ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **19**, Numbers 1-2, 2016, 65–77

Data Driven Decision Support Systems: An Application Case in Labour Market Analysis

Claudiu BRANDAS[1], Ciprian PANZARU[2], Florin Gheorghe FILIP[3]

[1]West University of Timisoara, Department of Business Information Systems, Timisoara, Romania
[2]West University of Timisoara, Department of Sociology, Timisoara, Romania
[3]Romanian Academy, Information Science and Technology Section, Bucharest, Romania

E-mail: claudiu.brandas@e-uvt.ro

Abstract. The paper reviews recent advances in the domain of computer supported data driven decision-making and proposes a conceptual framework for the construction and implementation of a class of practical *decision support systems* named *Labour Market Decision Support Systems* (LM-DSS) to be applied to labour market analysis. LM-DSS are meant to support the analysis the information and knowledge about dynamics of job vacancies and skills, concentration on territorial units and economic activities, most requested skills, education and training programmes, unemployment and labour force statistics, wage information and projections. The experimental results are presented in the context af academia labour market.

Key-words: big data, business intelligence& analytics, data mining, web mining, labour market.

1. Introduction

A *Decision Support Systems* (DSS) can be viewed as "an anthropocentric and evolving information system which is meant to implement the functions of a human support system that would otherwise be necessary to help the decision-maker to overcome his/her limits and constraints he/she may encounter when trying to solve complex and complicated decision problems that count"[14]. The development of Internet-connected devices, the dramatic volume increase of stored data sets and development of *Information and Communication Technologies* (ICT) have led to the new application fields for DSS. So far, *Business Intelligence and Analytics* (BI&A) and the related domain of *Big Data* (BD) analytics have been all deployed in the field of banks, insurances, health, industry, transport, GIS and IT.

Recently, BI&A and BD have also become increasingly useful in the labour market analysis [1], [19], [20] especially because of the existence of huge volumes of online data in the form of job ads and job demands. Early works related to BD and the labour market addressed the prediction of the unemployment phenomenon [4]. Other studies were focused on correlations between Google searches for the word "unemployment" and unemployment data from official statistics [17]. Based on this information, unemployment rates could be forecast [23].

In this context, BD has become an important source for information about the workforce. This information concerns mainly the job websites. The history of portals on the demand and supply of labour already spans the period of 20 years. Since 1995, with the advent of the first website of its kind – *CareerBuilder* – their number has increased significantly. It is difficult to estimate how many such portals are active at the moment or the number of their users. The number of visitors, however, is significant. The number of unique monthly visitors for the top 10 job sites in the world counted, at the beginning of 2015, about 145 million. Also, we frequently find labour market information on social networks. They accumulate by far the most users. For example, *LinkedIn* had about 350 million users in December 2014 (`www.statista.com`). These data show that the Internet and related technologies such as social networks and web technology play an increasingly important role in configuring the supply and demand on the labour market. Basically, huge volumes of data are produced in the form of job ads and job applications. The analysis of this tremendous volume of data is an immense challenge for the socio-economic but also for the political environment. Thereby, the online job portals have become an important and valuable data source for a Labour Market Information System.

On the other hand, online job portals provide useful information for designing and implementing new models and tools for innovating Labour Market Information System and Labour Market Intelligence and Services [12]. *Labour Market Information* includes descriptive data such as statistics or survey results. *Labour Market Intelligence* includes analysis, interpretation, conclusions and policy recommendations. Labour Market Intelligence means that the information has already been analysed and reduced to the important and relevant aspects for decision-making [22]. Because usually Labour Market Information and Labour Market Intelligence are used in a similar way, to avoid confusion, a solution is to use LMI for both information and intelligence in the labour market [34]. We will consider that an LMI provides quantitative and qualitative information and intelligence, to help labour market agents in decision making.

