Web Mining techniques and applications : literature review and a proposal approach to improve performance of employment for young graduate in Morocco

K.SELLAMY , Y. FAKHRI , S. BOULAKNADEL[1] , A.MOUMEN, K. HAFED
Laboratory of Research in Computer Science and Telecommunications
Faculty of Sciences - University Ibn Tofail , Kenitra-Morocco
[1] IRCAM, Allal El Fassi Avenue, Madinat Al Irfane, Rabat-Instituts, Morocco
sellamyinfo@gmail.com

H JAMIL
Secretary of State for Water, Information System Division, Rabat, Morocco

Y. Lakhrissi
National School of Applied Sciences of Fes, Fes, Morocco

Abstract— **View the large amount of data presented in the web pages, there is a great focus of different organizations and companies, to gather this information to use them in their best interest. This is achieved through Web mining which uses the techniques and algorithms of Data Mining to extract information and knowledge directly from the Web.**

The Web data mining is not only focused on the achievement of commercial information, but it is also used by various organization to make the good predictions and decisions in a different areas. The present paper deals with primary discussion of web mining, it focuses on a short presentation of web data mining its techniques, tools and applications. Then, we will propose an approach for the cross-analysis between the skills acquired in university training and the skills sought by employers in Morocco.

Keywords— ***Web data mining, data mining, web usages mining, web content mining, web structure mining.***

I. INTRODUCTION

The first use of the term Web mining (WM) goes to Oren Etzioni who tried to apply data mining technology on the Web. In this work [1], he defines Web mining as the application of data mining techniques to Web data for the extraction of relevant information from the resources available in the Web, a Web resource can be a document or a web service.

Since 1996, several other works and publications have focused on this subject. The almost all of this research, it is widely accepted that Web mining is a subject of many other fields (information retrieval, databases, artificial intelligence ...), which aims to extend and adapt data mining techniques [2],[3].

The data are generally stored in a Data-Wabehouse, including the objective of construction is to collect specific data to analyze the behavior of navigation we can classify the data used in four types :

- Content Data: Data contained in the Web pages (texts, images, graphics…).
- Data relating to the structure: structure of the page, structure inter-page…
- Data relating to the Use: data providing information on the use such as IP addresses, the date and the time of queries.
- Data relating to the profile of the user.

In this work we will present a review of the literature on Web mining, since its techniques: last trend, its tools, and an overview on some areas of application. Then, we will propose an approach for the cross-analysis between the skills acquired in university training and the skills sought by employers in Morocco.

II. THE TYPES OF WEB MINING

978-1-5386-4396-9/18/$31.00 ©2018 IEEE

According to the types of data to be extracted, we can divide the Web mining into three axes:

A. *Web Content Mining*

The Web content mining is the mechanism for extracting knowledge from the actual content of documents and web pages such as structured recordings, images, texts, videos, etc… [4].

B. *Web Structure Mining*

Web Structure Mining is an analysis of the structure of Web i.e. the architecture and the links that exist between the different sites. The analysis of the paths travelled allows, for example, determining how many pages to consult the internet users on the average and thus adapt the site tree for that the pages of the most sought after are in the first pages of the site. Similarly, research associations between the pages consulted allows improving the ergonomics of the site by creation of new links. We cited two well-known structures mining algorithms, PageRank and HITS [4].

C. *Web usage mining*

The purpose of this methodology is to analyze user behavior and to extract interesting usage patterns from the interaction with the website. There are three steps in web usage mining that helps to navigate the web in an effective way. These steps include the preprocessing, pattern discovery and pattern analysis.

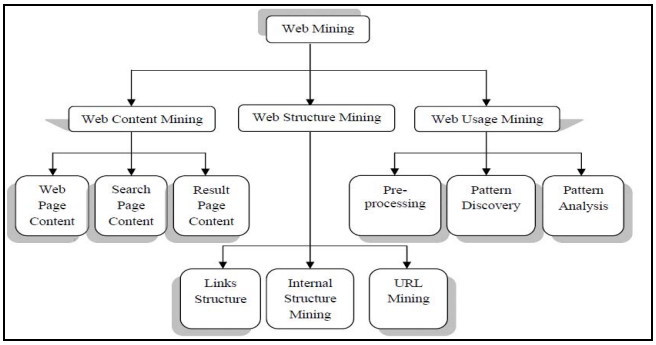

Fig1: Classification of Web Mining

III. WEB MINING TECHNIQUES : LAST TRENDS

Several techniques and algorithms are used by the web mining for the exploration of the data-web we present someone [5]:

A. *Types of classification models:*

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

B. *Types of clustering methods*

- Partitioning Methods.
- Hierarchical Agglomerative (divisive) methods.
- Density based methods.
- Grid-based methods.
- Model-based methods.

C. *Types of association rule*

- Multilevel association rule.
- Multidimensional association rule.
- Quantitative association rule.

Many results have been obtained by implementing the different web mining algorithms which help businesses in structuring and organizing the website. The following table provides insight into the findings from the previous research in the data and web mining.

TABLE 1: LITERATURE SURVEY ON WEB MINING

Year	Researchers	Algorithm/ Method	Input	Results
2017	Hesham Abdo Ahmed Aqlan Shoiab Ahmed Ajit Danti [6]	MLPReg-ressor, Multilayer Perceptron, Linear Regression, SMO Regression	-	these methods is used for time series analysis to predict future events based on past events' Time series forecast'
2016	Radhika Bairagade, Nikita Afre, Nirmala singh, and Durga Bhamare [7]	Smart-Crawler	Web site	SmartCrawler is a focused robot for location and exploration of the site. it is for snatching deep web interfaces.
2016	S.Sharma and S.SLodhi [8]	Decision Tree Algorithm	Web log files	The proposed web mining method withstand with all the applied input Parameters.
2016	S. SPatil and HP. Khandagale [8]	Various web mining algorithm [	Web site navigat ion log files	The proposed method helps to identify the user pattern in an effective way and reduces the time expended by the developer.
2016	A. Raiyani and S.S. Pandya [8]	Reviewed Knowledge Discovery in Database	-	provide information about approaches used in Knowledge disco-very process through different

Year	Researchers	Algorithm/ Method	Input	Results
2016	P.Sukumar and al [8]	Investigate and implement-ed various web mining algorithm	Raw Log Files	Provided benefits and limitations of the reviewed algorithms
2016	Sunena and K. Kaur [8]	Compared various data mining techniques	-	The comparison is helpful for providing web services to the users
2015	C Bull etal [8]	Software Architec-ture for Mental Health Self-Manage-ment	Hardw are log files	The proposed methods allow asecurestorage and large-scale data collection.
2015	B. K. Malviya and J. Agrawal [8]	Studied various techniques of W U M	-	Described the whole web mining proves and the problems encountered in it.
2015	Ms Shashi Sahu1 Leena Sahu2 [9]	EPLogClea -ner Two-level clustering Noise Detector	web log	EPLogCleaner filters better than traditional methods of data cleansing it exceeds 30% of URL requests.
2013	A. Raiyani and Prof . S.S. Pandya [8]	Distinct User Identifica-tion (DUI)	User Log files	Proposed method is effective for fraud detection and unusual, suspicious activities on the secure data.
2013	, Mehul P. Barot, Shaily G.Langhnoja, Darshak B. Mehta, [10]	DBSCAN	Web log files	The proposed web mining method used to find user's having common behavior and access patterns.
2012	Loraine Charlet Annie M.C , Ashok Kumar D [11]	Apriori K-Apriori K-means	Web pages	Experiments are performed using real and synthetic data, and found Apriori algorithm is less efficient compared to K-Apriori algorithm.

IV. WEB MINING TOOLS

The tools of Web mining allows to download the data from the web, also it gathers appropriate information and perfectly adjusted data. In this part we offer various types of instruments used in this field and do not need to encoded by robots in their entirety, which would require a strong knowledge in PHP or Java / Python,JavaScript, HTML, CSS and XPath some of them are free; some of them have trial periods and paid offers.

According to [12] [13] [14] [15], the table following represents some tools used in the web mining their characteristics /tasks, languages and supported Operating Systems.

