Conceptual Approach for Development of Web Scraping Application for Tracking Information

Plamen Milev[*]

Summary:

The paper focuses on the issues of development of web scraping applications. The main features of such an application are defined with some specific functionalities that should be implemented. The theoretical framework of the article is based on web scraping as a part of data mining and on data mining as a part of business intelligence. The paper reveals some possible approaches to web scraping in terms of the development of software solutions. The article makes an overview of the research on scraping applications and the functional characteristics of several traditional solutions for web scraping that have proved successful. The author presents a conceptual approach for the development of software solution for web scraping in eight steps. The proposed concept allows for the identification of its strengths, weaknesses, opportunities and threats, that is, the performance of SWOT analysis to provide for the comparison between the traditional solutions for web scraping and the proposed conceptual approach. In conclusion, the paper shows some functional advantages of the proposed conception over the traditional software solutions in the field.

Key words: web scraping, tracking information, business intelligence

JEL: C8, L86.

1. Introduction

Nowadays there are millions of sources of information. According to some extreme views, it is no longer relevant whether the information by various sources published is correct. Sources of information nowadays tend to seek fast provision rather than precision. Hence the wealth of articles are published unchecked, given that no one wants to waste time and seeks to be the first to provide information. There is no doubt that the Internet today offers the channel through which information is spread the fastest. There is a huge number of websites providing various types of information. Logically, it should be assumed that the more often we come across certain information, the likelier it is that the information is correct, especially if it is presented on reliable sources. Right and on-time information on the Internet is available in web-based systems, among which the most important for this study are web portals for news, blogs, message boards, social networks, among other websites. According to many authors, including Yordanova and Stefanova (2017), social media and review sites are possible

[*] Senior Assist. Prof. PhD. at Department of Information and Communication Technologies, University of National and World Economy, 1700 Sofia, UNWE, +359 2 8195 312, pmilev@unwe.bg

sources of unstructured data, where users can express their opinion about products and services by posting comments. According to other sources (Marzovanova, 2015), there is a wide variety of research and projects in the area of structuring unstructured data, as well as developed and functional tools and systems that are generally regarded as laying the laws and principles in knowledge organization. Many authors (Kirilova, 2014) focus their research on the field of online public services, sharing the idea that these services are associated with the use of Internet and the status and trends of electronic interaction. Information that could be found on web sources is presented in ways other than the structural perspective. There are lots of reasons for the differences in the presentation of information on the worldwide web. On the one hand, a website might have a specific design, on the other hand - there is always the possibility that the same design is implemented differently by website developers. There are many cases in which some mistakes in structuring information on various websites are due to accidental inaccuracies made by developers. Many of the web browsers have tools that can rectify errors and accordingly visualize web content in a better way.

The aim of this study is to propose a conceptual approach through which any web content should be extracted in a structured way, despite differences across sources, and this concept could be used as a basis for the development of web scraping application for tracking information. In this sense, the aim of the article pertains to the implementation of the following sequence of tasks:

- Making an overview of the development of web scraping applications and the key role of business intelligence;
- clarifying the main points in the conceptual possibilities for developing a web scraping application;
- defining the features in the composition of the concept that may be considered as advantages of this approach.

2. Theoretical framework

According to some sources (Quora, 2017), web scraping is a method for extracting textual characters from websites, so that they could be analyzed, though not exclusively. According to the same source, web scraping is different from data mining, given that the latter involves data analysis and in this context getting data is irrelevant. Also, data mining involves the use of complex statistical algorithms. In this study the assumption is held that web scraping could be a part of data mining. In particular web scraping could be defined as a first step within the data mining process. Data mining itself is considered to be a part of business intelligence. Business intelligence (BI) was pioneered by Howard Dresner of Gartner Group in 1989. In his opinion (as citied in Power, 2007), business intelligence is a set of concepts and methods to improve the process of decision making in management, using information systems that use real business data. According to Lifecycle Software Ltd. (2002, as citied in Stefanova 2008), there are two elements that distinguish business intelligence systems from other systems, namely:

- data integration, which means merging data from different sources and in different formats, and providing coherent data access;
- providing techniques for analysis and visualization of information in a new way that is comprehensible to the user.