In this article, we present a framework for the construction and implementation of a Labour *Market Decision Support System* (LM-DSS) based on web mining, data mining and *spatialization* (spatial mapping of results) with application case in labour market analysis. LM-DSS is meant to provide information and knowledge for decision-making processes carried out by diverse users, such as job seekers, researchers, consultants or business and governmental policy makers, educators and training institutions, public and private employment agencies and social services.

2. Computer Supported Data-driven Decision-Making

At the moment, Big Data is getting ever more attention, generating new opportunities in processing and analysing unstructured data within decision-making processes. The concept of Big Data is defined in various ways. The survey contained in this section is based on the presentation made in a forthcoming book (Filip F.G., Ciurea C., Zamfirescu C.B., *Computer Supported Collaborative Decision-making*, Springer).

A well-known definition [21] considers Big Data as a combination of three characteristics: *volume, velocity*, and *variety*. A fourth dimension, called *veracity*, was subsequently added [35], to indicate that the data are accurate and true. Madden [24] provides a simple but intuitive definition. Big Data represents a data set that is: a) too big (it has a large scale and various sources); b) too fast, (and needs quick processing); c) too hard (and cannot be managed by traditional database management systems). Regardless of definition, Big Data is commonly referring to the massive amounts of unstructured data and the information that can be retrieved from it. In this way, Big Data also include the technology used for extracting and analysing useful information.

Big Data application fields are numerous and diverse [7], [8], [9],[15], [25], [30], [31]. For example, the web-based companies, as Amazon.com, e-Bay, and Google, intensively use Big Data applications. Financial companies, manufacturing firms and other organisations from public and private sectors are also taking advantage of the immense volumes of stored data and available technologies. Some references note that the world is increasingly driven by insight derived big data [29],[33].

The new Big Data era and development of Big Data Analytics brought to the fore the well established term of *Business Intelligence* (BI). B I is viewed both in industry and academia since early 1990s as a tool meant to provide the possibility to transform raw data into valuable insights. In 1989, H. Dresner coined the term BI to name, "all technologies that help the business make a decision on facts" [27]. According to Gartner Group, Business Intelligence (BI) is defined as a software platform that delivers a set of capabilities organised into three classes of functionalities [9]:

- *Integration,* that includes: BI infrastructure, metadata management, development tools and enabling collaboration;

- *Information delivery*, that includes: reporting, dashboards, ad-hoc query, Microsoft Office integration, search-based BI, and mobile BI;
- *Analysis*, that includes OLAP (Online Analytical Processing), interactive visualization, predictive modeling, data mining and scorecards.

To highlight the key analytical part of BI, Davenport [10] introduced the term *business analytics* (BA) to include [big] data analytics, text analytics, web analytics, network analytics, mobile analytics . Thus, the original term of BI became BI&A.

Chen et al. [9] identify three generations of BI&A:

- BI&A 1.0, adopted by industry in the 1990s, is characterised by the predominance of structured data which are collected by existing legacy systems and stored and processed by RDBM (Relational DataBase Management Systems). The majority of analytical techniques were using well established statistical methods and data mining tools developed in the 1980s. The ETL ("Extract, Transformation, and Load") of data warehouses, OLAP (On Line Analytical Processing) and simple reporting tools are commonly met in BI&A 1.0.
- BI&A 2.0 was the next stage. It was triggered by the emergence of Internet and Web technologies in particular text mining and web search engines, social networks and the development of the e-commerce in the early 2000s.
- BI&A 3.0 is a new stage of development. It is characterised by the massive usage of mobile devices and applications. The main characteristic feature of BI&A 3.0 is the intensive data collection enabled by the "Internet of Things".

In the Big Data context, BI&A methods and corresponding tools spread very rapidly and are used in many fields. Moreover, technology and analytics support represent the basis for data-driven decision making. Brynjolfsson et al [6] developed a measure of the usage of *data-driven decision* making that was meant to capture business practice surrounding the collection and analysis of external and internal data of an enterprise. Power [28] notices that, at present, data-driven decision support is used for a broad spectrum of purposes including operational and strategic business intelligence queries, real-time supervision and performance monitoring, and CRM (Customer Relationship Management).

