Data Visualization of Moroccan Recruitment Offers

Hanae Mgarbi, Mohamed Yassin Chkouri, Abderrahim Tahiri

SIGL Laboratory, ENSATE, Abdelmalek
Essaadi University, Tetouan, Morocco

`hanae.mgarbi@etu.uae.ac.ma`

Abstract. This paper introduces data visualization of the Moroccan recruitment domain that helps in making decisions, starting with data extraction using Data Miner then data preprocessing, and moving to the data visualization stage using big data techniques and Python libraries.

Keywords: Job offers, Recruitment criteria, XPath, Python, Data visualization.

1. Introduction:

Recruitment in Morocco has gone from the classic procedure of preparing a CV [1] and physically submitting it to companies in the event of spontaneous applications or seeing job offers published in newspapers to apply in the event of job offers being offered. To digitalization through recruitment sites and professional networks has greatly facilitated the recruitment process and optimized working time for job seekers and entrepreneurs. In the labor market field, despite the digitalization it has experienced, there is no reference that can help the actors in this field. There are always questions asked by students and future graduates and entrepreneurs such as: what is the most demanded sector of activity on the job market? Which function is most in demand? The diploma or the level of study is the most requested? The type of contract is the most and/or least requested? …, and there are no answers given to the criteria which remain essential for job seekers and entrepreneurs. Our objective of research work concerns this aspect of the job market which allows to give visibility to the actors of the job market and especially of recruitment and to answer their questions which are always asked. We chose the recruitment site "rekrute", it is the most used recruitment site now in Morocco and we extracted all the published job offers which are 1409 job offers. After extracting the data, we used Big Data [2] techniques to visualize the data in graph format for several purposes such as the creation of decision-making systems [3] for job seekers (even before orienting themselves during the course academic) or entrepreneurs (especially before launching the project). The paper presents the steps to visualize the data through section 2, which concerns the extraction and preprocessing of data, as well as section 3 describes the visualization of recruitment data according to several criteria. Section 4 gives the conclusion and the perspectives.

2. Theoretical background

2.1. Data extraction

Web data extraction [4], also known as web scraping, is the process that extracts data from web pages. It typically involves several functions that work together to collect, transform, and deliver the extracted data to various applications or databases. There are two methods to accomplish the data extraction task either using the XPath [5] or the web wrapper [6]:

XPath. The addressing element in the document tree is The XML Path Language (The XPath) [5], which provides a powerful syntax for addressing specific elements of an XML (Extensible Markup Language) document. This method presents web data in textual and semi-structured form by ordered labeled trees called DOM (Document Object Model) [7]. XPath can be used to address elements of a web page. There are two ways to use XPath: to identify a single element in the document tree, or to process multiple occurrences of the same element. XPath was defined by the World Wide Web Consortium (W3C) [8], making it a widely recognized and accepted tool in the web development community. This allows great flexibility in how elements are selected and

extracted from a web page. For example, in the first case, XPath can be used to target a specific element on a web page, while in the second case, it can be used to target multiple instances of the same type of element sharing the same hierarchical location, like all cells in a table.

Web wrapper. Web data extraction is extracting data from web pages automatically and repeatedly whose content changes and passing the extracted data to an external application. Using a web wrapper [6] for data extraction is usually divided into different functions:

-   web interaction: it is the navigation to predetermined target web pages containing the desired information;
-   wrapper generation and execution: a wrapper is a program that identifies the desired data on target pages, extracts the data, and transforms it into a structured format;
-   Scheduling: which allows the repeated application of previously generated envelopes on their respective target pages;

2.2.   Data preprocessing

Data preprocessing [9], [10] is the set of techniques that prepares data to serve as input to a certain input. this step is mandatory to convert previous data into new data ready to use. If the data is not processed, one could surely have errors during its execution, or offered results will not make sense or will not be considered accurate information. this step aims to:

-   Data cleaning: this involves eliminating unnecessary, incorrect, redundant, or outlying data and all data that is not relevant for analysis.
-   Data transformation: This involves the conversion of data into a desired format for use.
-   Data integration: This is the combination of data from different sources to create a single, consistent set of data.
-   Data normalization: This is the putting of data on a common scale like Min-Max normalization [11] or other methods of normalization.
-   Imputation of missing data: This involves replacing missing values in the data with estimated or predicted values so that important information is not lost.
-   Noise Identification: This involves detecting and preprocessing noisy data that can skew analysis results.

3.   Data extraction and preprocessing:

3.1.   Data extraction

We began our study by extracting data concerning job offers in Morocco, directly from the recruitment site "rekrute". We were able to point to all the information we'll need as the job title; field of activity; mission; level of education; years of experience; city; the type of contract that makes up the data model [12], then they are saved in a structured format in the Excel form file. We have collected 1409 job offers using the browser extension software Data Miner [13] which helps extract the HTML data in the browser window and save it to an Excel spreadsheet file. Data is always secure and private and never leaves the local computer.

To extract data from the web using Data Miner, the web page can be extracted all its details in a single row, or have rows extracted from a list or table, then we located each data that we will need on the web page with identifiers declared in the HTML source code.

There are HTML elements that do not have an ID, which requires applying the addressing element in the document tree: XPath which is able to retrieve powerful, standardized syntax and flexibility in element selection to locate the html element even if there are no IDs.

The result is the Excel file with the data structured inline format. Each line represents a job offer published on the site. And each column represents a characteristic of the job offer such as job title, company representation, mission, and date of publication on the site, sector of activity, function, years of experience, level of study, and the type of contract.

3.2. Data preprocessing

The cleaning of data extracted from the web is an essential step to use for several purposes. We started with the first operation, which is the deletion of rows with missing values so as not to noise the subsequent preprocessing. In the "rekrute" recruitment site, the job title is composed of the profile title, gender, and city. We retrieved the genre and the city, and created two columns on the Excel file to store them for later use. A program is written in Python [14] to add two columns to the file with their names, and also to tokenize the job title which contains the information of the city where the job offer is located and the gender requested by the position, the code is as follows:

```
from openpyxl import load_workbook


wb_access = load_workbook('AllDataExtraction.xlsx')

sheet = wb_access.active

sheet.insert_cols(idx=2)

sheet.insert_cols(idx=3)

sheet["B1"] = "Gender"

sheet["C1"] = "City"

wb_access.save('AllDataExtraction.xlsx')

data_rows=data["Title of Job"]

value_job_title=[]

value_job_city_country1=[]

value_job_city=[]

for value in data_rows:

value_list=value.split('|')

value_job_title.append(value_list[0])

value_job_city_country1.append(value_list[1])

value_job_city_country2=value_job_city_country1.split('(')

value_job_city.append(value_job_city_country2[0])

data["City"]=value_job_city

wb_access.save('AllDataExtraction.xlsx')
```

[Program code used for data preprocessing]

4. Data analysis and data visualization:

In this part, the exported data is processed and ready for transformation into statistical graphs using Python libraries such as Matplotlib [15].

4.1. Activity sector:

The job seeker always has the choice to work in his field in different sectors of activity, so the choice of the sector of activity can be among the criteria of the job search. From the data we have extracted, we find as shown in Figure 1, that the IT sector with the most job offers with a percentage of 32.43% followed by the call center sector at 12.49% and the banking and finance at 6.81%. The other sectors such as automotive, education,

distribution, and the consultancy and studies sector have percentages between 5.5% and 3%, the least offering job offers. Job offers for the rest of the sectors remain minimal, such as cosmetics, insurance, agri-food, etc.
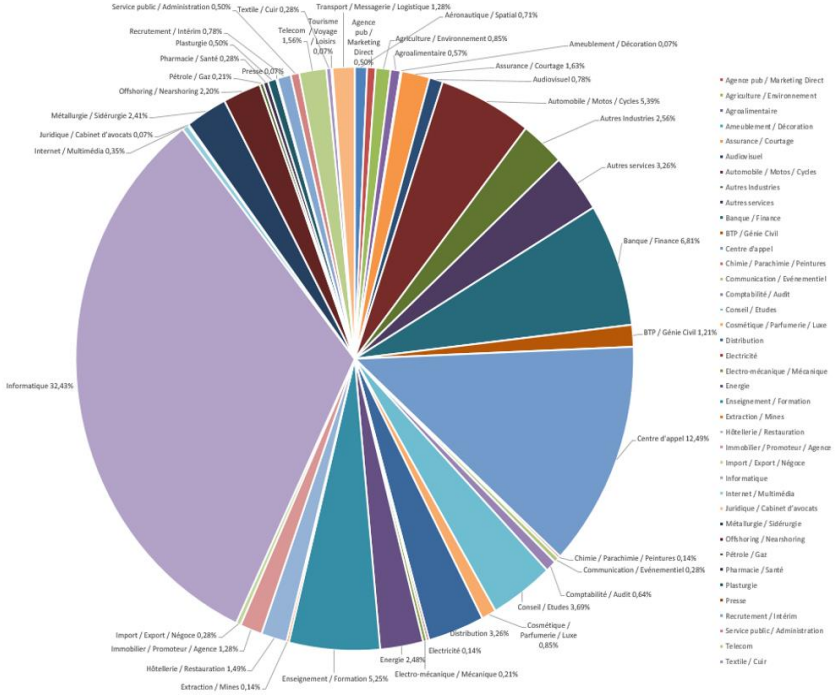