According to Solomon Negash from Kennesaw State University and Paul Gray from Claremont Graduate University (2003 as citied in Stefanova 2008), business intelligence systems replace some existing solutions such as decision and executive support systems as well as management information systems. Their opinion gives a new perspective on the concept of business intelligence. According to these authors, business intelligence systems perform data storage and knowledge management with the help of analytical tools in order to present complex and critical business information, which is needed for planning and decision making management. A different perspective on the concept of business intelligence is suggested by Quarles (2002) from the University of Amsterdam. In the researcher's view, the concept of business intelligence should be seen as a pyramid where data warehouse constitutes the base with three analytical layers above it, namely queries and reports, online analytical processing and data mining. Moss and Atre (2004) offer a definition for business intelligent systems, according to which these systems represent the architecture of integrated applications for service operations and support of decision making. Eckerson and Howson (2005) define the concept of business intelligence as processes, tools and technologies, needed to transform data into information, and information – into knowledge and plans that increase business efficiency. Back-end business intelligent systems include technologies and processes, associated with data warehouses, while front-end ones – tools and processes, used to carry out queries, reports and analysis of information. The authors pay attention to the transformation of data into useful information as a major contributor to business intelligence. This study will adhere to the definition for business intelligence provided by Stefanova (2008) from University of National and World Economy (UNWE), according to which business intelligence is a general term, including processes, tools and technologies used to convert data into information and knowledge needed for the support of business decisions.

The paper holds the view that, despite the wide variety of definitions, they all share

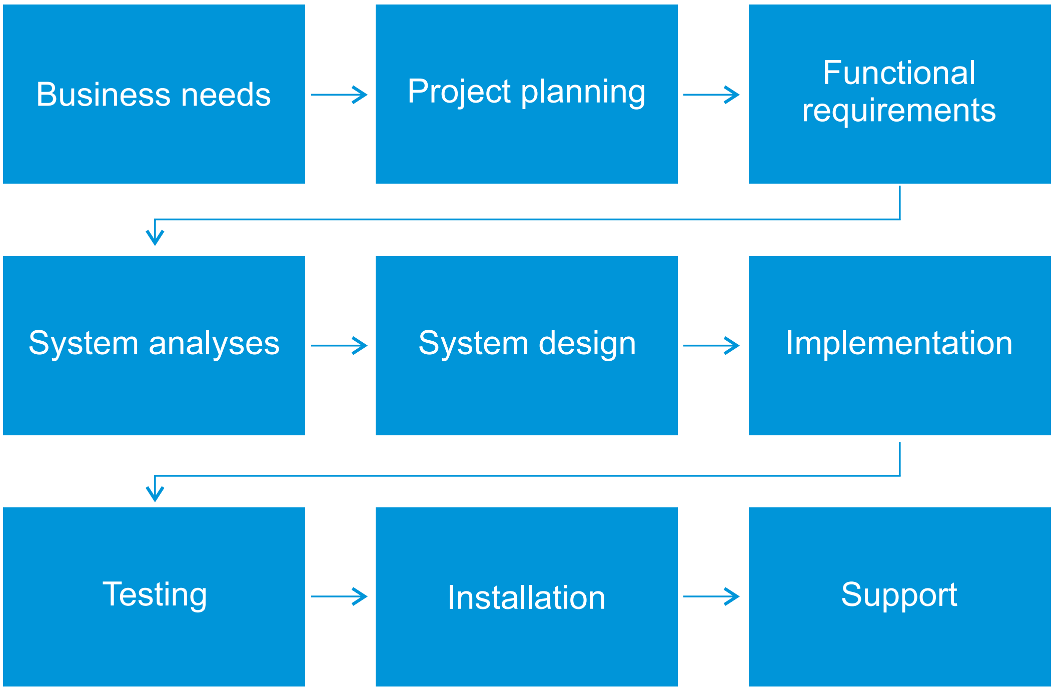

Fig. 1 Traditional approach for development of applications in 9 steps

the common assumption that the essence of business intelligence consists in the use of the data, generated by the enterprise, and their transformation into information and knowledge that support the decision making process and improve business performance. According to Stefanov (2015), in many cases decision making requires data that are found in multiple and independently developed data sources.