Given the complexity of Big Data (BD) and its potential usage in decision making activities of the present day, it is reasonable to consider developing DSS that are based on artificial intelligence technologies for knowledge discovery. This development supports not only the reporting but also the automatic learning for adaptive decision-making. Considering the data mining capabilities, the authors think that integrating this technology within DSS creates great potential in optimising and updating the models and knowledge base of the system. Implementing data mining methods in DSS application in a BD context is an essential element of the BI&A. Such integration can provide effective support for decision-makers in the analysis of large and heterogeneous data. By using data mining techniques, the system can uncover hidden patterns within large amounts of data. These hidden patterns can be used to predict future behaviour based on the given data set [13].

At the same time, recent business models and technological have developments led to the concept of *Labour Market Analytics* (LMA) [11]. The majority of new LMA systems can be noticed in the software and recruiting companies. For example, the *Labour Market Intelligence Predictive Analytics Platform* (LMI-PAP) of *Focal HR* (`http://focalhr.com/`) is a "granular reporting cloud-based intelligence and analytics software product, focused on labour market metrics". A similar example is reported by Javed at al [18] who describe a "machine learning-based semi-supervised job title classification system for the online job recruitment domain" released by CareerBuilder.com (`http://www.careerbuilder.com/`). The above information tools use sophisticated technologies to convert unstructured data into meaningful analytics, to allow the implementation of a *real-time data collection and labour market information* (RT-LMI) service from web portals. The objective of this type of analytics is to diminish of the *Time to Answer* (TtA) and *Data-to-Decision* (DtD). The solution is to provide relevant data in real time to make a decision. The value extraction from big data volumes is carried out by using Business Intelligence & Analytics. The analysis includes OLAP (Online Analytical Processing), interactive visualisation, predictive modeling, data mining and scorecards.

3. LM-DSS Framework Development

The LM-DSS model could be described not only as an innovative but also as a complex solution for providing information concerning highly important economic elements related to labour force and employment. The innovation resides in the creating a systemic framework of explanation regarding the context above, taking into consideration vectors such as processes, tools, and outputs. At the same time, the model complexity derives from the number of elements it focuses on. As such, it has to be said that LM-DSS is meant to provide information and knowledge about labour market dynamics, such as job vacancies and skills, their concentration on territorial units and industry sectors, the most requested skills, education and training programs, unemployment and labour force statistics, wage information and projections. The conceptual model of the framework is presented in Fig. 1.

One of the main ways to get information about labour market is based on the extraction of knowledge from large unstructured data collection [5]. Web mining and data mining of online job ads could be a valuable way to get real-time labour market information. The extraction and processing of data through web mining and knowledge discovery through data mining represent the effort to understand, analyse and, possibly, use an enormous amount of available data

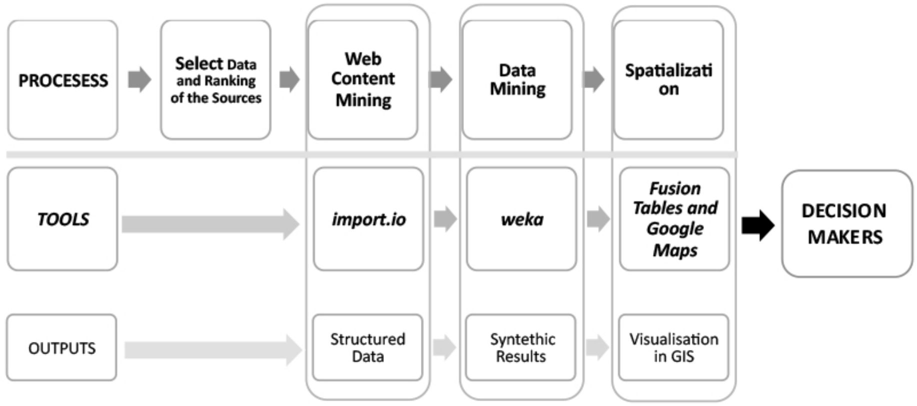

Fig. 1. Framework of a Labour Market Decision Support System (LM-DSS).

in an effective manner[13].

In the research, five processes to get knowledge about the labour market are run as follows.

3.1. Process #1: Select data and rank the sources

In the testing experiment carried-out, we only focused on data about academic job vacancies, extracted from the academic job portals. We selected the sources taking into account the number of available jobs on site. The main source was `www.academicpositions.eu`, one of the biggest network for academics in Europe having separate portals for each country such as `www.dk.academicpositions.eu`, `www.ge.academicpositions.eu`, `www.it.academicpositions.eu`, `www.fr.academicpositions.eu` and so on. Information was also extracted from other portals such as: `www.academicjobseu.com`, `www.universitypositions.eu`, `www.academics.com`, `www.jobs.ac.uk/categories/academic-jobs-europe`, `www.eurosciencejobs.com/jobs/academic`, `www.unijobs.com`, `https://www.akadeus.com`, and `www.careeredu.eu`.

3.2. Process #2: Web Content Mining

The second process consisted in extracting and structuring data using web content mining. Therefore, we considered *Import.io* (`https://www.import.io/`) as a software tool that allows the extraction and conversion of semi-structured data into structured data. The collected data can be exported as CSV (Comma-Separated Values), Excel, Google Sheets or JSON (JavaScript Object Notation). Consequently, we scraped data from the selected sites in the first process. In this stage, the duplicates were eliminated. The outputs of this process were the collection of tables with structured data.

3.3. Process #3: Data Mining

In the third process, we employed data mining technique on data which was epreviously extracted and structured. Two categories of data mining techniques can be used: predictive and descriptive ones. *Predictive data mining* is meant for analysing historical data sets and constructing models based on discovered hidden patterns. *Descriptive data mining* enables analysing data sets to find the descriptive patterns. Within the process, both the predictive and descriptive methods can be used.

More specifically, data was processed by using data mining clustering techniques [3] and simple unsupervised k-means clustering algorithm [16], [32]. The method involves starting with k values (random) and building clusters according to them. It is an algorithm for classifying or to grouping data objects based on their attributes into k clusters by minimising the Euclidean distance between each data objects and clusters centroids (mean vectors).

For data mining through clustering, we used WEKA (Waikato Environment for Knowledge Analysis) application. It is a machine learning software for data mining processes and contains tools for data pre-processing, classification, regression, association rules, clustering and visualisation (`http://www.cs.waikato.ac.nz/ml/weka/`).The outputs of data mining process are the synthetic results representing the clusters and each cluster's centroids.

3.4. Process #4: Spatialization

In the fourth and last process of our application, for a better visualisation, the results are spatialized (placed on the map). In this matter, one can use *Google Fusion Tables* (`http://support.google.com/fusiontables/answer/2571232?hl=en`), which is an experimental data visualisation web application that allows the gathering, visualisation, and sharing of data tables using Google Maps.

4. Experimental Results

In the *Web Content Mining* process using Import.io, applying the Extractor function on the research and academic jobs websites, we obtained the following *research_and_academic_job_data_set* table with the following attributes: job title, location.

In the *Data Mining Analysis* process, we used WEKA and simple k-means clustering algorithm with the following parameters: a) *distance function*: EuclideanDistance; b) *maxIterations*: 500; c) *numClusters*: 10; d) *seed*: 10 We got 10 clusters for the *research_and_academic jobs_data_set*. There was one instance for each cluster representing the cluster centroid (mean vectors for each cluster) (Table 1).

