Company Predictor Using Resume

Yusoof Ali[1], Suraj RK[2], Noor Ahmed[3],Rajendra M4
[4]Assistant Professor
[1, 2, 3, 4] Atria Institute of Technology

Abstract- *A user needs to contribute time and energy looking for their dream work that matches their skills. Because of this the talent and ability of the user may go unnoticed by an organization. Our solution additionally tackles the issue of a recruiter experiencing a countless number of resumes every day. Along these lines, organizations are not ready to employ the best possibility for their association and need to contribute additional time and asset to find the best candidates for their association. The primary undertaking is to separate the occurrence of different data, for example, keywords, skills, work experience, etc. were made reference to in the CV. The keywords should be positioned by whether they are required and additionally rank them depending on their order of the CV. A UI is created for the user to submit their resume. In the proposed model we develop a system which will suggest the job for the user according to the user's resume as well as allow recruiters to view the candidate list based on the job requirements.*

Keywords- Web, Resume Parser, KNN Query Processing, Content Keyword Weightage.

I. INTRODUCTION

In this busy world time is everything for people and it is very difficult for people to find appropriate company they need to apply for a job as they are unaware of the requirement of the company due this people spend a lot of time in searching for the company by visiting websites of different companies. But by using the website, people can save time as they just need to upload there resume and they will get list of companies they can apply for a job. This website uses many technologies to predict appropriate company such as web scrapping, data mining and regression algorithm.

A user needs to contribute time and energy looking for their dream work that matches their skills. Because of this the talent and ability of the user may go unnoticed by an organization. Our solution additionally tackles the issue of a recruiter experiencing a countless number of resumes every day. It might likewise be that the user's abilities don't coordinate the organization's prerequisites or maybe the user applying for the job, isn't sufficiently skilled for the particular employment.

Along these lines, organizations are not ready to employ the best possibility for their association and need to contribute additional time and asset to find the best candidates for their association. The primary undertaking is to separate the occurrence of different data, for example, keywords, skills, work experience, etc. were made reference to in the CV. The keywords should be positioned by whether they are required and additionally rank them depending on their order of the CV.

KNN query processing is used to investigate several aspects of keyword processing from the resume. We will study the cost of (k, δ)-Range algorithm, which mainly contributes to the server-side cost. We will show the overall cost distribution over the cloud side and the proxy server. Very often, especially when measuring the distance in the plane, we use the formula for the Euclidean distance.

Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. The enhancement of predictive web analytics calculates statistical probabilities of future events online. Predictive analytics statistical techniques include data modelling, machine learning, AI, deep learning algorithms and data mining. It is used in the web app to seek job according to the users resume.

The existing solution is to use third-party recruiters but that comes with several disadvantages. Disadvantages of third-party recruiters:

1. The cost of recruiting

Working with third-party recruiters requires a fee for their services. If you use a temporary to hire employee you are paying the mark up over the employee’s salary which covers the administrative costs, taxes, and some profit for the agency. If you chose to hire directly from their pool of candidates you will be required to pay a percentage of the employee’s salary directly to the staffing company. These costs can be negotiable but are unavoidable. Before utilizing third-party recruiters, consider your budget and assess ROI potential.

2. Lack of control

For many HR or hiring managers it is difficult to give up the control over the hiring process. You want to read every resume and interview every candidate.

While third-party recruiters are in the business of consolidating the top candidates for your review it can be frustrating to take yourself out of the process altogether.

3. Indirect candidate access

Without being part of the staffing experience from the very beginning it is difficult to assess the individual employee before they arrive at your office.

4. Communication issues

Probably the biggest disadvantage of working with a third-party recruiter is the communication. Each of the previous three bullets is also a product of this potential disaster. Individuals communicate in different ways and if you and the service are not on the same page it can be difficult to come to a shared agreement.