There are different approaches to the design and development of software systems, including web scraping applications. According to Mihova (2015), the main problems of business applications are relevant to their productivity and performance. According to other authors (Kirilov, 2016), the development of software solutions poses a host of challenges, because on the one hand there are many versions of software platforms offered on the market, developed by different companies. On the other, some software solutions are very specific. A traditional coherent approach to the construction of a system is shown in figure 1. This approach includes the analysis of the business needs of the system's user with regard to its development and maintenance. According to some authors (Moss and Atre, 2004), this approach is not effective for the development of systems for business analysis, including web scraping applications.

Business intelligence systems serve the entire enterprise and are improved continuously based on the feedback obtained from their users. Traditional approaches that have been regarded as successful in the development of information systems are not appropriate for the development of business intelligence systems, including web scraping applications, considering they do not include functionalities for the support of decision making management. According

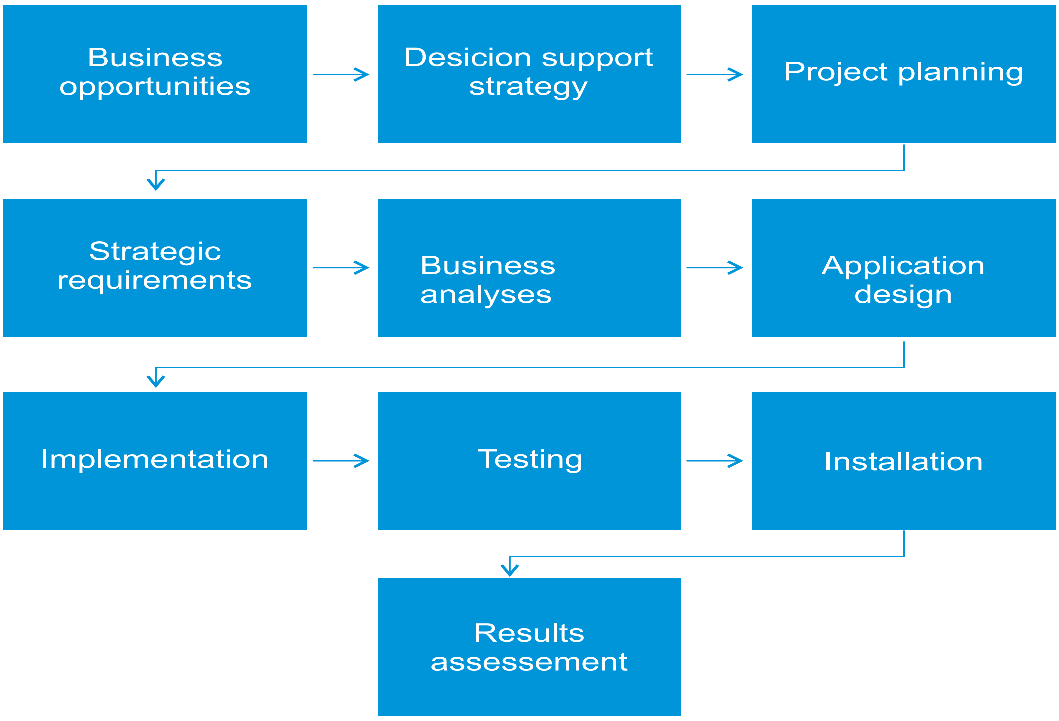

Fig. 2. Approach for development of BI applications in 10 steps

Table 1. Traditional solutions for web scraping

#	Solution	URL
1	Outwit Hub	https://addons.mozilla.org/en-US/firefox/addon/outwit-hub/
2	Web Scraper Chrome Extension	https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn
3	Spinn3r	https://www.spinn3r.com/
4	Fminer	http://www.fminer.com/
5	Dexi.io (CloudScape)	https://dexi.io/
6	ParseHub	https://www.parsehub.com/
7	OctoParse	http://www.octoparse.com/
8	Import.io	https://www.import.io/
9	Webhose.io	https://webhose.io/
10	Scrapinghub	https://scrapinghub.com/
11	VisualScraper	http://www.visualscraper.com/
12	80legs	http://80legs.com/

to an approach, proposed by some authors (Moss and Atre, 2004), a dynamic integrated environment that supports decision making management, cannot be established at once. Instead, the data and functionalities should be developed in several iterations, where each iteration leads to the occurrence of new requirements for the next iteration. Such an approach is presented in figure 2.

Traditional software solutions, including web scraping applications, are based on one of the examined approaches. Many authors, including Radoev (2016), mention in their research different components for collecting sets of data. According to PromptCloud and HKDC (2017), there are several traditional solutions for web scraping that are generally seen as successive. They are presented in Table 1.

