USING DATA MINING TECHNIQUES TO CREATE AN AUTOMATED MODEL THAT MAKES COMPARISONS BETWEEN MARKET DEMANDS AND UNIVERSITY CURRICULA

M.Sc. Ylber Januzaj PhD (C).[1], Prof. Assoc. Artan Luma PhD.[1], Prof. Assoc. Azir Aliu PhD.[1], Prof. Assoc. Besnik Selimi PhD.[1], Prof. Assoc. Bujar Raufi PhD.[1], Prof. Assoc. Halil Snopce PhD.[1], Prof. Ass. Vehbi Ramaj PhD.[2]
Faculty of Contemporary Sciences and Technologies – South East European University, Macedonia [1]
Faculty of Economics – University of Peja "Haxhi Zeka", Kosovo [2]

{yj16535, a.luma, azir.aliu, b.selimi, b.raufi, h.snopce}@seeu.edu.mk [1]

vehbi.ramaj@unhz.eu [2]

Abstract: *Studying the right field has great importance in human life and perspective. Rather than affecting the greatest employer's ability, some studies see higher education as one of the leading factors that directly affects the style of life that we do. Therefore, today's demands have increased significantly for skilled people, and prepared in complex areas, and a consolidation between market demands and university curricula is needed. This paper examines Data mining techniques which are used in order to create an automated model which makes comparisons between market demands and university curricula. We also present how proposed model is able to give recommendations, based on the comparison between market demands and university curricula.*

Keywords: TECHNOLOGY, DATA MINING, JOB MARKET, UNIVERSITY CURRICULA, WEB MINING

1. Introduction

It is precisely technology that has affected the lives of people today to be more problematic and more complex, making the vast majority of information available today in the electronic format. And precisely this large volume of information, and the exceeded number of technological devices, has made demand for the labor market in the field of technology to be increased every day and more.

Latest achievements in Data Mining are huge, contributing directly to the development of different fields. Increment of data from time to time has made it necessary to apply Data Mining as a field in order to produce results as accurate and timely as possible. In [1,3] it is mentioned that we are now living in information time.

Lastly, we can easily see that we are dealing with digitalization of almost all areas, ranging from social to scientific ones, and that is precisely why we are dealing with such a large volume of data.

If we stop at social networks, nowadays we can see how from seconds to seconds we are dealing with increasing data in different formats: textual, photo, audio, video, etc.

On the other hand, if we stop at the scientific area, we know that the various exact sciences have reached in a high point of research by applying the technology and digitization their data [6,7,9]. Therefore, this huge amount of data that is crossing the world day by day, it is necessary to apply professional techniques that enable us to process accurate results in a short time.

2. Related Work

The data we are talking about, which grow dynamically at any moment may not be understandable to us when they are in the original format, but applying Data Mining techniques they are processed and converted into a format that is understandable to us. Therefore, as we can see, the main purpose of Data Mining is to adapt to the human language by becoming more and more inseparable part of us.

When we say inseparable part of ours, it should be noted that thousands of applications today function as an inseparable part of the human world by giving different people recommendations on different activities [2]. These recommendations given to different people through various applications today are made possible through Data Mining techniques. And it is precisely the collection and processing of various information that has influenced these applications to give us recommendations and conclusions through the Data Mining techniques.

A concrete example is the case of medical analyzes, where digitized devices are able to produce results on our health condition based on earlier samples of different persons that are stored in the system. The greater the capacity of these data stored in the system, the greater the precision of the result determined by the digital system. This is an exact example of how Data Mining converts data into a format that is understandable to the human world as well.

In [4,5] it is mentioned that evolution and technology development has occurred precisely because of the tremendous development that Data Mining has suffered. This dependence between technology development and Data Mining lies in the fact that the data available today on the internet is of a great variety.

The data that Data Mining can handle are: textual, numeric, structured data and unstructured data, graphical data, and data that are distributed through web systems.

In the following we explain how Data Mining has found the application on each of them.

2.1. Textual data

When we talk about textual data, the Data Mining application process in the text is known as Text mining or Text Data mining. Text mining includes the process of text processing to the presentation in a professional and high quality. The process of text processing through Data mining techniques is the process of removing a word in the text, adding a new word, or completely structuring a document based on a model that the system possesses as a training model [8,11,14]. And once the text has been processed, then it becomes ready for presentation in a format that is understandable to the human eye.

So the main purpose of the Data mining application in text is to convert textual data into data that are readily available for analysis. A simpler method for Text Data mining is converting documents from hard copy to electronic format in order to enable the application of Data mining techniques. Some of the areas that directly depend on Text mining are state intelligence, search engines, publishing houses, social networks, and so on.

2.2. Numeric Data

Creating different models that are able to provide predictive results based on preliminary results is precisely the application of Data mining to numerical data. Nowadays, many predictive analyzes are needed that help us to make decisions based on the result that the system exits. A method by which we can make predictions based on preliminary results is through Bayesian. Through this method we can make data sorting by dividing the data

into two or more groups that the system classifies. Even in this case, the more data that the system possesses the more accurate the classification that the system determines. Some of the domains where Numerical Data mining has found application are: math's, medicine, physics, chemistry, and so on.

2.3. Structured and unstructured Data

The data that can be found in a database can be structured and unstructured [1,16,17]. Structured data are the data that are organized in the best possible way. While unstructured data are data that do not have an organization and a structure. When we are in structured data, it's easier to apply Data mining techniques as we have tables that are related to one another, so when a given data is needed then we know exactly which table and we which line should we ask for it. While the main problem and concern lies in unstructured data as we have no information on where to find it. However, through Data mining techniques we can create intelligent systems using unstructured data and creating in this form of structured and understandable human data forms.

2.4. Graphical Data

A kind of data which is difficult to process are the graphical data due to the complexity they possess and because of the difficulty of conversion in data for analysis. Currently in almost all areas, the results are converted into graphs in order to be more understandable during different presentations. With Data mining techniques we can also process these data in order to obtain desired results. Data mining application in graphical data is known as Graph mining, and unlike other Data mining techniques this technique is less accurate [10,12].

2.5. Web Data

Nowadays companies of all areas promote themselves through the web platform [9]. Apart from the information that should contain a website, designers today are also focused on the dynamics that the website incorporates, including information as part of design and multimedia. It is precisely this fact that has troubled the work of extracting and processing information from web sites. Our automated model depends exactly on the websites, as the main information will be extracted from them. And the best solution is to apply Data mining techniques to the web. Two techniques that enable us to extract this content are Web Scraping and Web Crawling, which we will later discuss in more detail.

2.6. Data mining evolution

In the following we will present the evolution of Data mining and its history of how it has reached this peak point where it is today.

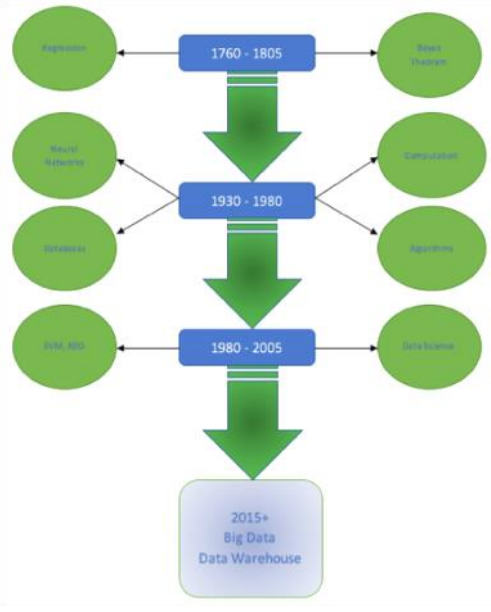

Fig. 1. Data mining evolution over centuries

In Fig. 1 we can see how Data mining has been developed since the 17th century, where from 1760 to 1805 we have the discovery of two theorems: Bayesian and Regression. Currently, through these two theorems, science has reached high discovery by yielding accurate results and predictions regarding certain cases [19,22]. We can also note that in the 19th century there have been great discoveries and great advancements in the Data mining field. Specifically, from 1930 to 1980 is the period when the first databases were created, also the Neural Networks was discovered. It is precisely this time when important algorithms that day-to-day use are being discovered, and since then, they have been sophisticated and advanced in different languages. Also this period is well known for the digitization of these discoveries since all the theorems that were discovered and all the algorithms that were discovered began to be applied through the computer, directly affecting them at the time of their implementation and in the accuracy of the processed results.

Not by chance that these algorithms were applied to the computer, since at this time the volume of data that was stored on the Internet began to increase as the Internet has begun to evolve. Also, in this period, from 1980 to 2005, other discoveries of various algorithms that are known today and are very applicable were made, such as SVM, Deep Learning, KDD, etc., Therefore, this period is also known as the period of the Data Science, at that time scientific methods began to be implemented to produce data and to present the data in a human-accessible format, whether structured or unstructured.

In Fig. 1 we can also note that the last period since 2015 is known as the Big Data period, where the data format that is stored is petabyte above. So this is exactly the time when different businesses and corporations began to use technology in order to create intelligent systems that represent their business in the best possible way [21]. Also in this period, the Data Mining application is started on giant and data sets known as Big Data. Also, the Data Mining application at Big Data has made it possible to make data processing in a precise and fast way. Finally the areas where Data mining has found application on Big Data are different, such as: Medicine, Education, Social Science, Marketing, Finance, etc.

3. Techniques used to build automated model that makes comparisons between market demands and university curricula

Technology development has directly influenced other areas as well. It has greatly influenced the application of different techniques in order to make consolidation between the labor market and the study programs offered by universities has greatly influenced. Next we present the techniques used to create an automated system that will make a comparison between the labor market and curriculum requirements offered by universities.

3.1. Clustering

The division of data into certain groups based on the similarity of objects is known as Clustering. Like other Data mining techniques, clustering is one of the techniques that has managed to develop alongside other techniques. Developing and advancing clustering techniques has also led to the development of many other areas such as medicine, education, finance, marketing, machine learning, etc.

The way clustering works is by creating as many groups of objects which are the same with each other. The greater the likeness of being within a group, and the bigger the difference between the groups, the greater the clustering accuracy.

In addition to clustering there are many other techniques that make data collection in different groups, but in some literature we can also find that clustering can also be known as a form of classification. We say that it is known as a form of classification, as it divides into data groups by classifying them into different groups. But distinction from the classification technique that classifies the

data into different groups, and when it comes to a data that is classified as a new group, clustering these new data tries to group them with the actual data being compared to the most similar figure. In the following we will show graphically how clustering works.

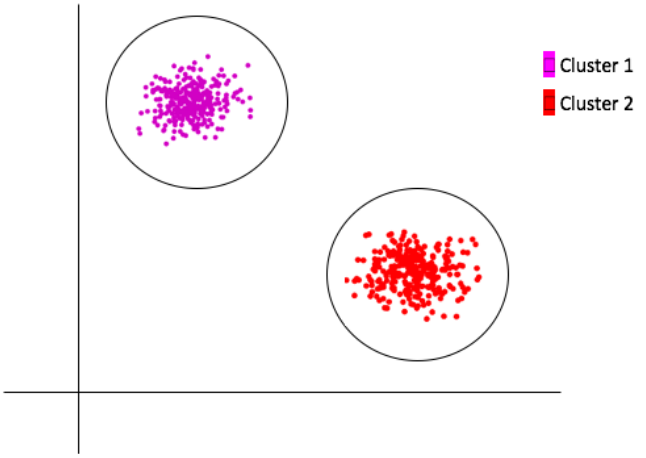

Fig. 2. Clustering different groups

In Fig. 2 we can see how the clustering technique is divided into two sets of data that are located in our dataset. These data are divided into two groups that differ with each other, but with data that are close to each other. In our case the data are divided into two groups only because of the difference between them, but in other cases we have more groups. Also in Fig. 2 we can see that we have a high level clustering because the data are completely separated, but there may also be times when the data cannot be completely separated but only join to the group which is more closest. When we are in clustering, we can point out that clustering types are: well separated, center - based, contiguity - based, density - based and conceptual clusters.

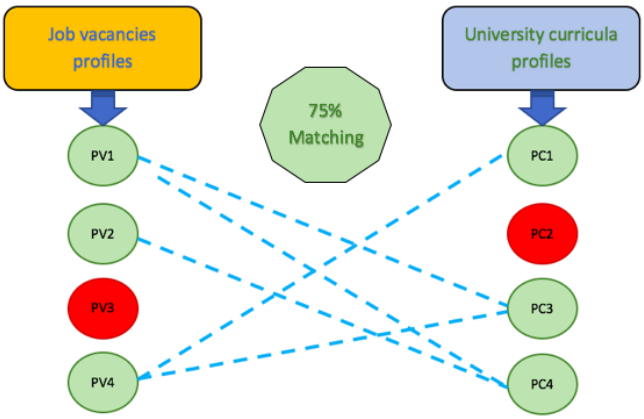

Fig. 3. Matching between market demand and university curricula

In Fig. 3 we can see how the matching algorithm works, which in this case is illustrated how it will be applied to our model. As we can see on the left hand we have job vacancies profiles that are named from PV1 to PV4, and under the right hand we have university curricula profiles that are named from PC1 to PC4. After applying the algorithm we can note that the PV1 has PC3 and PC4 adaptability, as well as PV2 has PC4 compatibility, and PV4 has PC1 and PC4 compatibility. As we can see profiles vacancies 3 and profile curricula 2 have no relation to any of the profiles. And what results from this result is that profile 3 of the job remains uncovered, and in that case our system will be able to give recommendations that this profile should be covered so that a new curriculum can be added.

Or in the other case we have profile curricula 2 which also does not have any links to any of the profiles of work, then we can conclude that this curriculum is not needed in the labor market, and our system will be able to make recommendations on the changes that need to be made in this curriculum so that it responds to the demands of the labor market.

Also based on the results that derived our algorithm after the application, it can be concluded that out of 4 profiles, 3 of them respond to labor market demands or 75% is the level of adjustment between the demands of the labor market and the curricula offered from universities in the field of technology.

3.3. Web crawling

The program that allows us to automate browsing through websites that are active is known as web crawling [14]. The way the web crawler works is by checking webpages that are active in certain phrases that we define by themselves. Some websites today use web crawling as a perfect way to keep their web pages updated.

The other way web crawlers work is by visiting the web pages automatically and downloading them to a local disk that we set as a destination to maintain the content. The content of the webpage we can download, starting from static to dynamic, but depending on the dynamics of the webpage depends also on the script we need to build in order to have as much information as possible. Below we will graphically show how the web crawler works.

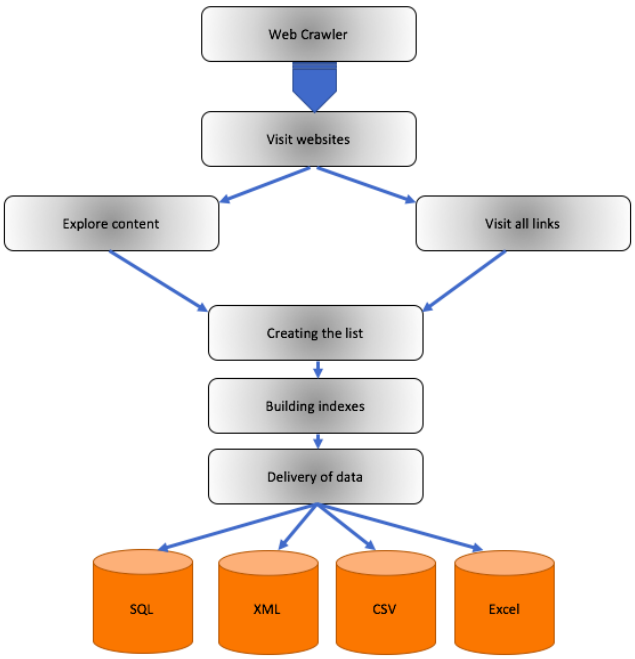

Fig. 4. Extraction of vacancies websites

In Fig. 4 we can see that the first step in which the web crawler application passes is the site visit that we define at the beginning. After a visit to the web, the crawler makes an exploration of the content based on the phrases that we have assigned at the beginning, and at the same time the crawler visits the other links that are defined within the webpage.

Once the textual content has been extracted, the crawler creates lists of those web pages, in order to create indexes that will be downloaded later.

Once the indexes are created then the last step is to save the data, or as we have presented it as delivery of the data, where the data we extracted from the content of the webpage can be stored in any format that we have need. The most used formats that can be converted to the data extracted from the web site are: SQL, XML, CSV, Excel, etc [20, 21].

Based on this, the main reason for web crawling will be applied in our research is to extract information that is published by some websites on vacancies. These data will be extracted in certain keywords, where their clustering will also be based on them.

Initially in the region there are web pages that contain data on university programs, these data also contain descriptions of the subject and description of the study program.

Based on this information, will be the clustering of study programs that are offered by universities. These later clustering will be used to match the market requirements.

On the other hand, Crawling will also be applied on web sites that provide information on competitions offered by different companies [13,17]. Also, the web crawling application will be made to specific keywords, knowing that each contestant has additional information on the specific position requests.

So, every position that is required to keep information in addition to the required position, the applicant must have knowledge in several different areas such as: programming, databases, networking, etc. All of these descriptions of later positions will be used to gather different positions.

4. Conclusion

The relevance of adapting the study programs to the labor market demands has a great importance. In our paper, we presented the tools that helps us to create a model that will compare the demands of the labor market and the curricula offered by universities.

First we presented the latest achievements in the Data mining field and the stages that Data Mining has passed to reach where it is today. Then we have seen what are the techniques we have used to create the model that will make comparisons between market demands and the curricula offered by universities.

Also in our paper we presented parts of the automated model such as: the application of the algorithm that will find the level of the actual adjustment between the labor market demands and the curricula offered by the universities. Then we presented the steps that are passed in order to extract the webpages that publish competitions in the field of technology.