Cluster #6 contains the majority of instances (18%). It is followed by the Cluster #4, which contains 14% of the total number of instances. In Cluster

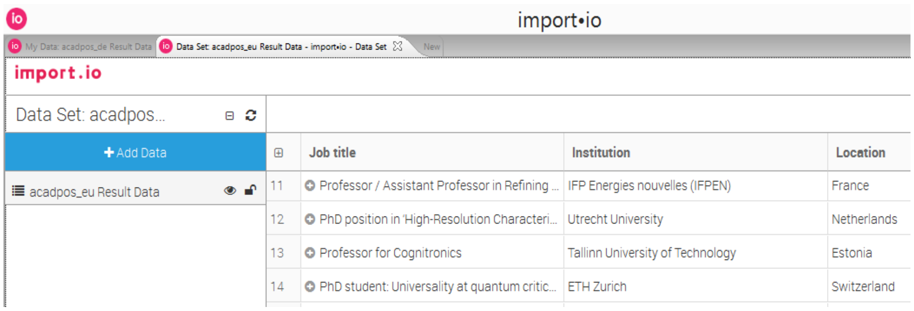

Fig. 2. Web content mining using *Import.io*.

#0 and #1 there are the fewest instances (5%). Cluster analysis using *WEKA* shows the densities of the most academic required jobs by each country (Fig. 3).

Table 1. Academic jobs (*W*EKA) clustering results

Attribute	Cluster #0	Cluster #1	Cluster #2	Cluster #3	Cluster #4
Location	Luxembourg	Germany	France	Netherlands	United States
Job title	Postdoc in Mathematics	Research fellowships	Research department chair	Senior Lecturer	Clinical instructor/clinical professor

Attribute	Cluster #5	Cluster #6	Cluster #7	Cluster #8	Cluster #9
Location	United States	Sweden	Australia	United States	United States
Job title	Lecturer	Lecturer	Research Assistant	Research scientist/engineer	Research Associate

Clustered Instances

0	492 (5%)	5	1354 (13%)
1	490 (5%)	6	1875 (18%)
2	931 (9%)	7	955 (9%)
3	1242 (12%)	8	947 (9%)
4	1521 (14%)	9	726 (7%)

The results obtained by using WEKA can be analysed with respect to at least two aspects: a) the diversity of the job demand, and b) quantity. The United States and Sweden stand out as the countries with the largest and , at the same time the most diversified offer of jobs available. We find that the most diverse job offers are in the United States, a fact which corresponds to multiple research opportunities, due to a large number of research institutes. It should be noticed that four clusters out of ten (namely clusters #4, #5, #8, and #9) are placed in the United States. We have also noticed a large number of academic vacancies in the Netherlands. In Europe, a significant number of academic vacancies are in France, Germany and Luxembourg. Australia is another country with an important academic job demand.