Outwit Hub (2017) is a Firefox extension that can be easily downloaded from the Firefox add-ons store. This tool can automatically browse through pages and store the extracted information in a proper format. Outwit Hub offers a single interface for scraping tiny or huge amounts of data per needs. It is one of the simplest web scraping tools, which is free to use and offers the convenience to extract web data without writing a single line of code. Web Scraper (2017) is an alternative to Outwit Hub and it is a Google Chrome extension, that can be used for web scraping. It can scrape multiple pages simultaneously and even has dynamic data extraction capabilities. Web Scraper can also handle pages with JavaScript and AJAX. The downside to this solution is, that it doesn't have many automation features built in. Spinn3r (2017) is an application for scraping entire data from blogs, news sites, social media and RSS feeds. It can filter the data that it scrapes using keywords, which helps in weeding out irrelevant content. Spinn3r works by continuously scanning the web and updating their data sets. It has an admin console packed with features that can perform searches on the raw data. Fminer (2017) is probably one of the easiest to use web scraping tools. Its visual dashboard makes extracting data from websites simple and intuitive. It can scrape data from simple web pages or carry out complex data fetching projects that require proxy server lists, AJAX handling and multi-layered crawls. Dexi.io (2017) was formerly known as CloudScrape. It is a web-based scraping application that does not require any download, because it is a browser-based tool that can set up crawlers and fetch data in real-time. Dexi.io also supports scraping the data anonymously using proxy servers. ParseHub (2017) is a web scraping software that supports complicated data extraction from sites, using AJAX, JavaScript, redirects and cookies. It is equipped with machine learning technology that can read and analyze documents on the web to output relevant data. OctoParse (2017) is a visual web scraping tool that is easy to configure. OctoParse gives the option to run an extraction on the cloud and on a local machine. It can export the scraped data in TXT, CSV, HTML or Excel formats. Import.io (2017) is a web scraping application that offers a builder to form own datasets by simply importing the data from a particular web page and exporting the data to CSV. Import.io offers free apps for Windows, Mac OS X and Linux to build data extractors and crawlers, download data and sync with the online account. Webhose.io (2017) is a browser-based web application that provides direct access to real-time and structured data from crawling thousands of online sources. This web scraper supports extracting web data in more than 240 languages and saving the output data in various formats including XML, JSON and RSS. Scrapinghub (2017) is a cloud-based data extraction tool that fetches data. Scrapinghub uses Crawlera, which is a smart proxy rotator, supports bypassing bot counter-measures to crawl huge or bot-protected sites. Scrapinghub can convert an entire webpage into organized content. VisualScraper (2017) is a web data extraction software that can be used to collect information from the web. The software can extract data from several web pages and fetches the results in real time. VisualScraper can also export in various formats like CSV, XML, JSON and SQL. 80legs (2017) is a flexible web crawling tool that can be configured to fetch huge amounts of data along with the option to download the extracted data instantly. This web scraper claims to crawl more than sixteen hundred domains and it is used by PayPal.

