Machine Learning Methods for Solving Complex Ranking and Sorting Issues in Human Resourcing

Arun Kumar
Big Data and Analytics Dept - ERS
HCL Technologies
Noida, India
Arunav.kumar@gmail.com

Anurag Pandey
M.Tech Software Engg
BITS Pilani
Rajasthan, India
anurag.pandey382@gmail.com

Suman Kaushik
Human Resource Manager-HR
TerminalApps Software Private Ltd
Ghaziabad, India
Suman.k@terminalapps.com

***Abstract*—Every organization doesn't necessary to have the common point of view of a particular resume while considering for a job description (JD). Keeping the same role in place, while some stress on technical skills, the other give importance to professional experience and domain expertise. Understanding these hiring patterns are becoming important in today's head hunting. The traditional job search engines offers resumes which matches to the input keywords. As the search outcomes from these search engines grows, the problem in selecting the best profile surges. The role of Human Resource (HR) staff becomes more important in understanding these hiring patterns and suggesting the suitable profiles. HR staff proposes these profiles which are ranked manually. The proposed method is to understand the intelligence behind the hiring pattern and apply the machine learning to accommodate the identified intelligence. The proposed method offers the ranking system according to the hiring patterns. Highly trained models along with the traditional search method, predicts the ranking and sorting of resumes with high accuracy and simplifies the job of human resourcing efficiently.**

***Index Terms*—Machine learning, supervised clusters, unsupervised clusters, human resourcing, hiring patterns, Cosine Similarity, latent semantic analysis (LSA), K-means, support vector machines (SVM), association rule mining (ARM)**

I. Introduction

As the new technologies are evolving day by day, the human resourcing is facing peculiar challenges in meeting the requirements from client to client. The same set of resumes for a same JD doesn't work for all the clients. As every organization carries a different point about a resume while reading through the resume. Marely matching skills and experience is no more important alone for the serious organizations. For example, some companies consider the Domain expertise but some other gives more importance to the number of skills and total years of professional experience.

Human Resource (HR) agencies use various head hunting tools and online search methods. These search methods connected with the database of millions of resumes. These are the simple search engines who parses the resumes against the given keywords and offers the best match results. The list of the searching keywords is usually prepared by the HR after reading the job description several time. The HR downloads these searched resumes and does the manual work by opening and reading the resumes. By this ways, HR person tries to find the resumes which are best match to the JD. This is a cumbersome process and requires reasonable time and multiple discussion with the candidate before offering the resume to the client. Usually, due to the complexity of the database, many efficient resumes missed out from the search results or not considered due to stringent timelines of closure.

The proposed method selects the resumes not only based on the skills and experience but also considers the hiring pattern of the recruiters. The hiring patterns are prepared by understanding the implicit requirements from the client and preparing the Hiring Matrix based on the important factors. The factors consist the numerical weightages about the *Current Location, Preferred work Location, Technical Domain corresponding to the Skills, total years of experience, and Education Qualification*. While searching the database, the hiring pattern is used to find the best matches. The hiring patterns are prepared with the help of machine learning algorithms with the flavor of data analytics. The proposed method will solve the complex ranking issue and will sort the resumes efficiently over the existing traditional methods.

In the result section, a rationale over the results obtained from the proposed method is outlined. The proposed method will demonstrate the effectiveness of supervised and unsupervised learning methods and will also implement the dynamic rules to accommodate the hiring Furthermore, to fine-tune the outcomes, the ANN is proposed, which will the future work to make the system more effective with self-learning and to self-overcome the errors and deviations as the database grows in future.

II. Ease of Use

The proposed method will automate the manual work involved in ranking and reselecting the candidates against a JD. Also, if the requirements in the JD changes on the fly, the proposed method will be able to accommodate the changes dynamically. The proposed method is designed using the Text Mining and analytics packages available in open source tool R-Analytics. The proposed method will try to remove the dependency and prolonged experience required in the field of Human Resourcing and will enable the opportunities for the

978-1-5090-1560-3/17 $31.00 © 2017 IEEE
DOI 10.1109/IACC.2017.16

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on July 12,2025 at 04:10:39 UTC from IEEE Xplore. Restrictions apply.

new people to work in the field of HR without special knowledge and experience needed.

III. Implementation

The proposed method starts with the preparation of the clusters of the fundamental search attributes required to prepare the Hiring Patterns.

A. Preparation of the raw corpus

Raw corpus is the plain text form of resumes the indexed data of resume, which leaves behind the complexities involved in reading the data from different sources. It normalizes the input data and removes the impact of data formatting, fonts, colors and styles. The Text Mining (TM) package of R-tool is used to make the raw corpus from the resumes.

B. Preparation of the refined corpus

The raw corpus contains lots of common words like punctuations, special characters, verbs, auxiliary verbs, junk words, non-dictionary words like candidate names and email ids and misspelled words. Having these words in searching mechanism will reduce the search performance, hence, it is required to clean up the raw corpus with this noise. So a filtering is applied to remove these junk and stop words and the remaining data will be free from this noise. Thus obtained reduced corpus is named as refined corpus.

C. TF-IDF matrix: Inverted indexed searching

The reduced corpus is further input to the TF-IDF. TF-IDF is called as Term Frequency and Inverse Document Frequency matrix. TF-IDF is used to extract the frequency of a particular keyword in the whole reduced corpus. TF-IDF matrix helps out in sorting out the relevant resumes based on the inverted index search. This TF-IDF matrix is the input to further modules, K-Means, SVM (Support Vector Machines) and ARM (Association Rule Mining). The TM and NLP R-packages are used to create the TF-IDF matrix.

D. K-Means: Locations and Technical Skills clusters

K-Means [1] is preferred to prepare the clusters on TF-IDF matrix data as it offered very favorable results for analytics. It is observed through results that, when the dataset is larger, the purity of the clusters created by K-Means is very high. K-Means produces the output where all the grouped keywords are semantically similar. We could draw the implicit requirements of the job description using the K-Means by taking the advantage of these unsupervised clusters. Tab-IV shows the K-Means clusters formed from the input of 4000 resumes. Showing from K1 to K5 only.

E. Support Vector Machines (SVM): Technical Domains mapping of the Skills clusters

The clusters created from K-Means algorithm are unsupervised clusters and they do not have the appropriate labels. They are categorized under the unknown labels from K=1 to K=n, the value n is found from the Elbow rule, which gives the optimal value of K, where the K-Means converges. It was observed that for the set of 10,000 input resumes collected against the 100 different job descriptions the value of K is 31. To label these K1 to K31, the human knowledge is used. With the labeled clusters, the SVM [2] model is trained. To train the SVM model efficiently, out of 10,000 resumes, 4000 resumes are selected randomly, and confusion matrix is calculated over the test and prediction calibration. It is observed, as more test resumes participate in training of SVM, higher accuracies of prediction are achieved as seen in the results of *confusion matrix*. Once the prediction accuracy is achieved up to 99% and above, the SVM model is labeled as trained model, and can be used to predict the *Technical Domains* for the unknown resumes.

F. Regular Expressions to find the Experience in years

Regular expressions are required to fetch one of the important attribute, *Experience*. The total *Years of Experience* is mentioned in different styles as shown in Tab-I

TABLE I. