International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Classification of Resumes by Machine Learning Methods to Solve Complex Issues in Human Resourcing

Deepak Kumar C V
MCA Department
RNS Institute of Technology,Bengaluru
mca.deepakkumarcv@gmail.com

Mohammed Junaid
MCA Department
RNS Institute of Technology,Bengaluru
mca.mohammedjuaid@gmail.com

Meghana R Halliyavar
MCA Department
RNS Institute of Technology,Bengaluru
Mca.meghanarhalliyavar@gmail.com

Naman Kumar
MCA Department
RNS Institute of Technology,Bengaluru
mca.namankumar@gmail.com

Kushal P
MCA Department
RNS Institute of Technology,Bengaluru
mca.kushalp@gmail.com

Abstract*-- *Differently to the serious companies every company has a different point of view while reading through your resume, barely matching skills and experience is no more important alone. Domain expertise plays an important role for some companies or others may count the total number of skills, years as professional. HR agencies use multiple headhunting tools and online search technologies. In the database of millions and searches of data streams. By using this technology, these are automatically classified by a pre-determined machine-learning classification and the usual search methods of sorting resumes according to rank measures with high accuracy allow for better way working competence human resources.

***Index terms*-**Machine learning, clustering, human resourcing, hiring patters, K-means, support vector machines(SVM), association rule mining(ARM).

I. INTRODUCTION

The Human resource is facing new faced of problems as days role on in meeting the client to client requirement for which they have a unique solution every time. Not every job description can be satisfied with the same set of resume for each client. Because every organization has varied perception of a resume when reading an job description. Experience and skills which is being considered for more serious organizational matching. Some company sees number of skills and total years as majar role but some other compaines wants the domain expertise. Recruitment Agencies use various head hunting tools and online search methods. These search methods were tied to the

millions of resumes in a database. Yet here you'll find basic search engines that only analyze resumes based on applied keywords and recommend the 5 best resume matches. These Keywords are generally decided by the HR person, who will read and re -read the Job description. It means they download the needed resumes and do all that manual labour work of opening, reading the resumes. This traditional resume by best match with the job description is pretty a old way of doing, and its more time consuming as well need n number of discussion to identify if we should offer this resume copy, or just leave it. And this way best resume could be missed or ignored because of less time. In the advised means, candidate presents their resume. Data Mining algorithms are applied and the attributes: Years of experience, Current location, Education, programming skills, Domain etc. When querying the database, pass in how you want to hire it will get best matches for that style of hiring.This method will solve the complex ranking issues and will select resumes efficiently over the existing traditional method.

II. BIG DATA

Big data is the word that appears everywhere. You store day-to-data like structured and unstructured vaolume of data, then it becomes bigdata It is about more data. but that does not really matter That is all of what an organization does with itthe data that matters. One of the values lies in analyzing big data to derive actionable insights that enable more informed decision-making and strategic enterprise moves. Until yesterday, the data that reside in your companies servers was just “data” – organized and cataloged. One day, the definition of Big Data suddenly became popular, and nowadays it is no longer customary to call your data 'the business-relevant object' - now you are allowed to say that this is Big Data. It includes any and every single piece of data your company has stored until this very moment. That data includes the ones you store in clouds, and even so far as URLs that of which you bookmarked. Maybe all data have not been digitized at your company. You may still not have

already structured all of the data. Well, then everything digital and none structured with your company is now Big Data.In other words, the data — whether categorized or not in your servers is BIG DATA. The possibilities with these kinds of data are endless in terms of getting different results through types analysis. Not all data have to be used in every analysis On other hand the different analysis of BIG DATA is then used with part of big data to get the result predictions. So what is Big Data; it means the data that You are doing to process for some results and use with predictions etc. At once and wham your Company or Organization is doing the Big Data i.e., High kahuna Information technology in an effort to deduce different kinds of results From The Same ol' data you have been storing intentionally OR Unintentionally for ages.