The application of all of the examined software solutions for web scraping has both positive and negative aspects. There are also many more possibilities for use of web scraping software. However, we are now proposing a conceptual approach for the development of web scraping application that differs from the presented in the paper traditional approach for the development of applications and from the approach for the development of BI applications.

3. Milestones of the conceptual approach

The concept that is the subject of this paper aims to develop a web scraping application. Hence its implementation should have functional advantages over the traditional software solutions for web scraping. The application of business intelligence in the concept is determined by the need to serve large volumes of data that differ in structural terms. These differences should be overcome and the data should be subsequently stored and processed in a uniform way. The conceptual approach for the development of web scraping application is presented in figure 3. It includes the following steps that are achieved successively:

1. Definition of the data sources for web scraping – these are the wanted websites;
2. Analysis of the defined data sources – it is necessary to make sure that web scraping is possible for these websites;
3. Start of the data extraction process – this process includes saving the content of the websites in its original form;
4. Saving of the extracted data into a temporary database during the web scraping;

5. Transformation of the extracted data into a form that could possibly be loaded into a specific structure of a data warehouse;
6. Loading the transformed data into the data warehouse;
7. Now, with all the data available in the warehouse, it would be possible to search for a specific list of keywords and to send notifications for a presence of wanted keywords within the updated data warehouse;
8. Cleaning of the temporary resources that have been used for web scraping, as only the transformed data loaded into the warehouse will be used.

The proposed conceptual approach for the development of web scraping application in eight steps requires the following technological conditions:

- the high frequency of data update by the use of selected sources that should be