In the end we conclude that our country and the region has a great need for a better adjustment between study programs and labor market demands. We also conclude that the creation of such an automated model will help meet the demands of the labor market, and our universities will be able to offer students also for the European market.

References

[1] M. Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education". In IEEE Access, May 2016.

[2] T. Xie, Q. Zheng, W. Zhang, H, Qu, "Modeling and Predicting the Active Video – Viewing Time in a Large – Scale E – Learning System". In IEEE Access, June 2017.

[3] A. M. Njeru, M. S. Omar, S. Yi, "IoTs for Capturing and Mastering Massive Data Online Learning Courses". In IEEE Computer Society, ICIS, Wuhan, China, May 2017.

[4] R. Heartfield, G. Loukas, D. Gan, "You are probably not the weakest link: Towards Practical Prediction of Susceptibility to Semantic Social Engineering Attacks". In IEEE Access, October 2016.

[5] E. J. Fortuny, D. Martens, "Active Learning – Based Pedagogical Rule Extraction". In IEEE Transaction on Neural Network and Learning Systems, Vol. 26, No. 11, November 2015.

[6] A. Mukhopadhyay, S. Bandyopadhyay, "A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I". In IEEE Transaction on Evolutionary Computation, Vol. 18, No. 1, February 2014.

[7] Zh. Song, A. Kusiak, "Optimization of Temporal Processess: A Model Predictive Control Approach". In IEEE Transaction on Evolutionary Computation, Vol. 13, No. 1, February 2009.

[8] S. Malgaonkar, S. Soral, Sh. Sumeet, T. Parekhji, "Study on Big Data Analytics Research Domain". In International Conference on Reliability, Infocom Technologies and Optimization ICRITO,Noida, India, September 2016.

[9]K. P. Anicic, B. Divjak, K. Arbanas, "Preparint ICT Graduates for Real – World Challenges: Results of a Meta – Analysis". In IEEE Transactions on Education, Vol 60, No. 3, August 2017.

[10] A. Haskova, D. V. Merode, "Professional Training in Embedded Systems and its Promotion". In IEEE Transacions on Education, 2016.

[11] S.C. Smith, W. K. Al-Assadi, J. Di, "Integrating Asynchronous Digital Design into the Computer Engineering Curriculum". In IEEE Transactions on Education, Vol. 53, No. 3, August 2010.

[12] M. D. Koretsky, D. Amatore, C. Barnes, Sh. Kimura, "Enhancement of Student Learning in Experimental Design Using a Virtual Laboratory". In IEEE Transactions on Education, Vol. 51, No. 1, February 2008.

[13] B. G. Member, V. S. Sheng, K. Y. Tay, W. Romano, Sh. Li, "Incremental Support Vector Learning for Ordinal Regression". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 26, No. 7, July 2015.

[14] J. Li, T. Zhang, W. Luo, J. Yang, X. T. Yuan, J. Zhang, "Sparseness Analysis in the Pretraining of Deep Neural Networks". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, No. 6, June 2017.

[15] Y. Qian, F. Li, J. Liang, B. Liu, Ch. Dang, "Space Structure and Clustering of Categorical Data". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 27, No. 10, October 2016.

[16] Y. Xiao, B. Liu, Zh. Hao, "A Maximum Margin Approach for Semisupervised Ordinal Regression Clustering". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 27, No. 5, May 2016.

[17] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, Sh. Li, "Incremental Support Vector Learning for Ordinal Regression". In IEEE Transactions on Neural Networks and Learning Systems, Vol. 26, No. 7, July 2015.

[18] P. Navrat, L. Molnar, "Curricula Transformation in the Countries in Transition: An Experience from Slovakia". In IEEE Transactions on Education, Vol. 41, No. 2, May 1998.

[19] S. Nalintippayawong, K. Atchariyachanvanich, "IT Management Status in Public Higher Education Institutions in Thailand". In IEEE ICIS 2016, June 26-29, 2016, Okayama, Japan.

[20] J. I. Godino – Llorente, R. Fraile, J. C. Gonzales de Sante, V. Osma – Ruiz, N. Saenz – Lechon, ""Design for All in the Context of the Information Society": Integration of a Specialist Course in a Generalist M.Sc. Program in Electrical and Electronics Engineering". In IEEE Transactions on Education, Vol. 55, No. 1, February 2012.

[21] M. Dolores Cano, "Students' Involvement in Continuous Assessment Methodologies: A Case Study for a Distributed Information Systems Course". In IEEE Transactions on Education, Vol. 54, No. 3, August 2011.