IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Survey on Extracting and Analyzing Online Data for Enhancement in Recruitment Process

[1]Hardik Kokate, [2]Tushar Dhawal, [3]Dhatrik Reddy,[4]Tanishq Bhargava,[5]Sparsh Giri,
[6]Mrs. Dhanashree Phalke,[7]Mr. Jayant Shimpi
[1-5]Student, [6-7]Assistant Professor,
Department of Computer Engineering, D.Y. Patil College of Engineering Akurdi , Pune, India

Abstract: The paper is a review on recommendation system for the selection of an interviewed candidate. The paper contains the survey of various methods & techniques used by different author to derive insights from real time data . Looking at the current interview scenario , it is biased towards "candidate's performance during interview" and doesn't take other factors into account such as candidate's competitive coding abilities, contribution towards developer community and so on. Therefore, a recommendation system is needed to make hiring process more efficient and effective. A system that will effectively use data available on global platforms and does evaluation , aggregation & visualization of real time data . Thus, the system would help the interviewer get a full view of candidate's abilities and therefore helps them make an unbiased and informed decision.

Keywords – data evaluation , data aggregation , data visualization , hiring process , effectiveness

I. INTRODUCTION

In today's era data online is of more importance than the data offline. In [1] author has mentioned the more than 50% of the world population will be on the internet therefore the data requirement, extraction and analysis are some of the biggest components that are to be measured and deployed. Talking about the tech related member in the tech community, they invest most of their time learning and coding online, though it might not be of that importance to others, but it certainly can help an interviewer to make a better decision. Extracting the data from the internet by using a technique called web scrapping[6]. Now what's web scrapping? So, Web scraping is the process of collecting structured web data in an automated fashion. It's also called web data extraction. Web scrapping is mainly used in price monitoring, price intelligence, news monitoring, lead generation and market research among many others. In general, web data extraction is used by people who want to make use of the vast amount of publicly available web data to make smarter decisions. In case of hiring process , it would be the data of a technical guy working on some globally available platforms like GitHub, Hacker-rank etc. These scrapped data could be further used to analysis candidate's performance and thereby giving interviews an idea about candidate's contribution on such platforms. There will be a discussion on which techniques and methods are used by different author for scraping, visualization of data and furthermore.

II. LITERATURE SURVEY

There has been lot of discussions on data extraction and how it can be used to improve efficiency of the results that users want to search. In [1], author mainly focuses highly on modularized framework for information retrieval, processing and presentation of data. Communication between the modules is done similarly to the pipes and filters architectural pattern i.e., the modules are chained so that the output of a module is the input of another one. Author [1] proposed three-tier application where the communication between tiers is achieved via micro web services and where each component is highly modularized. The component that the author proposed are the distributed crawler, the persistent storage (i.e., Data Server) and the product web server. The final goal is to have a fully autonomous system that detects websites, correctly extracts required data and links those data on different websites so to provide an excellent view to user by having a vast variety of information related to subject.

The author [2], proposed a recommendation system requires only interviewer's preference vectors (v) in addition to other known variables i.e., number of posts (K), number of interviewers (n), total number of candidates (m) and their objective evaluation scores. The proposed recommendation system uses two algorithms, the Hungarian Aggregated Method (HAM) and the Greedy Aggregated Method (GRAM) for computing the final decision
vector (r). The final decision vector contains selected candidates in ascending order. The GRAM follows the local best approach hence compute an efficient solution but not optimal as compared to HAM that follows the global best approach hence require computational complexity but finds an optimum solution.

Further, an effective attribute-based opinion algorithm has been proposed which enables the user to make well-informed decisions [2]. The proposed tool is a generalized tool, nevertheless the comments about products from e- commerce website is considered for the sake of discussion. In current ecommerce websites namely Amazon, Flipkart and Snapdeal, the recommendation system works in two ways. To begin with, they provide an overall rating for the product based on customer satisfaction. Added to that, they apply the Associative- Rule-Mining algorithm to track customer buying habits, to suggest the items that are frequently bought together. From these two methodologies, it can be inferred that there is no recommender system which gives attribute-wise opinion for the product. Score vector is a vector of k entries where each value corresponds to the candidate score. Score vector is used to measure the satisfaction score of the interviewers for a specific position [2].