III. DATA MINING

The quantity of information that's been saved in databases as well the complete number of database packages (in organization and technological know-how) which built over remaining two decades have similarly long gone up by wayover 10-folds. The growth in electronic data was driven by the dramatic success of the relational model for data organization and through advances in both storage technology as well as retrieval/manipulation technologies. Although technology for storage needed to catch up with this demand (and did, at an unparalleled rate), developing software that could analyze the data was not of great concern until companies started to recognize that a gold mine lay within these mounds. That data contains years of business discipline knowledge in various areas just waiting to be leveraged for smarter Business Decision Support. Current database management systems used to manage these data sets do not allow access anything beyond the explicitly represented information in databases, that is just their contents. The information behind this is like the tip of an iceberg and we are only seeing data stored in database part. A great deal of business insight, often hidden knowledge about aspects of their business that can be used in more effective decision support for the company is implicitly contained within this data This extraction of knowledge from data is known as Data Mining or Knowledge Discovery in Databases and it refers to the non-trivial activity of extracting implicit, previously unknown information that provides added value over time for customers. This is evident with the incredible benefits that come out of Data Mining there has indeed been much effort put in its development. Almost concurrently with the growth of database technology, research in machine learning was also maturing and intelligent solutions were developed using various sophisticated techniques based on human Learning by example, casedbased reasoning, learning by observation and neural networks are some of the most popular learning techniques that were being used to create the ultimate thinking machine.

IV. IMPLEMENTATION

The proposed approach starts with the preparation of the clusters of the fundamental search attributes required to prepare the hiring patters. Initially, the admin who is registered can view the file path, articles, stopwords, data cleaning of resume, future vector (token name, frequency, IDFT), clustering(K means cluster, SVM cluster),Remove old data. As switching to HR can register using Register Form, he can search the resume based on skills(Programming,testingetc),domain(Bigdata, Networking, Embedded system etc).HR can view group analysis graph separately for domain and skills. Coming as a candidate can Submit resume, view resume, delete resume.

A. ADMIN

The admin can view output of all the data mining techniques and classification output in the form of grid. The different data mining techniques and classifications are as listed below:-

1) Resume module

The resume module is responsible for storage of resumes. Resume name and resume description acts as an input.

2) Data cleaning

Data Cleaning algorithm: It is included for stop words removal from each of tweet resume which are not significant in meaning. A list of stop-words: DATA MINING,STOPWORDS:(the stopwords defined by the data mining forum) (These are eliminated before or after processing natural language data.) Not all use ere same stop words and there is not a definite list of top stop words. The stop words array that algorithm uses a : able about across after all...and so on Once the data clean is done, we can take this as a set, (Clean Id,Clean Data,Resume ID). Where Clean Id is the Tweet unique ID, Clean Data will be the clean data after cleanup on Ethical resumes and Article id as resume_unique_id.

3) Tokenization

Tokenization is a process of converting the clean data into a set of words known as tokens. Each of the token can be represented as Token Id, Token Name and Resume ID. This table has set of words after cleaning.

4) Frequency Computation

This is a process in which the frequency computation is performed. For each of the resumes the frequency is computed. Frequency is number of times a i^{th} token appears in resume j^{th} . The frequency matrix is computed in the following format.

Freq ID	Resume ID	Token Name	Frequency

Table IV.1:Frequency Table

5) Feature Vector Computation

This module is responsible for feature vector computation in which each of the token IDFT and FV is computed. The IDFT is computed using the equation 1

$$IDFT=log[N/f] \text{ ---(1)}$$

Where,

N = *number of pages in which tokenis present*

f = *frequency of word*

The Feature vector is computed using the following equation 2

$$FV = f*IDFT \text{ ---(2)}$$

Ranking of Resumes using TF-IDF

- Divide the search string into words
- For the list of unique resumes uploaded
- Find the feature vector for each words of search and do a summation for the sequence of words for a resume
- Repeat the process for all the resumes
- Rank the resumes based on sorted order of the values

6) K Means Resume Classification

List of category along with training data set for each of the category is taken whose sample is shown in the figure 2

