrieval systems and collaborative filtering; web structure mining is matched with the ranking of relevant web pages in the context of the hyperlink structure; and web usage mining is associated with the extraction of the association rules.

Furthermore, an overview of the specific tasks that can be fulfilled by some web data mining approaches and the kinds of the knowledge that could be revealed are described in the paper as well. Such knowledge and information could help a company to better plan its business strategies and can enable a faster expansion. Numerous opportunities and benefits of the utilization of web mining vs. the other analytical approaches are outlined in various business areas: a deeper insight into social processes, measuring the status or an influence (of documents, individuals, or other entities), the categorization of web pages, an insight into the hierarchical structure of web pages, web traffic analysis, revealing common navigation paths through a company's website, the assessment of a website efficiency and the improvement of its utilization, an insight into the user-website interaction and the personalization of websites/item catalogs, a recommendation for items/services, the user profiling for improved relationship management, extracting advanced information on demand valuable for demand, the inventory and production planning, the improvement of web servers' performance and applications.

The practice so far, as well as the selected examples of the best practice, have shown that every aspect of the use of intelligent techniques contributes to improved business operations, and also that the quality of the achieved results is significantly improved by a combination of several methods (Zaiane, Li & Hayward, 2004). The foregoing speaks in favor of the starting assumption that the application of web mining is extremely heterogeneous, with evident benefits in many business spheres. Many intelligent technologies, however, web mining included, are still at an early stage of adoption and broader acceptance can only be expected once more software providers include the achievements of machine learning in their solutions, or once the offer of specialized tools for this purpose is expanded. Until then, web data mining remains an advantage in market competition, so the companies already exploiting this advantage and harvesting web mining benefits are unwilling to publicly share their experiences. Therefore, there is only a small number of the best practices described in the references, amongst which e-commerce applications dominate.

In the field of artificial intelligence, various approaches to data analysis are constantly developed and the range of possible applications is being expanded. For example, Tableau and VoiceBase (Tableau-VoiceBase, 2019) have jointly developed an analysis tool allowing

conversations made through call centers to be analyzed without transcribing, visualizing voice messages and providing an insight into unstructured telephone conversations and the client base (Sevilla, 2019). In this manner, the services of marketing, sales, production, etc. can improve decision-making and taking adequate actions. Since the paper presents an overview of the current situation in the domain of web mining, such innovative approaches are not included in it. Similarly, new classes of the data generated through mobile applications, social media, sensors, and mobile devices are the segment which an ever-increasing number of organizations will rely on in their everyday operations. Understanding the utilization of these multiple channels and creating analytical data processing capabilities in order to improve online and offline business operations are the subject matter of further research.

REFERENCES

AstraZeneca, Co. (2019). *AstraZeneca: Early Hypotheses Testing Through Linked Data*. Retrieved Jun 6, 2019, from https://www.ontotext.com/knowledgehub/case-studies/causality-mining-pharma/

Babu, G. P., & Mehtre, B. M. (1995). Color indexing for efficient image retrieval. *Multimedia Tools and applications, 1*(4), 327-348.

Cheng, G., Healey, M. J., McHugh, J. A. M., & Wang, J. T. L. (2001). *Mining the World Wide Web - An Information Search Approach*. Dordrecht, The Netherlands: Kluwer Acaddemic Publishers.

Dudovskiy, J. (2017). *IKEA Segmentation, Targeting and Positioning: Targeting Cost-Conscious Customers.* Retrieved Jun 6, 2019, from https://research-methodology.net/ikea-segmentation-targeting-positioning-targeting-cost-conscious-customers/

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics: quantitative methods in library, documentation and information science*. Amsterdam, The Netherlands: Elsevier Science Publishers.

Fayyad, U. M., Djorgovski, S. G., & Weir, N. (1996). Automating the analysis and cataloging of sky surveys. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.). *Advances in Knowledge Discovery and Data Mining*, (pp. 471-493). Menlo Park, California: AAAI Press.

Fan, G. L., Liu, Y. W., Tong, J. Q., Zhao, S. H., & Nie, Z. Q. (2016). Application of K-means algorithm to web text mining based on average density optimization. *Journal of Digital Information Management*, *14*(1), 41-46.