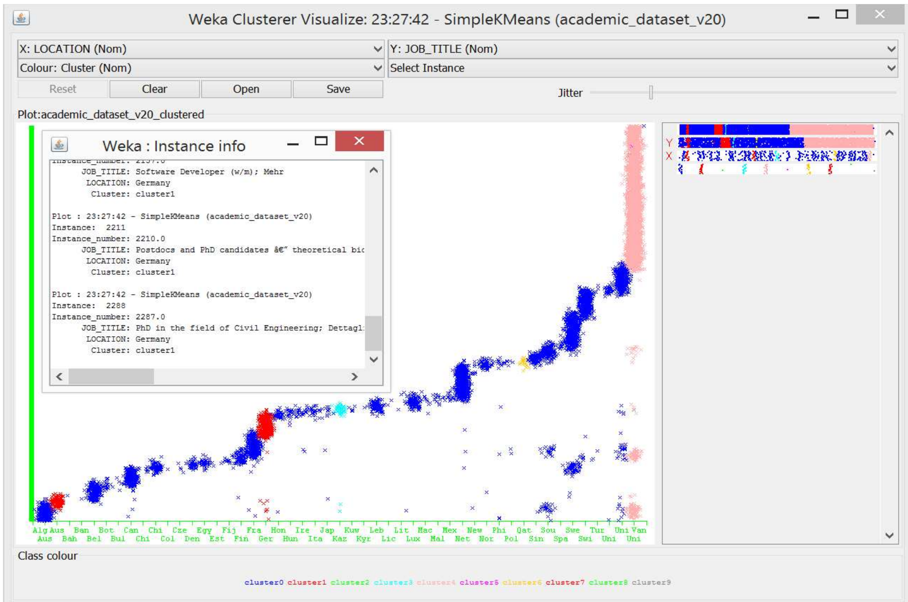

Fig. 3. Academic job density.

Fig. 4. Data spatialization of the academic jobs.

In the process of *Data Spatialization*, Fusion Tables and Google Maps were used for a better visualisation of the results. Figure 4 shows the spatialized data for the academic labour market demand.

The decision makers could analyse the above results, also taking into account supplementary information, such as academic domain, expenditure on education, brain drain phenomenon, unemployment, and so on. Determining the labour force needs and the type of employee are important issues not only to eliminate disfunctions in the labour market but also to develop strategic plans. The results provided by the system could be useful for a wide category of decision-makers. Thereby, they could be used by local and central public administration to develop labour market policies, to education and training providers to calibrate their supply to the market needs, to jobseekers to find a job, to economic players to make an investment, relocation, development, etc. They could also be used to identify decision trees for modeling career paths.

The next level in analysing of the Labour Market should be based on the idea of *Smart Labour Market* (SLM), which integrates in an intelligent way the actors (the decision-makers), technology and information (Fig. 5).

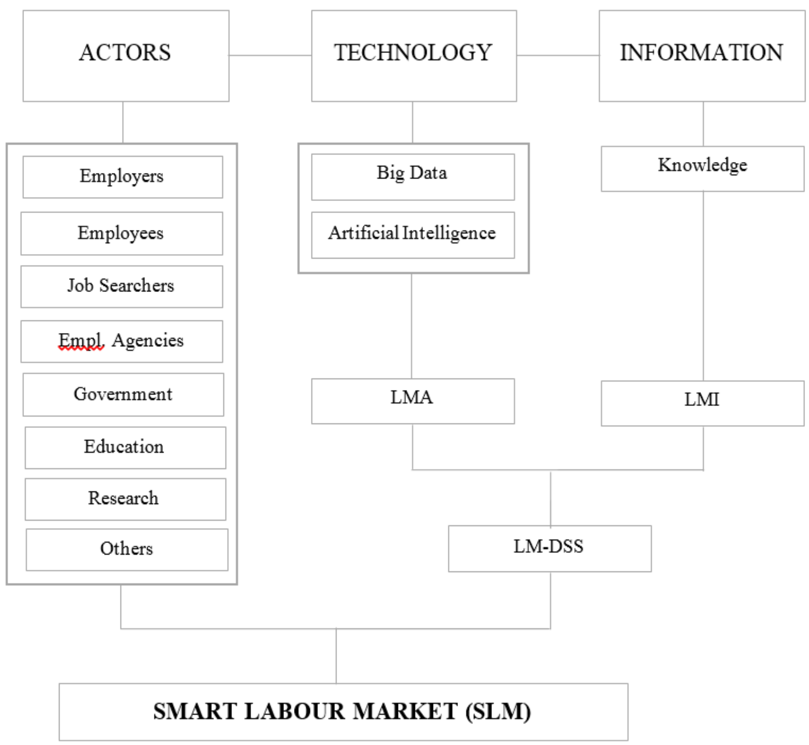

Fig. 5. Conceptual map of a Smart Labour Market.

The main characteristics of the Smart Labour Market for decision makers are personalised information, meaningful, adaptive and relevant. One may view Labour Market Decision Support System as a part of the Smart Labour Market.

5. Conclusions

In this paper, we showed that big data analysis and extraction of knowledge from data collection could be successfully used to analyse labour market. In this context, the Labour Market Decision Support System can be a solution for advanced Labour Market Analysis.