Fig. 1. Job offers by sector of activity

4.2.   Working years:

We know that the criterion of years of experience is the main recruitment and job market in Morocco and elsewhere, so it is always mentioned in the job description and also in the CV.

Our data set contains the data concerning the years of experience, so we can visualize this parameter according to the number of job offers shown in Figure 2 below.
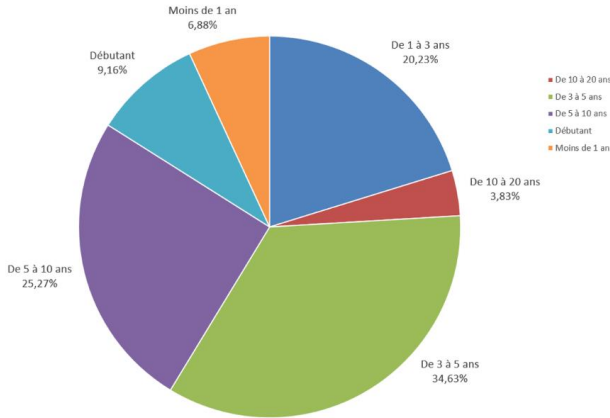
Fig. 2. Number of job offers by years of experience

Positions requiring 3 to 5 years of work are the most important, their number is around 34%, while those requiring 5 to 10 years of work represent 25.27%. Work of 1 to 3 years is requested, representing 20.23%, and experience of less than 1 year represents about 6%. For starters, there is 9.16% that don't require experience. It can be noted that experienced profiles are more in demand on the job market compared to beginners and profiles with less than 1 year of experience that require training and support during work.

4.3.   Studying level:

The job offers always mentioned the level of education in order to properly target the profile suited to the position offered, as shown in Figure 3 the levels of education required for the job offers in our dataset: are 58.69% for the level of education bac+5 and more, this is the most dominant level compared to the other levels of study, this level concerns engineers and masters or doctorate profiles. There are 12.14% of dedicated job offers for the BAC+3 level, equivalent to the license in Morocco. The BAC+2 level of study offers approximately 13% of job offers, most of them for technicians or generalist university graduates. The BAC+4 level of studies is requested by 8.87% of job offers. There are 2.84% of job offers for the BAC+1 level and 4.26% of job offers for the baccalaureate level. We can see that the Moroccan job market published on the recruitment site focuses more on profiles with BAC+5 and more, on the other hand, qualifications before the baccalaureate are more likely to be poached by publications on social networks, etc.
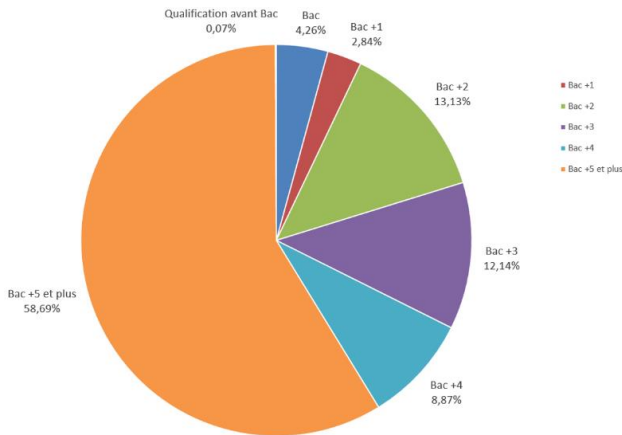

Fig. 3. Number of job offers by the level of study

4.4.   Contract type:

The types of contracts used in Morocco are the open-ended contract (CDI), the fixed-term contract (CDD), the freelance contract, or another type of contract such as the Anapec contract offered for beginners. There are also internship offers that are published on the recruitment site, which may or may not be pre-employment internships. In Figure 4 below, we have the index "?" concerns job offers that do not mention the type of contract so their contract types remain unknown.
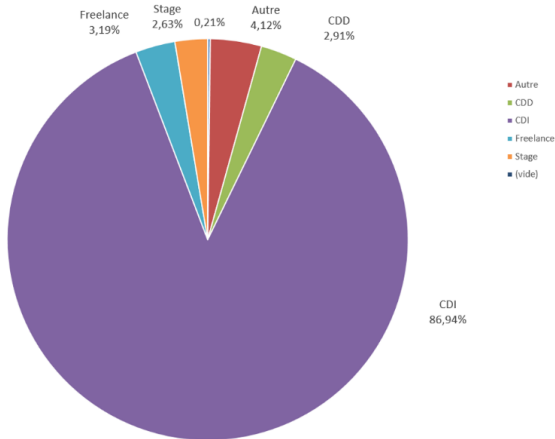

Fig. 4. Number of job offers by the level of study

We note that the type of contract most offered by job offers in Morocco is the CDI with a rate of 86.94%, the other types of contract remain minimal with 3,19% for Freelance and 2,91% for CDD.

5.   Conclusion:

This paper presents the steps of extracting data from the recruitment site "rekrute" using the Data Miner tool which helped us to extract all the necessary information then we moved on to data preprocessing to have the data ready to use, then, we used Python libraries to visualize the data according to several criteria used by the field of the labor market in order to present the analyzes of the data to the actors of recruitment which are useful for making decisions. Future work will be on several sources of data such as other recruitment sites used by Moroccans with different formats (e.g. PDF format) to have more general visibility on the job market in Morocco.

References

[1]   Sanchez M, Kephart K, Jones K and des Jardins M (2020) A Methodology to Analyze Self-Reflection in E-Portfolios. In: 2020 IEEE Frontiers in Education Conference (FIE), pp 1-5

[2]   Zerhari B, Lahcen A.A and Mouline S (2015) Big data clustering: Algorithms and challenges. In: Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15).

[3]   Rizk Y, Awad M and Tunstel E.W (2018) Decision Making in Multiagent Systems: A Survey. IEEE Transactions on Cognitive and Developmental Systems. vol 10. no 3. pp 514-529. doi: 10.1109/TCDS.2018.2840971

[4]   Baumgartner, R., Gatterbauer, W., & Gottlob, G. (2009). Web Data Extraction System. Encyclopedia of database systems, vol. 1.

[5]   Ferrara E, De Meo P, Fiumara G, and Baumgartner R (2014) Web Data Extraction, Applications and Techniques: A Survey. Knowledge-Based Systems, vol. 70, pp 301-323, doi: 10.1016/j.knosys.2014.07.007.

[6]   Azir, M. A. B. M., & Ahmad, K. B. (2017). Wrapper approaches for web data extraction: A review. 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI). doi:10.1109/iceei.2017.8312458

[7]   Keith J (2006) DOM scripting: web design with JavaScript and the Document Object Model. Apress

[8]   Brooks, T. (2017). World wide web consortium (W3C). In Encyclopedia of Library and Information Sciences (pp. 5034-5038). CRC Press.

[9]   García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. Springer Publishing Company.

[10]   Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.
[11]   Henderi, H., Wahyuningsih, T., & Rahwanto, E. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *International Journal of Informatics and Information Systems, 4*(1), 13-20. doi:10.47738/ijiis.v4i1.73
[12]   Goodchild M. F (1992) Geographical data modeling. Computers & Geosciences. vol 18. no. 4. pp 401-408. doi: 10.1016/0098-3004(92)90069-4
[13]   "Data Miner." dataminer.io.
          https://dataminer.io/ (accessed Dec. 10, 2021).
[14]   Python W (2021) Python. Python Releases Wind. vol. 24, Citeseer.
[15]   Bisong E (2019) Matplotlib and seaborn. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. pp 151-165.