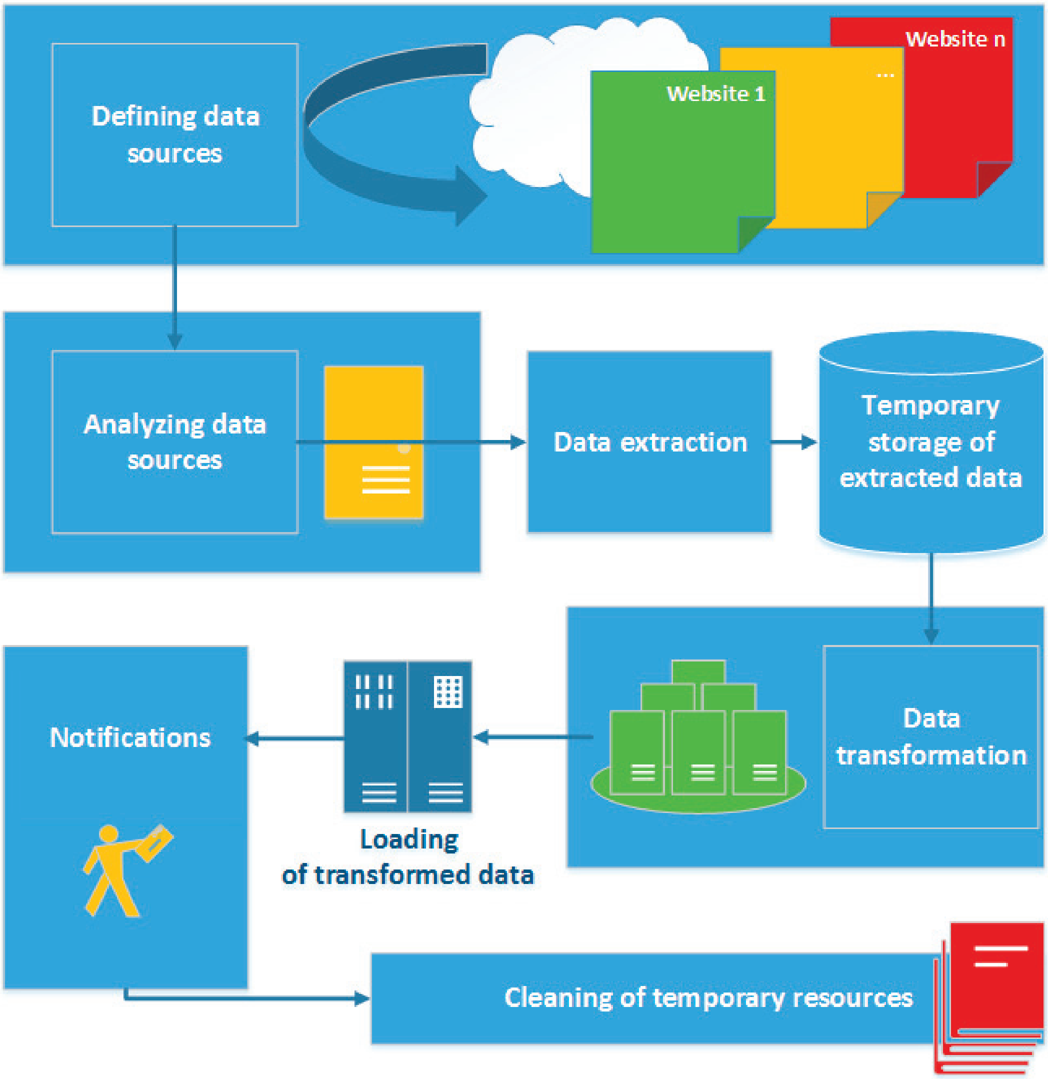

Fig. 3. Approach for development of web scraping application in 8 steps

scraped through an embedded intelligent algorithm for updating the data in short intervals and through embedded algorithms for limited data extraction for load balance at the data provider;

- precise categorization of the data by using classified sources that have their own categorization and by associating their categorization with the categorization of the web scraping application;
- large volume of indexed data, by means of a possibility of an unlimited expansion of new sources;
- Ease of use of the application, by means of implementation of web-based user interface, accessible over the Internet with just a limited number of functional tools;
- Lack of necessary level of qualification of specialists, who will work with the web scraping application, by means of not requiring a high level of training of users;
- specialized algorithm for a particular group of web sources, by means of usage of uniform model for describing and transforming different data structures into a uniform format;
- possible notifications, based on simple and complex searches, by means of combining many simple searches into the complex one with applying an optimized algorithm for quick search;
- option to filter the notifications by date or period by means of organizing data storage by date and setting analysis for periods of time;
- option to display similar results in the notifications by using a module for finding similar results within stored data;
- option to index data in different languages by means of customizing the language of the data of each source;
- Possible display of popular results in the notifications by using a module for finding popular results within stored data, based on criteria of importance.

The proposed requirements allow us to identify the strengths, weaknesses, opportunities and threats in introducing the use of the proposed conceptual approach for the development of web scraping application. The strengths of the proposed conceptual approach can be summed up as follows:

- Providing a large volume of information for analysis;
- Choosing a variety of online sources for data extraction;
- Availability of informal information for management levels;
- Specially developed software solution;
- Availability of good price – quality ratio in providing the necessary reliability of implemented technologies.

Weaknesses of the proposed conceptual approach can be summed up as follows:

- Necessary level of knowledge, skills and competencies of professionals who will work with the application;
- Dynamically changing range of sources of information and different structure of different sources;
- Low confidence of informal information.

Opportunities of the proposed conceptual approach can be summed up as follows:

- use of application with different sources of information;
- web-based architecture, based on a special methodology for extraction, transformation and loading of online information sources;
- continuous analysis of the conceptual line between risk factors in decision making management and common software model;
- real-time information based on an analysis of possible sources of informal data and their importance in the decision making process;
- Dynamic creation of datasets.

Threats of the proposed conceptual approach can be summed up as follows:

- Time required for the extraction of information;
- Creating large volumes of data files;

- The existence of different structures for different sources;
- Greater complexity in using the application.