Regarding the data processing, several strengths and weaknesses were noticed. Thus, one of the strengths is the capability of processing large amounts of data, which, once collected and sorted, identification of patterns by data mining are enabled without the need of using human and logistical resources for data collection through field research. Since we have often had to deal with unstructured or semi-structured data, the solution is to use the software applications, such as web scrapers, web spiders, crawlers or web data extractors. These applications are often specific programs specially designed for a particular data source and have the ability to retrieve structured data.

A weakness in the privacy policies of websites has been noticed. Some websites allow extraction of data while others are very strict in this way. Probably, the most important limitation is generated by the diversity of sources of information and attributes values. This explains the lack of structural and semantic compatibility of data which occurs frequently. In the future, the ontologies, web semantics and web 3.0 could be used as a adequate solution to solve this problem.

References

[1] ASKITAS N., ZIMMERMANN K.F, *The Internet as a data source for advancement in social sciences.* IZA Discussion Papers 8899. ftp: ftp.iza.org/dp8899.pdf. Accessed 10 Dec 2015.

[2] BANGE, C., GROSSNER, T. S. & JANOSCHEK, N., *Big Data Survey Europe: Usage Technology and Budgets in European Best Practice Companies.* BARC Institute, Würzburg, 2013.

[3] BERKHIN P., *A survey of clustering data mining techniques.* In: KOGAN J., NICHOLAS C., TEBOULLE M. (eds) *Grouping multidimensional data. recent advances in clustering.* Springer Berlin Heidelberg, Berlin, 2006, pp. 25-71.

[4] BOLLIER D., *The promise and peril of big data.* Aspen Institute, Washington DC, 2010.

[5] BRANDAS C., PANZARU C., *Analysing of the labour demand and supply using web mining and data mining in Romania.* In: LARSEN C., RAND S., SCHMID A., MEZZANZANICA M., DUSI S. (eds) *Big data and the complexity of labour market policies: new approaches in regional and local labour market monitoring for reducing skills mismatches.* Rainer Hampp Verlag, Munchen, 2015, pp. 71-86.

[6] BRYNJOLFSSON E., HITT L.M., KIM, H.H., *Strength in numbers: How does data-driven decisionmaking affect firm performance?* Available at SSRN 1819486, 2011.

[7] BUGHIN J., CHUI M., MANYIKA J., *Clouds, big data, and smart assets: Ten tech-enabled business trends to watch*. McKinsey Quarterly, 56(1), 2010, pp. 75-86.

[8] CHEN C.P., ZHANG C.Y., *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data*. Information Sciences, 275, 2014, pp. 314-347.

[9] CHEN H., CHIANG R.H., STOREY V.C., *Business Intelligence and Analytics: From Big Data to Big Impact*. MIS quarterly, 36(4), 2012, pp. 1165-1188.

[10] DAVENPORT, T. H., Competing on analytics. Harvard Business Review, 84(1), January, 2006, pp. 98-107.

[11] DORRER J., *Analyzing labor markets in real time*. In: SCHNEIDER M, DEANE KC (eds). *The university next door: what is a comprehensive university, who does it educate, and can it survive?* Teachers College Press, Amsterdam, 2015, pp. 149-172.

[12] DUSI S., MERCORIO F., MEZZANZANICA M., *Big data meets web job vacancies: trends, challenges and development directions*. In: LARSEN C., RAND S., SCHMID A., MEZZANZANICA M., DUSI S. (eds) *Big data and the complexity of labour market policies: new approaches in regional and local labour market monitoring for reducing skills mismatches*. Rainer Hampp Verlag, Munchen, 2015, pp. 3-44.

[13] FAYYAD U., PIATETSKY-SHAPIRO G., SMYTH P., *From data mining to knowledge discovery in databases*. AI magazine 17(3), 1996, pp. 37-54.