CATID	CATNAME	CATKEYWORD
3	PROGRAMMING	PYTHON
4	PROGRAMMING	JAVASCRIPT
5	PROGRAMMING	ANGULARJS
6	PROGRAMMING	ANGULAR2
7	PROGRAMMING	R
8	PROGRAMMING	MATLAB
9	PROGRAMMING	C++
19	PROGRAMMING	ANGULAR
28	PROGRAMMING	JUNIT
10	TESTING	SELENIUM
11	TESTING	QTP
12	TESTING	TESTCASES
13	TESTING	WEBDRIVER
20	TESTING	MANUAL
21	TESTING	TESTING
22	TESTING	RC
23	TESTING	RC
27	TESTING	JUNIT
29	TESTING	DRIVER
30	TESTING	ESTIMATION
31	TESTING	REVIEW
14	MANAGEMENT	MANAGE
15	MANAGEMENT	JIRA
16	MANAGEMENT	SKILLSSOFT
17	MANAGEMENT	TEAMS
18	MANAGEMENT	LEAD

Figure 2: K means Resume classification

- For each of the category the word count is obtained for the resume
- The distance is compute as maxValueCategory – resumecategorycount
- The matrix computed is shown in figure 3

Name
RESUMENAME
USERID
CATNAME
DISTANCE
COUNT

Figure 3: Matrix computed

- Finally the minimum distance is taken as the category for the resume and for each resume we compute in figure 4

#	Name
1	RESUMENAME
2	USERID
3	CATNAME

Figure 4:Minimun resume computed

7) SVM Classification based on probability

- The classifier training vectors for the various domains are chosen.

CATID	CATNAME
1	EMBEDDEDSYSTEMS
2	EMBEDDEDSYSTEMS
3	EMBEDDEDSYSTEMS
4	EMBEDDEDSYSTEMS
5	EMBEDDEDSYSTEMS
6	EMBEDDEDSYSTEMS
7	BIGDATA
8	BIGDATA
9	BIGDATA
10	BIGDATA
11	BIGDATA
12	NETWORKING
13	NETWORKING
14	NETWORKING
15	NETWORKING
16	WIRELESS
17	WIRELESS
18	WIRELESS
19	WIRELESS
20	TELECOMMUNICATION
21	TELECOMMUNICATION
22	TELECOMMUNICATION
23	TELECOMMUNICATION
24	TELECOMMUNICATION
25	AUTOMATIVE
27	AUTOMATIVE

Figure 5:Resume classification based on probability

- The probability is computed for each of the category using the following

$$p(r|catname) = \frac{Number\ of\ words\ of\ category}{total\ words\ of\ resume} \quad ---(3)$$

The following matrix is computed as shown in figure 6

#	Name
1	RESUMENAME
2	USERID
3	CATNAME
4	PROBABILITY
5	COUNT

Figure 6:Resume classification based on matrix computed

- Once the probability is computed for each of the resume.
- The highest probability is found.
- The class label is assigned based on the respective category which is highest.

#	Name
1	RESUMENAME
2	USERID
3	CATNAME

Figure 7:Highest probability

B. CANDIDATE

The candidates has to register using register form, if it is for first time then login.If the candidate has registered initially then can login directly. As logged in to the page, will be able to upload the resume by giving name and submit in pdf format. During resume upload, sequence of data mining techniques like data cleaning, tokenization, frequency computation, feature vector computation and also classification of skills using k-means and domains using SVM is done. The candidate can even delete and upload new resume as needed.

C. HR(HUMAN RESOURCE)

The HR can register into the application and search based on association rule mining or based on the query. Once the search is performed the resumes are ranked based on the feature vector and domains and skills set related.

1) HIRING PATTERNS

Hiring patterns are the weighted as well as mandatory attributes which helps in ranking and re-ranking of resumes based on mentioned atrributes in mapping of k-means clusters. Parameters divided into Basic Search Elements for ex: Current Location, Years of Experience, Technical Domain and Skills & Education Qualification.

2) ASSOSIATION RULE MINING

Association rule mining: This is a procedure that involves the discovery of frequent patterns, correlations,associations or casual structures from data sets found in various types of databases such as relational database transactional database other forms of repositories. ARM: ARM is the methodology to merge multiple criterias from resume and rank them as best based on needs of multi attribute searches by performing intersection of various algorithms set. This time we will use ARM to uncover the association between educational qualification and years of experience for a candidate.

ADVANTAGES

- This way not only select skills and experience based resumes, but also takes care of hiring pattern of recruters.