Table 2 presents assessments in range from 1 to 6 of the presented conceptual approach in comparison to the approaches of traditional software solutions for web scraping, given subjectively by independent

Table 2. Assessments of traditional solutions for web scraping and the proposed conceptual approach, given subjectively by independent experts

Strengths			Weaknesses		
№	Description	Assessment traditional solutions – proposed conceptual approach	№	Description	Assessment traditional solutions – proposed conceptual approach
1	Providing a large volume of information to be analyzed	5 – 6	1	Necessary level of knowledge, skills and competencies of professionals, who will work with the application	3 – 2
2	Choosing a variety of online sources for data extraction	5 – 6	2	Constantly changing range of sources of information and different structure of different sources	2 – 2
3	Availability of informal information for management levels	4 – 3	3	Low confidence of informal information	5 – 4
4	Specially developed software solution	5 – 6			
5	Availability of good price – quality ratio in providing the necessary reliability of implemented technologies	2 – 4			
		4,20 – 5,00			3,33 – 2,67
Opportunities			Threats		
№	Description	Assessment traditional solutions – proposed conceptual approach	№	Description	Assessment traditional solutions – proposed conceptual approach
1	use of application with different sources of information	5 – 2	1	Time required for indexing information	6 – 4
2	web-based architecture, based on a special methodology for extraction, transformation and loading of online information sources	5 – 6	2	Creating large volumes of data files	1 – 1
3	continuous analysis of the conceptual line between risk factors in making management decisions in crisis situations and common software model	2 – 4	3	The existence of different structures for different sources	1 – 2
4	real-time information based on an analysis of possible sources of informal data and their importance in the decisions making management	2 – 5	4	Greater complexity in using the application	4 – 3
5	Dynamic creation of datasets	1 – 2			
		3,00 – 3,80			3,00 – 2,50

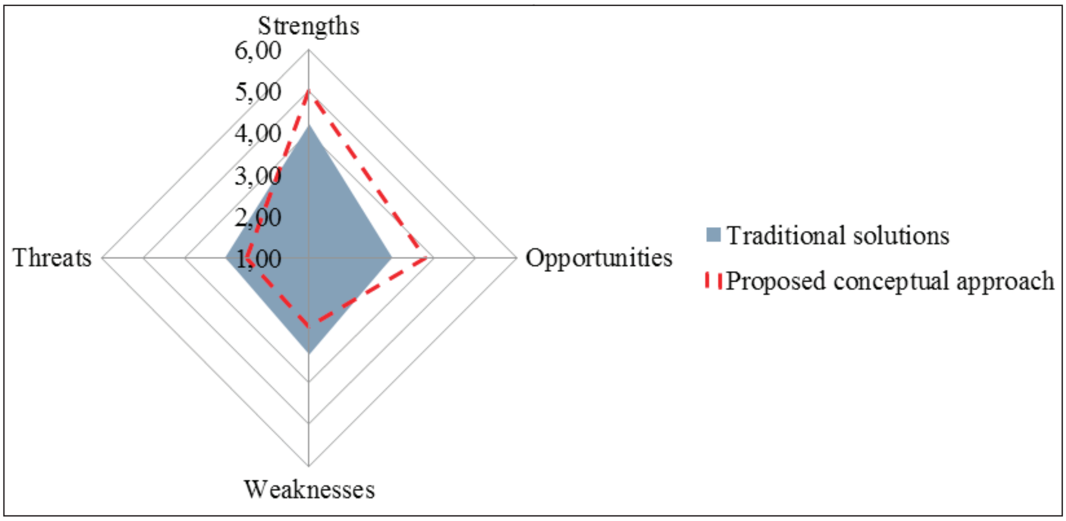

Fig. 4 SWOT analysis of the proposed conceptual approach

experts to all of the defined strengths, opportunities, weaknesses and threats.