The best way to solve these problems is to make use of HRMS software - either within your organization, or your recruitment company's system. The more documentation you have on all of these topics the less likely there will be disagreements and other communication breakdowns. Working with a third-party recruiter may or may not be right solution for you so do your homework and track your conversations and decide the best process for your next hire.
Seeing all these above listed issues we have decided to make a website which can overcome all these above listed issues and make a website which help in recruiting the candidate in cost effective process by using our website.

We create a UI is for the user to submit their resume. Before processing the resume, steps are taken to extract text from the resume. Unnecessary words are removed from the user's resume and the appropriate keywords are searched. The calculation of the keywords is done based on the content weight age of keywords and the Keywords are then converted into a hash code by using the MD5 algorithm and placed into an Index Array. This preparation information is fed into a model. In the proposed model we develop a system which will suggest the job for the user according to the user's resume as well as allow recruiters to view the candidate list based on the job requirements.

The user will be uploading their curriculum vitae to the website. Using regression algorithm and time series analyses it is possible to parse the cv into the database. The data obtained from the cv after parsing will compared with the company details which has been stored in the database. Using various machine learning algorithms, it will analyze whether the data provided by the candidate is meeting the requirements of the particular company. Based on the statistics the company predictions will displayed to the user.

II. RELATED WORKS

Ashish Dutt, MaizatulAkmar Ismail and TututHerawan used traditional data mining algorithms and concluded that it cannot be directly applied to educational problems, as they may have a specific objective and function. This implies that a preprocessing algorithm has to be enforced first and only then some specific data mining methods can be applied to the problems. One such preprocessing algorithm in EDM is clustering. Many studies on EDM have focused on the application of various data mining algorithms to educational attributes. Therefore, according to their paper provides over three decades long (1983–2016) systematic literature review on clustering algorithm and its applicability and usability in the context of EDM. Future insights are outlined based on the literature reviewed, and avenues for further research are identified.

As an interdisciplinary field of study, Educational Data Mining (EDM) applies machine-learning, statistics, Data Mining (DM), psycho-pedagogy, information retrieval, cognitive psychology, and recommender systems methods and techniques to various educational data sets so as to resolve educational issues [1]. The International Educational Data Mining Society [2] defines EDM as ''an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in'' (p. 601). EDM is concerned with analyzing data generated in an educational setup using disparate systems. Its aim is to develop models to improve learning experience and institutional effectiveness. While DM, also referred to as Knowledge Discovery in Databases (KDDs), is a known field of study in life sciences and commerce, yet, the application of DM to educational context is limited.

Winai Nadee, Korwait Prutsachainimmit made use of web archives, particularly rising of JavaScript Web advancement innovation that has altogether influenced the best approach to insert and rendering information of Web pages. In this paper, we propose a plan and execution of another Web Data Extraction framework that goes for removing information from JavaScript Web applications. The proposed framework empowers clients to choose significant information

from online Web records by characterizing information extraction standards and information change designs. The extraction motor naturally rub and changes semi-structure information into social information. The starter assessment results demonstrated that our proposed framework has effectively separate information from present day JavaScript Web applications.

Celikovic&Lukovic. used K-means clustering to an electronic log of 4096records featuring information on student login/logout actions according to the time table of class meetings. After clustering, it was found that students with high levels of spatial deployment (seat selection) have 10% higher assessment scores as compared to students with low spatial choice. Students typically write in the margins of books about their understanding of the text presented. This activity is called as 'annotation'. In one of a kind studies proposed by Ying, et al. [64] two simple biology inspired approaches of chromosome behavior were applied to 40 students' annotations text. Then, they clustered the data based on the similarity between annotations using K-means clustering and hierarchical clustering methods. They found that their proposed approaches are more efficient than the generic hierarchical clustering algorithms.

Buehl& Alexander [133] studied students' epistemological beliefs about knowledge acquisition and their learning process. The objective of this research was to examine epistemological beliefs and students' achievement motivation. The unique aspect of this study is that rather than examining whether or not individual beliefs are related or co-related to performance and motivation; the authors tested different configurations of beliefs that were related to students 'competence beliefs, achievement values and text-based learning.

The sample size was 482 undergraduate students whose beliefs on knowledge, competency levels, and achievement values in history and mathematics were analyzed. Ward's minimum variance hierarchical clustering technique was used to analyze the data. The results revealed that students with different epistemological beliefs vary with their competency beliefs and achievement values. They suggested that future research may apply cluster analysis to different configurations of beliefs related to various aspects of student learning

Dongcheng Peng, Tieshan Li, Yang Wang, C. L. Philip Chen proposed a solution of extracting information from the web or web mining to be one solution to gathering culinary tourism information that uses the web as a data source. The use of traditional methods, such as surveys, interviews or questionnaires, is often constrained by funding and geographical problems (Peng, Li, Wang, & Chen, 2018). In today's digital era, where a lot of information is shared by people through the internet, web mining is expected to be an effective solution for finding information or gathering information. Web crawling will search for web pages through URLs (Uniform Resource Locators), and return the data to the user directly. Users do not need to access information by browsing web pages one by one so as to save time and effort and improve.

Web and social media has open opportunities for business and organization to get the significant value that leads to efficient operations. As a result, Web Data Extraction has become an important tool for gathering and translating semi-structured documents into valuable information. However, one of the major challenges is dealing with changes from Web documents, especially emerging of JavaScript Web development technology that has significantly affected the way to embed and rendering data of Web pages. In this paper, we propose a design and implementation of a new Web Data Extraction system that aims for extracting data from JavaScript Web applications. The proposed system enables users to select valuable data from online Web documents by defining data extraction rules and data transformation patterns. The extraction engine automatically scrapes and transforms semi-structure data into relational data. The preliminary evaluation results showed that our proposed system has successfully extract data from modern JavaScript Web applications.

Zhaoli Wang, Xinhuai Tang, Delai Chenconsidered job interview as one of the most essential tasks in talent recruitment, which forms a bridge between candidates and employers in fitting the right person for the right job. While substantial efforts have been made on improving the job interview process, it is inevitable to have biased or inconsistent interview assessment due to the subjective nature of the traditional interview process. To this end, in this paper, we propose a novel approach to intelligent job interview assessment by learning the large-scale real-world interview data. Specifically, they develop a latent variable model named Joint Learning Model on Interview Assessment (JLMIA) to jointly model job description, candidate resume and interview assessment. JLMIA can effectively learn the representative perspectives of different job interview processes from the successful job interview records in history. Therefore, a variety of applications in job interviews can be enabled, such as person-job fit and interview question recommendation. Extensive experiments conducted on real-world data clearly validate the effectiveness of JLMIA, which can lead to

substantially less bias in job interviews and provide a valuable understanding of job interview assessment.

Zhang et al., 2014], talent mapping [Xu et al., 2016], and market trend analysis [Zhu et al., 2016] enhanced the quality and experience of job interview. A critical challenge along this line is how to reveal the latent relationships between job position and candidate, and further form perspectives for effective interview assessment. Intuitively, experienced interviewers could discover the topic-level correlation between job description and resume, and then design the interview details to measure the suitability of applicants. For example, a candidate for "Software Engineer", who has strong academic background, might be interviewed with questions not only about "Algorithm", "Programming", but also "Research". Meanwhile, compared with the technical interview, the vocabulary of comprehensive interview could be largely different.

III. OBJECTIVES AND SCOPE

The general objective is to build up a web application that causes CV to find the coveted activity effortlessly without squandering their time in visiting different sites. The site permits to likewise fending off extortion enrolment specialists from hopefuls.

This application causes organizations to spare their time and exertion to arrange occasions and drives for an occupation since they can discover numerous applicants that are qualified for the activity through this site. The individual doesn't have to go searching for different organizations; rather he/she can utilize the application to channel/waitlist organizations. The application makes utilization of the resume of the client and demonstrate the organizations that best match the client's abilities and experience. The client can apply to those organizations anticipated and have a higher possibility of getting enrolled. Subsequently, the organizations can enlist representatives that meet the organization prerequisites.
The client can likewise check whether their CV is valuable for enrolment of a specific organization or if their CV can locate the best organizations. In this manner the client can enhance their abilities or encounter and apply for the organization later. Organizations can include themselves and make work postings on the application. Therefore, it enables organizations to enroll representatives that meet their necessity.