[14] FILIP, F. G. (2008). *Decision support and control for large-scale complex systems*. Annual Reviews in Control, 32(1), 2008, pp.61-70.

[15] FILIP F.G., HERRERA-VIEDMA E., *Big data in the European Union*. The Bridge, 44(4), 2014, pp. 33-37.

[16] HARTIGAN J.A., WONG M.A., *Algorithm AS 136: A k-means clustering algorithm*. Applied Statistics 28(1), 1979, pp. 100-108.

[17] HILBERT M., *Big Data for Development: From information-to knowledge societies*. SSRN doi: 10.2139/ssrn.2205145, 2013.

[18] JAVED F., MCNAIR M., JACOB F., ZHAO M., *Towards a job title classification system*. In Proceedings of the 1st Workshop on Web-Scale Classification: Classifying big data from the web, New York City, February 2014.

[19] KREIBICH R., *Von big zu smart – zu sustainable?*. `http://www.bpb.de/apuz/202244/von-big-zu-smart-zu-sustainable?p=all`. Accessed 10 Dec 2015.

[20] KUREKOVÁ L.M., BEBLAVY M., THUM A.E., *Using Internet data to analyse the labour market: a methodological enquiry*. IZA Discussion Papers 8555, 2014, pp. 3-25.

[21] LANEY D., *3D Data Management: Controlling Data Volume, Variety and Velocity*. Meta Group. `http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf`, 2001, Accessed 11 Dec 2015.

[22] LANTRA, *Northern Ireland Labor Market Intelligence*. Lantra House, Warwickshire, 2005.

[23] LARSEN C., RAND S., *Introduction*. In: LARSEN C., RAND S., SCHMID A, MEZZANZANICA M., DUSI S. (eds) *Big data and the complexity of labour market policies: new approaches in regional and local labour market monitoring for reducing skills mismatches*. Rainer Hampp Verlag, Munchen, 2015, pp. 11-30.

[24] MADDEN S., *From databases to big data*, IEEE Internet Computing, May-June: 4-6, 2012.

[25] MANYIKA J., CHUI M., BROWN B., BUGHIN J., DOBBS R., ROXBURGH C., et al., *Big Data: The Next Frontier for Innovation, Competition and Productivity*, McKinsey Global Institute, May, 2011.

[26] MORISSETTE L., CHARTIER S., *The k-means clustering technique: General considerations and implementation in Mathematica*. Tutorials in Quantitative Methods for Psychology, 9(1), 2013, pp. 15-24.

[27] NYLUND, A. I., *Tracing the BI timely tree. Knowledge Management*, 60, July, 1999.

[28] POWER D.J., *A brief history of decision support systems*. DSS Resources. COM. Retrieved from http://DSSResources.COM/history/dsshistory.html, 2007.

[29] POWER D.J., *Creating a Data-Driven Global Society*. In Reshaping Society through Analytics, Collaboration, and Decision Support (pp. 13-28). Springer International Publishing, 2015.

[30] SAGIROGLU S., SINANC D., *Big data: a review*. Proceedings of the International Conference on Collaboration Technologies and Systems (CTS '13); May 2013; San Diego, Calif, USA. IEEE; 2013, pp. 42-47.

[31] SHI Y., *Big data: History, current status, and challenges going forward*. Bridge 44(4), 2014, pp. 6-11.

[32] TEKNOMO K., *K-Means Clustering Tutorials*. `http://people.revoledu.com/kardi/tutorial/kMean/`, 2007, Accessed 17 Dec 2015.

[33] WEF [World Economic Forum], *Unlocking the Value of Personal Data: From Collection to Usage*. Geneva, 2013.

[34] WOODS J.F., O'LEARY C.J., *Conceptual framework for an optimal labour market information system: final report*. Upjohn Institute Technical Report 07-022. doi:10.17848/tr07-022, 2006.

[35] ZIKOPOULOS P., EATON C., *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media, 2011.