International Journal of Mathematics and
Computer Science, **14**(2019), no. 2, 411–424

Proposition of an employability prediction system using data mining techniques in a big data environment

Mohamed Saouabi, Abdellah Ezzati

LAVETE laboratory, FST
University Hassan 1^{st}
Settat, Morocco

email: mohamed.saouabi@gmail.com, abdezati@gmail.com

(Received December 17, 2018, Accepted January 18, 2019)

Abstract

Employment is the main form of social integration, a factor of improving living conditions and preventing risks of poverty and vulnerability and the most appropriate indicator for assessing the level of social cohesion in a country. Graduates face every year real competitions to ensure their employability. We live now in a big data era, where the data is generated every day in large volumes, due to the use of technologies. But storing this data isn't the main objective, we need to use this data and take advantage of it by extracting valuable information from it in order to use it and improve a particular domain. Using data mining techniques will allow us to extract useful information and ameliorate employability by giving decision makers opportunities to make it better and predict in the future. In this paper, we present an intelligent system using data mining techniques on employability data, in a Big Data environment, which we used Hadoop ecosystem, and for data mining we used Rapid Miner Studio Educational Version 8.1.000. We presented first the characteristics of the proposed system, after that we presented the general architecture of the system and the tools and the technologies used, and finally we presented the system's process in details from the data collection till the results visualization.

Key words and phrases: Big Data, Data mining, Employability, Hadoop, Rapid Miner.
AMS (MOS) Subject Classification: 68T30.
ISSN 1814-0432, 2019, http://ijmcs.future-in-tech.net

1 Introduction

As data grow, traditional technologies and current database systems cannot handle this huge amount of data and process it efficiently. Big Data came in to resolve these problems and to offer solutions in order to efficiently process this huge amount of data and analyze it. Combination of big data and data mining can offers a lot of helps and it give opportunities for decision makers in order to take advantage of the results and predict into the future. We presented an employability prediction system (EPS), this system process and analyze data into Hadoop ecosystem using the different technologies that Hadoop offers, and we used Rapid Miner Studio Educational Version 8.1.000 for data mining. We have a classification problem, classifying graduates into working and not working. We used Rapid Miner as a data mining tool, we implemented three classification algorithms, Decision Tree, Regression logistic and Naïve Bayes. We presented the general architecture of the system as well as the system's characteristics and the process phases from the data collection till the results visualization.

2 Related Work

Many fields now are taking advantage of the powerful use of big data and data mining and the opportunities offered for future prediction and improvement. Sunitha and Sermakani [1] presented a system using big data and data mining in order to identify the cloud leakage responsible. They used data allocation which injects realistic but fake information records for improving the cloud leakage identification. The main objective of this work is to identify cloud leakage, provide security of cloud data, and protect the sensitive data from unauthorized access. Philip [2] presented a framework of knowledge creation with an objective of helping managers and researchers to understand the conversion of big data turning into knowledge that can be useful, by proposing elements for helping understand the conversion of big data into tacit and explicit employee knowledge. Also proposition of a solution for enhancing firm's dynamic capabilities using big data. Chien, Hong, and GUO [3] proposed a framework having for objective improving the manufacturing performance and achieve decision making ability of smart factory, by integrating knowledge and decision rules from big data analysis and simulations. The results they came up with have shown practical viability of the developed solutions enhancing productivity and flexibility. Founds [4] presented a framework in a big data environment dedicated for the guide of nursed and clinicians in the acquisition, modeling and management of multiomics data using big data from patients in illness. They propose a way how nurses can participate in the science and clinical translation of biological systems for precision health, and the opportunity for the patient to be engaged in their own care.

3 Big Data, Data mining and Employability

Nowadays, people and things are interconnected all the time, thanks to advances in communication technologies. The use of smart connected devices, such as cars with location sensors, smartphones and the use of social media, generates a lot of the data that the traditional database systems cannot manage. Big Data offers solutions, it can handle very large amount of data both structured and unstructured, on a variety of terminals. But the biggest challenge for Big Data isn't just storing this data, it is to explore this data and mine it in order to extract valuable information and knowledge for future actions. Which is why data mining is the most important process, it helps to make proactive actions based on the models generated and the knowledge discovered in order to answer problematic questions and predict in the future.