The following is divided into several sections, first we discuss about the problem statement, second, we discuss about the scope of the work, next the implementation and technologies used are shown. Finally, the data flow and results are discussed.

A. PROBLEM STATEMENT

In this busy world time is everything for people and it is very difficult for people to find appropriate company they need to apply for a job as they are unaware of the requirement of the company due this people spend a lot of time in searching for the company by visiting websites of different companies and now even the cost of recruitment is very expensive as working with outsider enrolment specialists requires a charge for their administrations.

In the event that you utilize an impermanent to procure worker you are paying the increase over the representative's compensation which takes care of the authoritative costs, charges, and some benefit for the organization. In the event that you contracted specifically from their pool of hopefuls you will be required to pay a level of the worker's compensation straightforwardly to the staffing organization. These expenses can be debatable however are unavoidable. Prior to using outsider selection representatives, consider your financial plan and evaluate ROI potential. But by using our website "company predictor people can save time as they just need to upload their resume and they will get list of companies they can apply for a job; they can even save their money from fraud recruiter.

Absence of control-For some HR or enlisting supervisors it is hard to surrender the command over the procuring procedure. You need to peruse each resume and meeting each hopeful. A recruiter needs to experience numerous quantities of resumes, which may not coordinate the organizations prerequisites or maybe the user applying for the organization, isn't sufficiently talented for the particular employment.

While outsider scouts are in the matter of solidifying the best possibility for your audit it very well may baffle remove yourself from the procedure by and large. Correspondence issues-Most likely the greatest weakness of working with an outsider enrolment specialist is the correspondence. Every one of the past three slugs is likewise a result of this potential catastrophe. People impart in various ways and in the event that you and the administration are not in agreement it very well may be hard to go to a mutual assertion.

Another issue because of this is that companies are not able to hire the best candidate for their organization and need to invest additional time and resource to discover best candidates for their organization.

B. SCOPE

Interface design-describes the structure and organization of the user interface. Includes a representation of screen layout, a definition of the modes of interaction, and a description of navigation mechanisms. Interface Control mechanisms- to implement navigation options, the designer selects form one of a number of interaction mechanism:

- Navigation menus
- Graphic icons
- Graphic images

Interface Design work flow- the work flow begins with the identification of user, task, and environmental requirements. Once user tasks have been identified, user scenarios are created and analyzed to define a set of interface objects and actions.

Aesthetic design-also called graphic design, describes the "look and feel" of the WebApp. Includes color schemes, geometric layout. Text size, font and placement, the use of graphics, and related aesthetic decisions. Content design-defines the layout, structure, and outline for all content that is presented as part of the WebApp. Establishes the relationships between content objects.

Navigation design-represents the navigational flow between contents objects and for all WebApp functions.

Architecture design-identifies the overall hypermedia structure for the WebApp. Architecture design is tied to the goals establish for a WebApp, the content to be presented, the users who will visit, and the navigation philosophy that has been established. WebApp architecture is defined within the context of the development environment in which the application is to be implemented.

C. IMPLEMENTATION

The primary undertaking is to separate the occurrence of different data, for example, keywords, skills, work experience etc. were referred to in the CV. The keywords should be positioned by whether they are required and additionally rank them depending on their order of the CV.

This data forms the preparation information and is fed into a model. We utilize numerous CVs which were successful to prepare the model. The resume is uploaded through the web app, the resume is cleansed of unwanted data and the appropriate keywords are stored into a database.