FeaturedCustomers, Co. (2019). *Vendor Directory*. Retrieved Jun 12, 2019, from https://www.featuredcustomers.com/vendors

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471-479. doi:10.1126/science.178.4060.471

Glover, E. J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D. M., & Flake, G. W. (2002, May). *Using web structure for classifying and describing web pages*. In Proceedings of the 11th international conference on World Wide Web, 562-569, ACM. doi:10.1145/511446.511520

Google. (2019). *Google analythics - Users (new, returning, unique) explained in great detail*. Retrieved Jun 12, 2019, from https://www.optimizesmart.com/understanding-users-in-google-analytics

Guandong, X., Yanchun, Z., & Lin, L. (2011). *Web mining and social networking: Techniques and applications*. Berlin, Germany: Springer

Harapiak, C. (2013). IKEA's International Expansion. *International Journal of Business Knowledge and Innovation in Practice*, *1*(1).

Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). *An algorithmic framework for performing collaborative filtering.* In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 230-237), Berkeley, California. doi:10.1145/312624.312682

Hilderman, R., & Hamilton, H. J. (2013). *Knowledge Discovery and Measures of Interest*. Berlin, Germany: Springer Science & Business Media.

Huang, T., & Mieghem, J. A. V. (2014). Clickstream data and inventory management: Model and empirical analysis. *Production and Operations Management*, *23*(3), 333-347. doi.org/10.1111/poms.12046

IBM Analythics. (2016). *Analytics solutions unified method - Implementations with Agile principles*. Retrieved Jun 12, 2019, from ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf

Jiawei, H., & Kamber, M. (2001). *Data Mining Concepts and Techniques*. San Francisco, SF: Morgan Kaufman Publishers.

Jokar, N., Honarvar, A. R., Hamirzadeh, S. A., & Esfandiari, K. (2016). Web mining and web usage mining techniques. *Bulletin de la Société des Sciences de Liège*, *85*, 321-328.

Kim, I., & Kuljis, J. (2010). Applying content analysis to web-based content. *Journal of Computing and Information Technology, 18*(4), 369-375. doi:10.2498/cit.1001924

Kokkoras, F., Jiang, H., Vlahavas, I., Elmagarmid, A. K., Houstis, E. N., & Aref, W. G. (2002). Smart videotext: A video data model based on conceptual graphs. *ACM Multimedia Systems, 8*(4), 328-338. doi.org/10.1007/s005300200

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. London, UK: Sage Publications.

Kumar, P. B. C., & Mohamed, D. (2018). Two-stage information filters for single and multiple sensors, and their square-root versions. *Automatica*, 98, 20-27. doi:10.1016/j.automatica.2018.09.001

Li, M., & Feng, C. (2010). *Overview of Web mining technology and its application in e-commerce*. In 2nd International Conference on Computer Engineering and Technology (pp. 277-280), IEEE Explore Digital Library. doi:10.1109/ICCET.2010.5485404

Leavitt, N. (2002). Let's Hear It for Audio Mining. *IEEE Computer Magazine on Technology News, 35*(10), 23-25. doi:10.1109/mc.2002.1039511

Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data*. Berlin, Germany: Springer Science & Business Media.

Ma, Y. F., Lu, L., Zhang, H. J., & Li, M. (2002). *A user attention model for video summarization*. In: Proceedings of the tenth ACM international conference on multimedia (pp. 533-542). doi:10.1145/641007.641116

Madhumathi, K., & Selvadoss Thanamani, A. (2014, March). Image mining: Frameworks and techniques. *International Journal of Innovative Research in Computer and Communication Engineering, 2*, (Special Issue 1), Proceedings of International Conference on Global Innovations in Computing Technology (ICGICT'14), Tirupur, Tamilnadu, India.

Markov, Z., & Larose, T. D. (2007). *Data Mining the Web - Uncovering Patterns in Web Content, Structure, and Usage*. Chichester, UK: John Wiley and Sons.

Mozenda, Co. (2018). *How To Scale Your Web Content Harvesting Operation*. Retrieved Jun 6, 2019, from https://www.mozenda.com/scale-web-content-harvesting-operation/

Palau, J., Montaner, M., Lopez, B., & de la Rosa, J. L. (2016). Collaboration analysis in the recommender system using social networks. *International Journal of Advanced Research in Computer Science and Software Engineering*, *6*(5), 137-151.