Figure 4 offers a graphic presentation of the SWOT analysis of the conceptual approach compared to the traditional software solutions for web scraping, through assessments from Table 2.

It can be concluded that the use of a web scraping application, developed under the proposed concept would possibly increase of strengths and opportunities, and reduce weaknesses and threats.

4. Conclusion

The paper highlighted some of the most important stages and components in the development of BI applications. The theoretical overview showed that web scraping could be considered as part of data mining and data mining itself is related to business intelligent systems. It can be concluded that the development of web scraping applications is a fairly complicated process. This paper is focused on the scraping of information from web sources. A key role in this area plays the data extraction process, as well as data transformation and data loading into a warehouse. The paper proposed a conceptual approach for the development of web scraping application. It can be concluded that the proposed model is implementable and could serve as a basis for the development of an effective software solution for web scraping. The practical importance of such a solution can be found in tracking information about online presence of a company by specific keywords in selected area of web sources.

References

Eckerson, W., 2003. Understanding Business Intelligence, TDWI

Eckerson, W. and Howson, C., 2005. Enterprise Business Intelligence: Strategies and Technologies for Deploying BI on an Enterprise Scale, The Data Warehousing Institute (DWHI)

Kirilov, R., 2016. Software solutions for managing projects co-financed under the european union's operational programs, *Business Management*, Issue 3, ISSN 0861-6604.

Kirilova, K., 2014. Methodological issues in development of public electronic services, *Economic and Social Alternatives*, Issue 4, ISSN 1314–6556

Loshin, D., 2013. Business Intelligence – The Savvy Manager's Guide, Second Edition

Marzovanova, M., 2015. Intelligent Tagging and Search as a Fully Automated System, *Economic Alternatives*, Issue 2, ISSN 1312-7462

Mihova, V., 2015. Methods of Using Business Intelligence Technologies for Dynamic Database Performance Administration, *Economic Alternatives*, Issue 3, ISSN 1312-7462

Moss, L. and Atre, S., 2004. Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications, Addison Wesley Information Technology Series

Power, D., 2007. A Brief History of Decision Support Systems, www.dssresources.com

Quarles van Ufford, D., 2002. Business Intelligence: The Umbrella Term, VU Amsterdam

Radoev, M., 2016. Comparison of Tools for Collecting Information About Query Performance in Microsoft SQL Server, *Economic Alternatives*, Issue 2, ISSN 1312-7462

Stefanov, G., 2015. Methods for Heterogeneity Detection During Multi-Dimensional Data Mart Integration, *Economic Alternatives*, Issue 1, ISSN 1312-7462

Stefanova, K., 2008. Factors and Main Directions for Business Intelligence Systems Design and Development, Yearbook of UNWE

Stefanova, K. and Yordanova, S., 2017. Knowledge Discovery from Unstructured Data using Sentiment Analysis, *Economic and Social Alternatives*, Issue 1, ISSN 1314–6556

http://80legs.com/, [Online, Accessed March 2017]

https://addons.mozilla.org/en-US/firefox/addon/outwit-hub/, [Online, Accessed March 2017]

https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn, [Online, Accessed March 2017]

https://dexi.io/, [Online, Accessed March 2017]

http://paper.ijcsns.org/07_book/201103/20110324.pdf, [Online, Accessed March 2017]

https://scrapinghub.com/, [Online, Accessed March 2017]

https://webhose.io/, [Online, Accessed March 2017]

http://www.fminer.com/, [Online, Accessed March 2017]

http://www.hongkiat.com/, [Online, Accessed March 2017]

https://www.import.io/, [Online, Accessed March 2017]

http://www.octoparse.com/, [Online, Accessed March 2017]

https://www.parsehub.com/, [Online, Accessed March 2017]

https://www.promptcloud.com/, [Online, Accessed March 2017]

https://www.quora.com/, [Online, Accessed March 2017]

https://www.spinn3r.com/, [Online, Accessed March 2017]

http://www.visualscraper.com/, [Online, Accessed March 2017]