The keywords are obtained using regular expressions. Details such as email, phone number, College, Skills, work experience are retrieved using regular expression and NLP. These keywords are converted into Hash Code and stored into an index Array. Java Hash Map is used which is a hash table-based implementation of Java's Map interface. A Map is a collection of key-value pairs. Java Hash Map allows null values and the null key. Hash Map is an unordered collection used to store the keywords and their Hash Code's.

The use case is shown in Figure 1, describing the user interactions with the website.

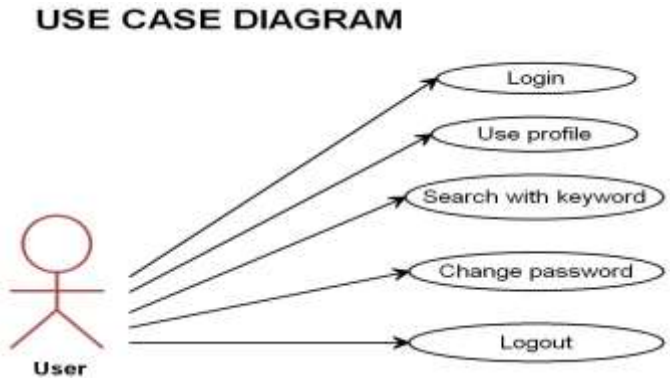

Fig 1. Use Case Diagram

The description of the module is as follows:

1) Keyword Indexing

Remove un-necessary words from the user resume and Find the keywords.
Calculate the Content Weight age of keywords Convert the Keywords into hash code by using MD5 algorithm; place the hash code in Index Array.

2) Search with Keyword-user

Admin has to Input the search keyword. Convert the keyword into HashCode. Send generated hash codes to server, based on the Hash codes received server has to check the keyword index and if any matching candidates are available, list all the jobs
to the user.

Calculations are done in the server by fetching the HashCode of the user's resume and using it to compare it with the details of the company's requirements. The company details are stored into MySQL database. The companies job requirements are compared with the HashCode of the users resume and matched companies are displayed and ranked to show the best job for the user.

KNN query classification is used to rank the companies based on the weightage derived from the

calculation. We investigate several aspects of KNN query processing. We will study the cost of (k, δ)-Range algorithm, which mainly contributes to the server-side cost. We will show the overall cost distribution over the cloud side and the proxy server. We will show the advantages of KNN-R over another popular approach: the Casper approach for privacy-preserving KNN search.

Step 1: Compute the Euclidean distance for one dimension. The distance between two points in one dimension is simply the absolute value of the difference between their coordinates. Mathematically, this is shown as |p1 - q1| where p1 is the first coordinate of the first point and q1 is the first coordinate of the second point. We use the absolute value of this difference since distance is normally considered to have only a non-negative value.
Step 2: Take two points P and Q in two dimensional Euclidean spaces. We will describe P
with the coordinates (p1, p2) and Q with the coordinates (q1, q2). Now construct a linesegment with the endpoints of P and Q.
Step 3: The distance between 2 points P = (p1, p2) and Q = (q1, q2) in two-dimensional space is therefore
((p1 - q1) ^2 + (p2 - q2) ^2) ^ (1/2).

Step 4: Extend the results of Step 3 to three dimensional spaces. The distance between points
P = (p1, p2, p3) and Q = (q1, q2,q3) can then be given as ((p1-q1)^2 + (p2-q2)^2 + (p3-q3) ^2) ^(1/2).

Step 5: Generalize the solution in Step 4 for the distance between two points P = (p1, p2, ...,pn) and Q = (q1, q2, ..., qn) in n dimensions. This general solution can be given as
((p1-q1) ^2+ (p2-q2) ^2 + ... + (pn-qn) ^2) ^ (1/2).

Formula used is,

Dist ((x, y), (a, b)) = $\sqrt{(x - a)^2 + (y - b)^2}$

Using this approach, we are able to rank companies based on the weightage of the keywords in the resume. The data flow of the operations is as shown in Figure 2. It begins with the user login and the uploading process of the resume. The specific keywords from the resume are retrieved and unwanted data are removed rom the resume data.