A lot of institutions and industries use big data and data mining in order to make their institutions better and ameliorate their conditions. Employability represents a serious problem for graduates, they face every year a big competition concerning the professional insertion and it is increasingly difficult. There are many explanations and causes for this matter, for example the poor economic performance of the country, the structure of the economy and its educational system take a big responsibility, or maybe the university fields of study which makes the professional insertion a bit difficult.

The use of big data and data mining will clarify the view and point the problems, and will also present solutions like identifying the determinants responsible of the professional insertion of the graduates, it may be because of the graduate's school curriculum, or maybe the job market, or the field of study chosen by the graduates. Answering such questions could be useful for graduates as well as researchers and public authorities for a better assessment of the training quality system and make the necessary readjustments.

4 The proposed system: Employability Prediction System (EPS)

4.1 The system's characteristics

Our Employability prediction system (EPS) offers many facilities and opportunities for decision makers in order to make improvement in the employability field. After analyzing the data and processing it in Hadoop ecosystem, this data will be mined with the data mining tool Rapid Miner and we'll present models and prediction information for making the future employability better.

The system will be able to:

- Supply the database with the needed data (baccalaureate, field, diploma, university, graduates, and graphs);
- Ingest the data from MYSQL database into Hadoop ecosystem;
- Query and process the data into Hadoop using the different technologies of the Hadoop ecosystem;
- Visualize data and create dashboards into Hadoop in order to understand the data ;
- Apply the data mining process to extract valuable information from the data;
- Share the results and the graphs generated by the prediction analytics;
- Propose model of employability prediction;
- Propose the variables which have the most impact on the employability;
- Predict if the graduates will be classified as working or not working.

4.2 The global architecture of the system for employability prediction

Now, we will present the general architecture of our proposed system. The system is based in Hadoop ecosystem. Hadoop offers a lot of tools and technologies making the processing and the analytics of the data an easy task. But, choosing the right technology for the right task is not as easy as it

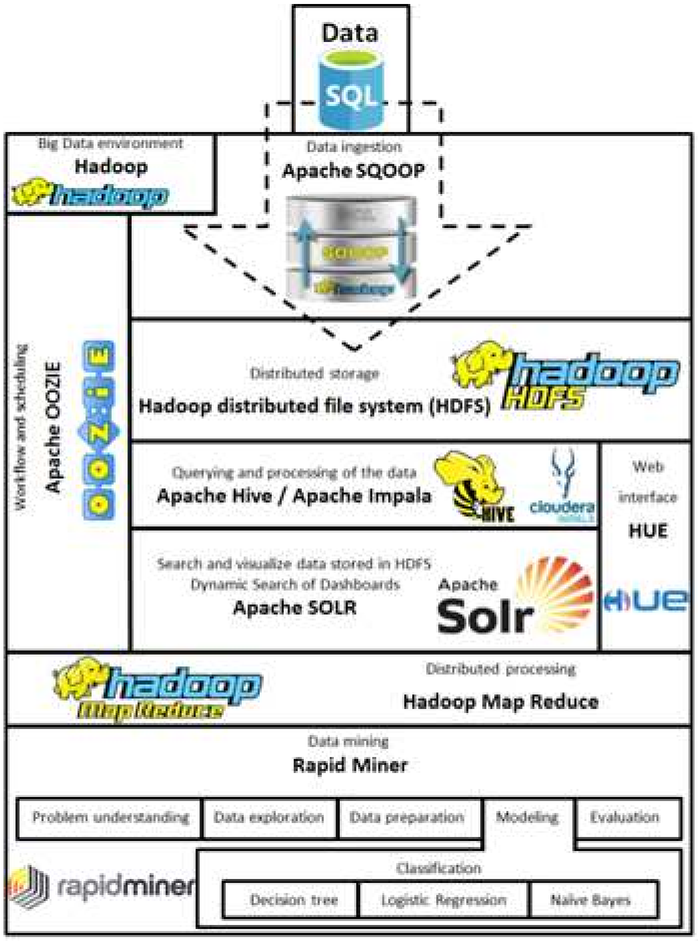

Figure 1: Employability Prediction System (EPS)

seems, we present here the architecture of the system, and all the tools and technologies used, explaining every technology aside and how it works.

4.2.1 HDFS

HDFS is a distributed file system dedicated to handle big data in combination with Map Reduce for distributed processing. The massive reliability and capacity for increasing the data management of the HDFS cluster making it very efficient in term of handling big data. HDFS handle a large amount of data in parallel and in fault tolerant manner and it offers an easy and fast access to data. HDFS presents a lot of advantages, including hardware scalability, data replication between cluster nodes, concurrent transactions, availability in case of node failure. It's based on master/slave architecture.

One of the big advantages of the distributed file system of Hadoop HDFS; the processing of the chunks of data in parallel, the slave nodes do all of the work of storing the data and running the computations. The slaves run data nodes and use the task tracker daemon, which takes instructions from the master nodes. The task tracker daemon is a slave to the Job tracker, and the data node daemon is a slave to the name node.

The master nodes on another side supervise the most important cores of Hadoop, which are storing the data in parallel in HDFS, and process it also in parallel using Map Reduce. The job tracker coordinates and supervises the parallel processing of the data using Map Reduce, while the name node coordinates and supervise the data storage function in HDFS.

4.2.2 Hadoop Map Reduce

As we explained before, Hadoop use HDFS in order to store the data. This data is stored in a distributed file system (HDFS), so to process this data Hadoop use Map Reduce to process this data in parallel. Map Reduce is a programming paradigm allowing the distributed processing in Hadoop cluster. It became dominant concerning the batch-processing, which consist of taking the input data and divide it into small chunks that are processed in a parallel manner.

Map reduce consists of two main functions, the Map and the Reduce. The first function, Mapper, is to process and create several small chunks of data of key/value pairs that are processed in parallel, and the outputs of the Mapper function will be shuffling and sorting and making the data ready for the next phase. The reduce phase, and here the data is aggregated to return the results.

4.2.3 Apache HIVE

Apache HIVE is a data warehouse system offered by Hadoop using the HIVEQL language to query the data stored in Hadoop and to facilitate adhoc querying, aggregation, and the analysis of large volumes of data stored in Hadoop distributed file systems. Learning HIVEQL is easy for the users who are familiar with the SQL language. Processing the data stored in HDFS needs Map Reduce; programming a Map Reduce is not simple, so HIVE also can convert queries HIVEQL into executable MapReduce jobs on Apache Tez, which is a framework of execution on Hadoop.

4.2.4 Apache IMPALA

In addition of Apache HIVE, impala also offers SQL syntax in order to send interactive SQL queries directly on Apache Hadoop data stored on HDFS. It provides a unified platform and familiar for batch-oriented or real-time queries. Apache impala offers a lot of advantages, besides the familiar SQL interface, the ability to query high large amount of data – Big Data- on Hadoop. Also the ability to data interchange between Impala and HIVE tables for read and writes, offering an easy analytics on Hive-produced data.

4.2.5 Apache SOLR

Apache SOLR allow the Hadoop user to explore the data stored in HDFS and discover it, visualize the data and offer a dynamic search of dashboards. It's optimized for high volume of data, making it capable of ingesting a massive amount of data to be available for the user in an intelligent way, in a few milliseconds.

4.2.6 HUE

HUE offers an interactive interface for analyzing data stored in Hadoop, offering a lot of features like an Editor for HIVE queries, Impala queries, browser for jobs designer, OOZIE interface to schedule jobs, etc. HUE offers another way to interact with Hadoop without the command prompt interaction for most Hadoop activities. HUE is developed by Cloudera.

4.2.7 Apache OOZIE

Apache OOZIE is a workflow scheduler used to schedule and manage Hadoop jobs, supporting different types of Hadoop jobs (HIVE, SQOOP, Java MapReduce, etc.). All sorts of programs involved in the Hadoop cluster can be organized in a specific order of execution using OOZIE, offering also a mechanism to run jobs at a given time following a defined schedule.

4.2.8 Apache SQOOP

Apache SQOOP plays an important role in the Hadoop ecosystem, as we know, applications and websites generally works with relational databases, which makes them one of the most important sources generating big data. Apache SQOOP provides interaction between relational databases and HDFS. It's designed in order to transfer data between HDFS and relational databases like Oracle, MySQL, TeraData, SQLITE, etc.

4.2.9 Data mining: Rapid Miner

Data mining refers to the extraction of information and hidden patterns from the data, but it's not the only process we need to perform, data mining involves other processes such as problem understanding, data exploration, data preparation, modeling, evaluation and deployment. Once all these processes are over, we would be able to use this information in order to improve a particular domain. We chose Rapid Miner which is open source system dedicated to machine learning and data mining processes, it also integrate Weka machine learning environment and statistical modeling schemas of the R project which can be installed as an extension on Rapid Miner.

4.3 The system's process

In this part, we will present the different phases of our proposed system in details, describing every phase apart.

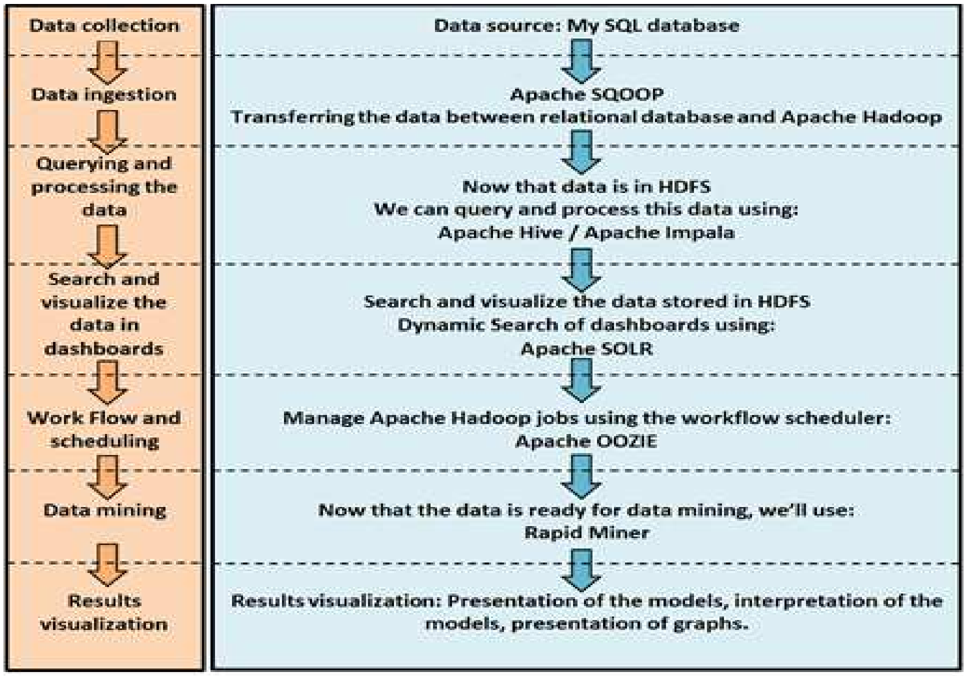
Data collection
Data ingestion
Querying and processing the data
Search and visualize the data in dashboards
Work Flow and scheduling
Data mining
Results visualization
Data source: My SQL database
Apache SQOOP
Transferring the data between relational database and Apache Hadoop
Now that data is in HDFS
We can query and process this data using:
Apache Hive / Apache Impala
Search and visualize the data stored in HDFS
Dynamic Search of dashboards using:
Apache SOLR
Manage Apache Hadoop jobs using the workflow scheduler:
Apache OOZIE
Now that the data is ready for data mining, we'll use:
Rapid Miner
Results visualization: Presentation of the models, interpretation of the models, presentation of graphs.

Figure 2: Employability Prediction System (EPS): The process

4.3.1 Data collection

In this phase, we collect the data; we used MYSQL as a relational database management system (RDBMS). In the next phase, we'll ingest the data from MYSQL into Hadoop distributed file system (HDFS) using Apache SQOOP.

4.3.2 Data ingestion

In this phase, we need to put the data into HDFS. The data we have is in a relational database format, we used SQOOP, which stands for SQL to HADOOP. First of all, we create a SQOOP job and then we launched it using the command below:

```
sqoop job –create create_db_job
```

And then we launch the job by executing the following command:

```
sqoop job –exec create_db_job
```

Now the data is transferred from MYSQL and stored into HDFS, it's ready to be queried and processed in Hadoop ecosystem.

4.3.3 Querying and processing the data

Hadoop offers a lot of tools and technologies making the process of the data easy. Our data now is in HDFS, we can process this data and even send SQL queries upon Hadoop, in order to see our data and look for the data problems. We can use the hive shell or the HUE interface.

4.3.4 Search and visualize the data in dashboards

In this phase, we'll explore and discover the data stored in HDFS by using graphs and dashboards with Apache SOLR. We used the interface HUE which making the creation of the dashboards easier and faster.

4.3.5 Workflow and scheduling

In this phase, we'll schedule the jobs to be executed using Apache OOZIE. We'll take for example SQOOP, which is the first step for ingesting the data from MYSQL database into Hadoop HDFS. Let's say that we want to launch the SQOOP job of data ingestion *Every week On Monday at 06.30,* the coordinator make sure that the job is execute in the exact same date, by giving the name of the coordinator as parameter, name of the job to execute and the time of execution.

4.3.6 Data mining

The most important phase of the process is data mining. In this phase, we'll apply the data mining process using Rapid Miner Studio Educational Version 8.1.000. The process involves the following phases: Problem understanding, data exploration or data understanding, data preparation, modeling and evaluation.

Once all this phases are completed, we'll be able to extract the knowledge and visualize the results in the next phase of the system's process.

4.3.7 Results visualization

This is the final phase of the process, where the presentation and interpretation of the models, presentation of graphs, the variables which have impact on the objective variable, those results will help decision makers and give them the opportunity to make proactive actions and predict in the future.

5 Conclusion

With the advances of technologies, prediction of the future is no longer a difficult task, the use of big data and data mining resolves every day a lot of problems and gives decision makers opportunities for improvements. Employment is one of the biggest problems graduates face every year. Pointing the problems will present solutions like identifying the determinants responsible of the professional insertion of the graduates, it may be because of the graduate's school curriculum, or maybe the job market, or the field of study chosen by the graduates. Answering such questions could be useful for graduates as well as researchers and public authorities for a better assessment of the training quality system and make the necessary readjustments. We presented in this paper an intelligent system using data mining techniques on employability data, in a Big Data environment, which we used Hadoop ecosystem, and for data mining we used Rapid Miner Studio Educational Version 8.1.000. We presented first the characteristics of proposed system, after that the general architecture of the system and the tools and the technologies used, and finally the system's process from the data collection tills the results visualization.

References

[1] J.Sunitha, A. M. Sermakani, Building an Authentication and Quality of Query Services in the Cloud, Procedia Computer Science, **50,** (2015), 122–127.

[2] Jestine Philip, An application of the dynamic knowledge creation model in big data, Technology in Society, **54,** (2018).

[3] Chen-Fu Chien, Tzu-yen Hong, Hong-Zhi Guo, A Conceptual Framework for "Industry 3.5" to Empower Intelligent Manufacturing and Case Studies, Procedia Manufacturing, **11,** (2017), 2009–2017.

[4] Sandra Founds, Systems biology for nursing in the era of big data and precision health, The official journal of the council for the advancement of nursing science, **66,** no. 3, (2018), 283–292.

[5] Shabnam Shadroo, Amir Masoud Rahmani, Systematic survey of big data and data mining in internet of things, Computer Networks, **139,** (2018), 19–47.

[6] Kamal Al-Barznji, Atanas Atanassov, Review of big data and big data mining for adding big value to enterprises, Science Engineering & Education, **2,** no. 1, (2017), 50–57.

[7] Rakesh Kumar, Bhanu Bhushan Parashar, Sakshi Gupta, Yougeshwary Sharma, Neha Gupta, Apache Hadoop, NoSQL and NewSQL Solutions of Big Data, International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE), 1, no. 6, (2014), 28–36.