TABLE2: WEB MINING TOOLS CLASSIFICATION

Tools	Area of Web mining	Langua-ge	Sup-ported O.S	Tasks and characteristics
Screen-Scraper	Web Content Mining	Java Python. Jython. NET,VB, ASP PHP	Linux, Win-dows	It is able to do a search in a database SQL, a SQL server, extract the contents of the Web with a knowledge of the proxy
Web Content Extractor	Web Content Mining	Python	Win-dows	This tool extracts specific data from a website such as business sector figures data about books and articles and download in extracts data from password protected websites.
Mozenda	Web Content Mining	Javriptas	Win-dows	With Mozenda, users can extract, store and circulate data to several destinations.
Web Info Extractor	Web Content Mining	--	Win-dows	It is Easy to characterize extraction assignment, Can run multi-undertaking in same time, Monitor pages and recover new substance
scrappy	Web Content Mining	Python	Linux, Win-dows, Mac and BSD	Scrapy is a collaborative open source framework that allows you to extract data from a website quickly and easily.
R	Web Usage Mining	Scripting: language like Python, Ruby, Perl	UNIX platfor ms Windo ws MacOS	R is a language or a free environment for statistical computing and graphics
Web Site Information Filter System (SIFT)	web usage mining	JAVA Procedu-ral SQL	Linux, Windo-ws,	Web SIFT automatically defines a set of beliefs exploit to find interesting patterns

Tools	Area of Web mining	Langua-ge	Sup-ported O.S	Tasks and characteristics
Web utilization miner	Web Usage Mining	Java	Linux, Win-dows,	It discovers the navigation patterns, visualization and it prepares log files, and creates basic report.
Redwood	web usage mining	Java and EJB techno-logy	Win-dows, Linux	Redwood is an ASP-based tool for web log, it's a open source.

V. WEB MINING APPLICATIONS

Lastly, the domain of web searches has increased, which gives importance and a variety of web mining application domains. We quote:

- *E-Learning*

E-learning provides an opportunity to learn online, this area has to exploit the web mining to improve and increase the quality of the learning process encouraging users [16].

- *Personalization of web content and a web design*

In general, a personalized web site identifies its users, collects information about their priorities and tailors its services to satisfy the needs of the user. And the web design can help to retrieve only relevant documents [13].

- *Fraud detection*

Unauthorized users can be traced using search results of web log data. A user unsuccessfully trying the access to any web site may be an intruder tries to break the password of restricted area of website [13].

- *Identify Web Robots*

Web Robots are software programs that behave like human for target website. These programs are very harmful for websites because they may crack a password or may breakdown the site by continuous fake requests [13].

- *E-commerce*

Several information extracted from the web is used in E- commerce, for example, improving marketing, improving the production and improving the customer relationships

- *Personalized Portal for the Web*

Personalized portal i.e. a web site designed to have the look-and-feel and content personalized to the needs of an individual end-user. Yahoo was the first to introduce this concept that allows us to create custom portals like Yodlee which led to the celebrity Yahoo website [17].

VI. FUTURE DEVELOPMENT OF WEB MINING

Some of the future scopes of web mining are:

- Digital forensics investigations [13].
- Crime investigation [13].
- Automated data cleaning [13].
- Robot detection and filtering [13].
- Transaction identification [13].
- Cloud mining [13].
- Temporal Evolution of the Web [17].
- Web Metrics and Measurements [17].

VII. CONCLUSION & FUTURE WORK

In this paper, we have presented the web data mining concept and techniques with an insight into the findings from the previous research in the data and web mining, we also listed the tools with their characteristics, and then we quote some web mining areas.

Our study aims to apply the latest trends of web mining and data mining in the education and employment field in Morocco, by the cross analysis between skills acquired in university and the skills sought by employers. More specifically, by conducting a quantitative and qualitative study of the skills taught during university period and the skills required by the job market, collected from web portals using web mining techniques and algorithms, then we will study the different variables. Later we will develop a model of proposal prototype to characterize the general properties.

The diagrams below summarize our proposed approach and prototype, which will be improved as this study progresses.

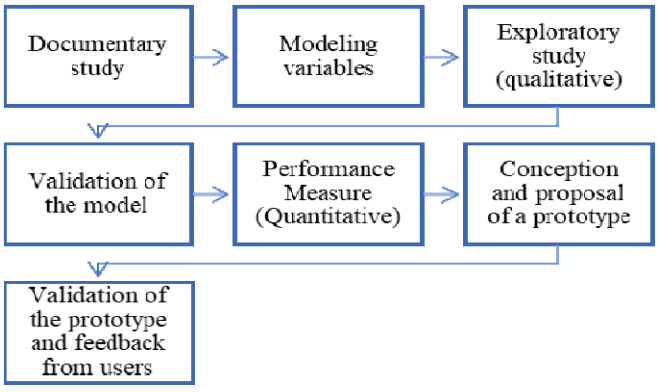

Fig2: A proposal approach of our research (quantitative – qualitative and prototype)

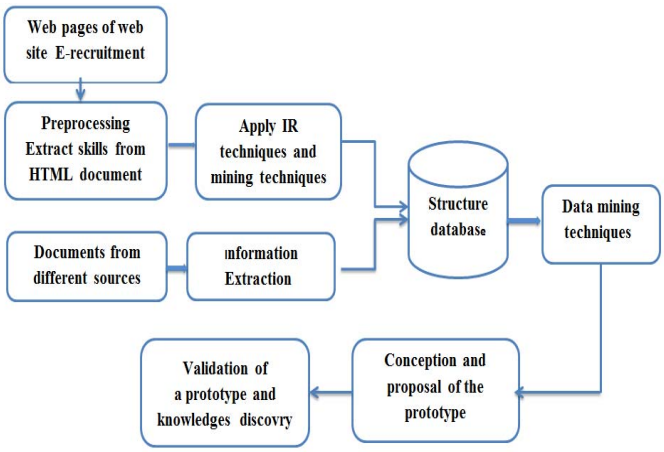

Fig3: A design of our proposal prototype for the cross-analysis between the skills acquired in university training and the skills sought by employers in Morocco.

REFERENCES

[1] Etzioni O..The World-Wide Web: quagmire or gold mine?. Communications of the ACM,1996, Vol. 39, No. 11, pp 65-68.

[2] Kosala R., Blockeel H. (2000). Web mining research: a survey. ACM SIGKDD Explorations,2000, Vol.2, No. 1, pp 1–15.

[3] Madria S.K., Bhowmick S.S., Ng W.K., and Lim E.P. (1999). Research issues in Web data mining. In Proceedings of the First International Conference in Data Warehousing and Knowledge Discovery, DaWaK, 1999, pp 303-312.

[4] Sunil B. Joshi, Dr. Shivaji D. Mundhe "Web Mining and Qualities of a Website Design to Be Evaluated for Customer Browsing Behavior: A Review" ,International Journal of Computer Applications Technology and Research Volume 6–Issue 6, 269-272, 2017, ISSN:-2319–8656

[5] Mrs. S. R. Kalaiselvi1, S. Maheshwari2, V. Shobana3 "Web Mining – Data Mining Concepts,Applications, and Research Directions" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 11, November 2015

[6] Hesham Abdo Ahmed Aqlan, Shoiab Ahmed, Ajit Danti"Death Prediction and Analysis Using Web Mining Techniques" ICACCS - 2015), Jan. 06 – 07, 2017, Coimbatore, INDIA- 978-1-5090-4559-4/17/$31-EEE

[7] Radhika Bairagade, Nikita Afre, Nirmala singh, Durga Bhamare "Smart Crawler: A Deep Web Harvesting Approach"International Journal of Innovative Research in Computer and Communication Engineering-Vol. 4, Issue 4, April 2016

[8] Arjun Sidana , Dr. Himanshu Aggarwal " Review of web usage of data mining in web mining " International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May – June 2017

[9] Shahu, M.S & Leena “A Survey on Frequent Web Page Mining with Improving Data Quality of Log Cleaner”, International Journal of Advanced Research in Computer Engineering & Technology, 2015

[10] Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta, “Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery”, International Journal of Data Mining Techniques and Applications, vol. 02, issue 01, June 2013

[11] Ashok Kumar D, Loraine Charlet Annie M.C.,” Web Log Mining using K-Apriori Algorithm”, International Journal of Computer Applications, vol. 41, no.11, March 2012

[12] Saranya A S, Geetharani S "A Study on Web Mining Tools & Techniques" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2016

[13] Neeraj Kandpal, Ripu Ranjan Sinha, M. S. Shekhawat " A Survey On Web Usage Mining: Process, Application and Tools ",Suresh Gyan Vihar University Journal of Engineering & Technology- Vol . 3, Issue 1, 2017, pp 19-25 .

[14] kalpana wani, archana shirke & parimita das "A Survey On Web Mining Tools" International Journal of Research in Engineering & Technology Vol. 3, Issue 10, Oct 2015, 27-34.

[15] https://www.octoparse.com/blog/7-web-mining-tools-around-the-web/

[16] S.Vidya, K.Banumathy"Web Mining- Concepts and Application" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3266-3268.

[17] Jaideep Srivastava, Prasanna Desikan, Vipin Kumarhttp," Web Mining Concepts, Applications, and Research directions " dmr.cs.umn.edu/Papers/P2004_4.pdf.

[18] https://touriaelouahabi.wordpress.com/web-mining/663-2.