Fig 2. Top-Level Class Diagram

The Top-Level Class Diagram is show in in Figure 2 above. It describes the various components and the functionality of each of them.

- Upload resume: The UI is created for the resume to be uploaded, the resume can be of type doc or docx.
- Parse details from resume: The resume is cleansed of unwanted data and only the required keywords are stored. The keywords are calculated based on their content weightage and their HashCode is created. This HashCode is stored into an index Array.
- Store processed data in DB: This index Array is stored into the MySQL database, which can be used later to fetch the information using its key.
- List of Company details from DB: The company details are inserted into the MySQL database and is used to search or fetch the company details.
- Apply algorithm on the HashCode: The HashCode is fetched from the server and checked for the keyword index. Based on the weightage of the keywords, the Jobs are ranked.
- Ranked Companies: The companies have been ranked based on the keyword's weightage of the user's resume.

IV. EXPERIMENTS AND RESULTS

The data flow of the system is shown in Figure 3 below. It starts with user login. The user uses his credentials to login to his/her account. The next step is to upload the resume, the resume can be of DOC, DOCX, PDF or TXT format.

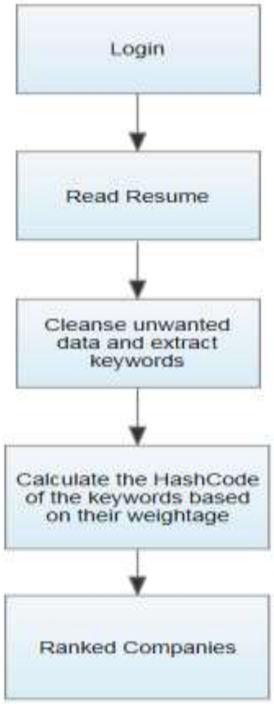

Fig 3. Data Flow

The keywords are obtained using regular expressions. Details such as email, phone number, College, Skills, work experience are retrieved using regular expression and NLP. These keywords are converted into HashCode and stored into an index Array. Java HashMap is used which is a hash table-based implementation of Java's Map interface. A Map is a collection of key-value pairs. Java HashMap allows null values and the null key. HashMap is an unordered collection used to store the keywords and their HashCode's.

Calculations are done in the server by fetching the HashCode of the user's resume and using it to compare it with the details of the company's requirements. The company details are stored into MySQL database. The companies job requirements are compared with the HashCode of the users resume and matched companies are displayed and ranked to show the best job for the user.

Based on this outcome, it puts a plan to utilize KNN Query processing for figuring out the best companies from resumes. Predictions of forecast continuously increments as the client information gather.

V. CONCLUSION

The web app is set up from various models, to foresee the best course of action of associations to the user. The user can view the best jobs according to his/her skills and experience and find their dream job. Anticipating rundown of organizations from the curriculum programs vitae of the client. Extracting useful data of organizations through procedure of web scraping from web crawlers. We can characterize the information dependent on the highlights separated.

VI. ACKNOWLEDGMENT

The work has been done in the university and home. I would like to thank Mr. Rajendra M for his guidance and support.

REFERENCES

[1] G. Demartini, D. E. Difallah, U. Gadiraju and M. Catasta, "An Introduction to hybrid Human-Machine information Systems", pp. 23-26, 2016.

[2] Ashish Dutt, Maizatul Akmar Ismail and Tutut Herawan, "A Systematic Review on Educational Data Mining," *August*, 2017.

[3] Winai Nadee, Korawit Prutsachainimmit, "Towards Data Extraction of Dynamic Content from JavaScript Web Applications" 10-12 Jan. 2018

[4] Dongcheng Peng, Tieshan Li, Yang Wang, C. L. Philip Chen, *Research on Information Collection Method of Shipping Job Hunting Based on Web Crawler, August 2018.*

[5] Zhaoli Wang, Xinhuai Tang, Debai Chen — *A Resume Recommendation Model for Online Recruitment*, March, 2016.