[8] Urmila R. Pol, Big Data, Hadoop Technology Solutions with Cloudera Manager, International Journal of Advanced Research in Computer Science and Software Engineering, **4,** no. 11, (2014), 1028–1034.

[9] Wissem Inoubli, Sabeur Aridhi, Haithem Mezni, Alexander Jung, Big Data Frameworks: A Comparative Study, Arxiv, **2,** (2017).

[10] Wei Fan, Albert Bifet, Mining Big Data: Current Status, and Forecast to the Future, SIGKDD Explorations, **14,** no. 2, (2013), 1–5.

[11] H. V. Jagadish, Big Data and Science: Myths and Reality, Big Data Research, **2,** (2015), 49–52.

[12] S. Barbosa, R. Souza, S. Cruz, L. Campos, Applying Data Warehousing and Big Data Techniques to Analyze Internet Performance, International Conference on Big Data Computing, Applications and Technologies, (2016), 67–92.

[13] Rikita Patel, Tanvi Desai, Big Data Analytics in Optimizing the Quality of Education: Challenges, International Journal for Innovative Research in Science & Technology, **3,** no. 6, (2016), 165–167.

[14] A. Tanuja, Swetha Ramana, Analyzing Big Data using Hadoop, International Journal for Innovative Research in Science & Technology, **3,** no. 3, (2016), 13–16.

[15] Gayathri Ravichandran, Big Data Processing with Hadoop: A Review, International Research Journal of Engineering and Technology, **4,** no. 2, (2017), 448–451.

[16] P. Muthukumar, Big Data in Data Mining, International Research journal of Management Science and Technology, **5,** no. 10, (2014), 54–59.

[17] Samson Oluwaseun Fadiya, Serdar Saydam, Emeka Joshua Chukwuemeka, Big Data in Education; Future Technology Integration, the International Journal Of Science & Technology, **2,** no. 8, (2014), 65–69.

[18] B. Tulasi, R. Suchithra, Big Data analytics and e-learning in higher education, International Journal on Cybernetics & Informatics (IJCI), **5,** no. 1, (2016), 81–85.

[19] Athanasios S. Drigas, Panagiotis Leliopoulos, The Use of Big Data in Education, IJCSI International Journal of Computer Science, **11,** no. 5, (2014), 58–63.

[20] B. Tulasi, Learning Analytics and Big Data in Higher Education, International Journal of Engineering Research & Technology IJERT, **3,** no. 1, (2014), 3377–3383.

[21] D. Sindhuja, R. Jemina Priyadarsini, A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder, International Journal of Computer Science and Mobile Computing, **5,** no. 5, (2016), 483–488.

[22] Surjeet Kumar Yadav, Saurabh Pal, Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, World of Computer Science and Information Technology Journal WCSIT, **2,** no. 2, (2012), 51–56.

[23] Xie Guangqiang, Li Yang, Extension Data Mining Knowledge Representation, Physics Procedia, **24,** (A), (2012), 240–246.

[24] B. Umadevi, R. Dhanalakshmi, A Comprehensive Survey of Students Performance Using Various Data Mining Techniques, International Journal of Science and Research IJSR, **6,** no. 4, (2017), 2233–2238.

[25] Rohit A. Kautkar , A Comprehensive Survey On Data, International Journal of research in engineering and technology (IJRET), **3,** no. 8, (2014), 185–191.

[26] Neha Midha, Vikram Singh, A Survey on Classification Techniques in Data Mining, International Journal of Computer Science & Management Studies (IJCSMS), **16,** no. 1, (2015), 9–12.

[27] Thair Nu Phyu, Survey of Classification Techniques in Data Mining, Proceedings of the International Multi Conference of Engineers and Computer Scientists, **1,** (2009).

[28] Petar Ristoski, Heiko Paulheim, Semantic Web in data mining and knowledge discovery: A comprehensive survey, **36,** (2016), 1–22.

[29] Kalpana Rangra, K. L. Bansal, Comparative Study of Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering, **4,** no. 6, (2014), 216–223.