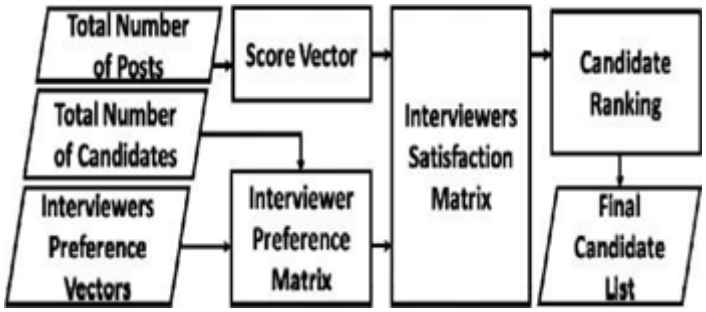

Figure 1: Ranked Candidate Recommendation System [2]

Recommender system is the part of web usage mining in which it predicts the "rating" or "preference" that a user would give to an item based on their interest. The main idea and goal of recommender systems is to utilize various sources of data to infer customer interests [5]. There are different approaches to web scraping. Collaborative filtering collects opinions of users and then recommend products or items using people's rated similarity. The users which have a similar opinion are the contributors. Collaborative filtering recommender systems mainly take details of more than one common product or items to define users, which affect the result. Collaborative filtering techniques mainly work on very huge amount data sets available on web. Content based filtering techniques recommends products based on a comparison between the detail data of the products with user profile.

The content of every product or item is defined as a group of terms or details of attributes, normally the characters which are available in a document. Hybrid filtering combines different recommender techniques to achieve good system utilization to avoid some problems and limitations of pure recommendation system techniques. The main idea behind hybrid filtering is to merge different algorithms will provide more accurate and effective recommendations than a single algorithm as the disadvantages of one simple technique can be overcome by another technique [5]. These generated user profile which represented the similar attributes and created by analyzing the details of products on which user gives the ratings. For score generating proposed recommender system considers interviewers preferences and the sequence in which they wanted their preferred candidates to appear in the final selection list. However, finding an agreement (maximum satisfaction) among multiple preferences and variation in ordering for preferred candidates makes the problem more challenging.

Data analysis is the method of extracting solutions to the problems via interpretation of data. The analysis process comprises of discovering problems, determining which method can help in finding the solution to the interesting problem and convey the result. So, in order to achieve data extraction or mining there is a process called web scrapping. Web scraping is the process of extracting or scrapping data off the web programmatically and converting it into a structured dataset. Web scraping allows for larger amounts of data to be collected in a lesser span of time and in an automated fashion that reduces error [6].

A web crawler or spider is the very first part of data extraction process. Web crawlers are used by various sites across the internet and search engines to change their contents regularly. The procedure of crawling first starts with the list of URLs which help to crawl, these URLs are mainly called as seeds URLs. To extract large and unstructured data from the web there are some data extraction technique and tools that can easily extract large unstructured data and convert them into a meaningful and structured format data extractor is used. It is basically a technique of extracting user- required information from the websites. The extraction process first gets started with indexing or generally used crawling. Data gets crawled from the web by using a web crawler. In the crawling process crawler could also transform an unstructured data on the web page into a structured data [6].

There are two ways for web scrapping, proposed by the author [3], first HTML parser (screen scraping, where you extract data from source code of a website) and through APIs (where a website offers a set of structured HTTP requests that return JSON or XML files). If the websites where the developer wants to extract data offers the APIs then an easy HTTP request are going to be sufficient for web scrapping and in other cases HTML parsing can be helpful. There are various difficulties in the process of scrapping. Sometimes it becomes difficult for student to update data which is first challenge. Second challenge is to handle missing values in the data since there is large data which is to be collected. A third challenge is the need for instructors to remember that they have no control of the online data. Thus, it is important to note potential issues related to connectivity and content. Last challenge is that instructors do not have control over the content of the web [3]. Without web scraping skills and rules, students are limited to hand scraping or to datasets that already come in as CSV or Excel files. In some situations, such files might contain

exactly the data students want. But in many situations having a limited amount of data that is already presented in a structured format can limit students' options.

In [7], author discuss issues of authentication by giving the example of Twitter API. Author mentions that the Twitter API allows to access the whole social graph, i.e., the graph representing users and their connections in Twitter, without authentication. Approaches based on the scraping of HTML pages can overcome the limitations above even if they are more complicated to design and implement. Collecting data from multiple Web platforms, the main technical challenges encountered by Web Data Extraction techniques to collect data from multiple Social Web platforms consists of linking information referring to the same user or the same object. It can be queried through an HTTP request having a URL called node as its parameter. The node specifies the URL of a Web page of a user u. The Google Social Graph API can return two kinds of results: (1) A list of public URLs that are associated with u; for instance, it reveals the URLs of the blog of u and of her Twitter page. (2) A list of publicly declared connections among users. For instance, it returns the list of persons who, in at least one social network, have a link to a page which can be associated with u [7].