• The factors include the numerical figure on current

• There is mandatory, which may not be done in case you have mentioned with older than (Year) of all other profile field on this filter like location, preffered work location and technical domain corresponding to the skills total years expereince educational qualification.

• Based on Hiring draft pattern best matching candidates are discovered during database search.

• It will resolve complex ranking difficulties and sort the resumes proficiently over previous traditional methods.

• The new technique would support reactive changes, for instance if the job description changes promptly.ecoreGen.

V. RESULTS

The user can register using SIGN UP-CANDIDATE portal,where in the user will be able to upload or delete the resumes.later the HR will login using SIGN-UP HR portal to search the eligible resumes based on his requirements and also download those resumes.LOGIN portal is the common portal for the HR and candidate to login once after the fresh registration.

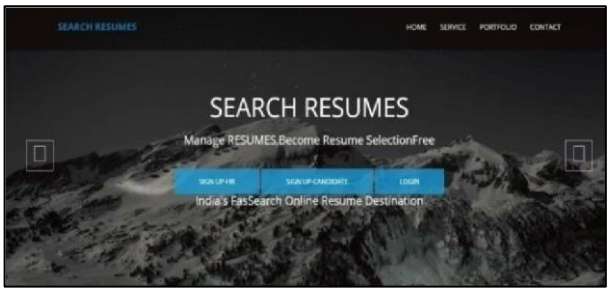

Figure 8:Resume search Machine for sorting

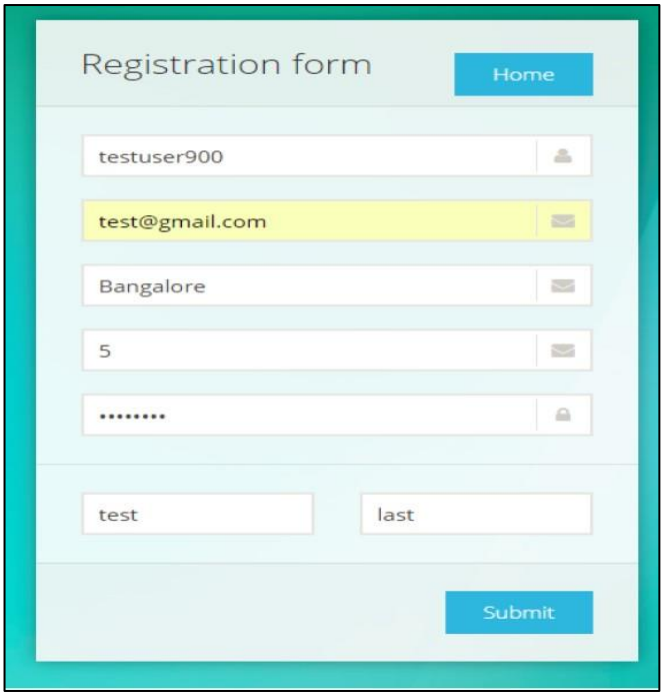

Figure 9:Resume search Technique

VI. CONCLUSION AND FUTURE WORK

This approach 3 different actors are used i.e., Candidate, HR and Admin. Able to upload the resume by candidate. During resume upload sequence of data cleaning, tokenization, frequency computation and also the feature vector calculation as well classification of skills using k-means and domains SVM is done through those string matching techniques. Even delete and uploads a new resume by the candidate All data mining techniques can be viewed in grid form which is formatted as the output of classification. All the HR need to do is just register into the application and he/she can search based on association rule mining or query ways. The actual feature vector is used to rank the resumes and domains as reported by Full Text Search are Novel's, Candidate features + Domains+ Skills set. The projectYou may add more domains to the request into additional dissciplies and types as a method for future development. If needed, you could extend the project so that you can analyze also sentiments.

VII. REFERENCES

[1] Junjie Wu, Advances in K-means Clustering. Springer-Verlag Berlin Heidelberg (2012)

[2]Leskovec J, Anand Raraman and Jeffrey D.Ullman Instructor, Mining of Massive Datesets Infolab Stanford 2014[2]

[3] Michael Steinbach, Vipin Kumarand Pang-Ning Tan, Introduction to Data Mining.Pearson Publications 2006.

[4] Yanchang Zhao, R and Data Mining: Examples and Case Studies, 2013