Petkovic, M., & Jonker, W. (2001). *Content-based retrieval of spatio-temporal video events*. In: Proceedings of multimedia computing and information management track of IRMA international conference. Retrieved Jun 6, 2019, from https://pdfs.semanticscholar.org/f2e0/f7557452c19e737a873b7444ca3e61e2b7fa.pdf

RapidMiner. (2019a). *Improving Customer Service with Text Mining and Auto-classification*. Retrieved Jun 6, 2019, from https://rapidminer.com/resource/customer-service-auto-classification/

RapidMiner. (2019b). *Optimization of Customer Support for Mobilkom Austria*. Retrieved Jun 6, 2019, from https://rapidminer.com/resource/optimization-customer-support-mobilkom-austria

RapidMiner. (2019c). *Sentiment Analysis at PayPal Using RapidMiner*. Retrieved Jun 6, 2019, from https://rapidminer.com/resource/sentiment-analysis-paypal/

Resnick, P., & Varian, H. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56-58.

Rui, Y., & Huang, T. D. (2000). A unified framework for video summarization. *Browsing and retrieval, image and video processing handbook*, 705-715.

Search Engine Land. (2017). *How Google measures the authority of web pages*. Retrieved January 23, 2019, from https://searchengineland.com/google-authority-metric-274231

Sevilla, C. G. (2019). *VoiceBase and Tableau Deliver New Insights through Speech Analytics*. Retrieved January 23, 2019, from https://www.pcmag.com/article/367331/voicebase-and-tableau-deliver-new-insights-through-speech-an

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13-22.

Shukla, R., Silakari, S., & Chande, P. K. (2013). Web Personalization Systems and Web Usage Mining: A Review. *International Journal of Computer Applications, 72*(21), 6-13. doi:10.5120/12664-9264

Siddiqui, A. T., & Aljahdali, S. (2013). Web Mining Techniques in e-commerce Applications. *International Journal of Computer Applications, 69*(8), 39-43. doi:10.5120/11864-7648

Splunk. (2013). Splunk for Digital Intelligence. Retrieved February 4, 2019, from https://www.splunk.com/web_assets/pdfs/secure/Splunk_for_Digital_Intelligence.pdf

Tableau-VoiceBase, Inc. (2019). *Leverage VoiceBase's Open Data Architecture to Visualize AI Powered Speech Analythics in Tableau*. Retrieved Jun 6, 2019, from https://www.voicebase.com/tableau-speech-analytics/

Tan, P-N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems, 29*(4), 293-313. doi.org/10.1016/S0306-4379(03)00072-3

Thomas, D. (2002). *Data mining boosts drug research at AstraZeneca.* Retrieved Jun 6, 2019, from https://www.computerweekly.com/news/2240046811/Data-mining-boosts-drug-research-at-astraZeneca.

Vijayakumar, V., & Nedunchezhian, R. (2012). A study on video data mining. *International Journal of Multimedia Information Retrieval, 1*(3), 153-172. doi:10.1007/s13735-012-0016-2

Zaiane, O. R., Li, J., & Hayward, R. (2004). *Mission-Based Navigational Behaviour Modeling for Web Recommender Systems*. In International Workshop on Knowledge Discovery on the Web (WebKDD 2004), Advances in Web Mining and Web Usage Analysis (pp. 37-55).

Yoneki, E., Tirado, J. M., Guo, Q., & Serban, O. (2016). MAKI: Tools for web data knowledge extraction. *Technical report UCAM-CL-TR-881*, Cambridge, United Kingdom: University of Cambridge Computer Laboratory.

Received on 11th April 2019,
after revision,
accepted for publication on 20th August 2019

Published online on 23rd August 2019

Zita Bosnjak is a full-time professor at the Faculty of Economics in Subotica, University of Novi Sad. She holds PhD in information science at the Faculty of Economics in Subotica. The areas of her scientific-research interests are: intelligent systems, knowledge management and intelligent data analysis.

Olivera Grljevic is an assistant professor at the Faculty of Economics in Subotica, University of Novi Sad. The areas of her research are: sentiment analysis, text mining and data mining.

Sasa Bosnjak is a full-time professor at the Faculty of Economics in Subotica, University of Novi Sad. He holds PhD in computer science at the Faculty of Economics in Subotica. The areas of his research interest are: databases, software development methods and internet technology.