Another key point is that which programming language to use for this data extraction. There are several languages that assist developers to scape data, the most common and efficient tool is the web crawler Scrapy utilizing the programming language Python adaptation 3.6. Web scraping software's such as Scrapy is available for whenever ease of access needed by the user and, it's an open-source web-crawling framework for the collection of any data as per user's needs. In a paper [8], author has done some projects in which he uses a methodology, to gather all the data extracted from various sources by using the vivid features of the web crawler scrapy using the scripts written in python language and further analyze it as per the requirements of the customer where the data is stored in the company's database.

The web crawler scrapy which is python based also may help in retrieving the desired result as the author analysis process by specific code and give the desired URL for the iteration to perform for scrapping the data from the source URL. The overall results of that project turn out to be helpful to understand. The Web scrapy extracted the data and made into csv file format. The script which was written to extract the data turned out to be both of finding each of these sources provided with great ease. Moreover, the analysis done has shown the most searched content in the site taken for test in the percentage format [8].

To visualize these scrapped data a system should have certain tools for visualization. Existing visualization tools are capable of visualizing only structured data, but more than 80% of business information is in unstructured format. Also, there is no integrated tool for structuring and visualizing Unstructured-Data in one- shot [4]. A tool called "ScrAnViz" is developed that could be applied to any domain accepting the domain specific and user specific attributes as inputs and plot graphs accordingly. The Unstructured data is obtained by web scrapping [4]. It is developed as single software tool which mines the interested data from the Unstructured- Data and visualizes it in a format required by the user. The Unstructured-Data can be obtained from any online website. The data extraction module accepts the URL of the website from where comments and reviews data must be scrapped. Web scraping has been implemented using Jaunt API. Create a User Agent object inside Jaunt package and call the visit () method by specifying the URL, the input received through the HTML form. The visit () function returns a Document object named userAgent.doc [4]. All these techniques, processes and tools will help to form a system that can provide the user a better view of their choices in any aspects related to their searches and requirements.

III. CONCLUSION

This paper gives a review on system and methods during recruitment and other different processes. We have surveyed a wide range of methods and techniques used by various authors. This paper also gives a summary of the work done till now.
There are many significant processes such as web scrapping, visualization and analysis of data that are significant to this proposed work and also to mention that the Python 3.6 is one of the best programming languages that is suitable for this kind of work. Looking at the proposed system by various authors, it is found that there is a lack of data analytics capabilities in recruitment processes. So, there is a need of system that will help in better decision making and thereby the recruitment process becomes more effective.

REFERENCES

[1] Alexandrescu, Adrian, "A distributed framework for information retrieval, processing and presentation of data",2018 22nd International Conference on System Theory, Control and Computing (ICSTCC) , pp. 267-272.

[2] S. Hammad, S. Ahmed, M. A. Abbas, I. Abbasi and M. T. Jilani, "A Recommendation System using Interviewers Preferences for Ranked Candidate Selection," 2018 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2018, pp. 230 -234.

[3] Dogucu, Mine & Cetinkaya, Mine. (2020). "Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities.", 2020 Journal of Statistics Education. 1 -24.

[4] Sriraghav, K & Jayanthi, Sriharsha & Vidya, N & Enigo, Felix. (2017)." ScrAnViz — A tool to scrap, analyze and visualize unstructured-data using attribute-based opinion mining algorithm", 2017 International Conference on Innovations in Power and Advanced Computing Technologies. 1- 5.

[5] F. Mansur, V. Patel and M. Patel, "A review on recommender systems," 2017 International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS), Coimbatore, 2017, pp. 1-6.

[6]Diouf, Rabiyatou & Sarr, Edouard & Sall, Ousmane & Birregah, Babiga & Bousso, M. & Mbaye, Sény. (2019). "Web Scraping: State-of-the-Art and Areas of Application". 2019 IEEE International Conference on Big data, pp. 1 -3.

[7] M. S. Parvez, K. S. A. Tasneem, S. S. Rajendra and K. R. Bodke, "Analysis of Different Web Data Extraction Techniques," 2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai, 2018, pp. 1-7.

[8] D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 450-454.

[9] David Mathew Thomas, Sandeep Mathur, "Data Analysis by Web Scraping using Python", Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019], 2019.

[10] K.Sundaramoorthy, R.Durga and S.Nagadarshini, "Newsone- An Aggregation System for News Using Web Scraping Method", International Conference on Technical Advancements in Computers and Communications (ICTACC), 2017.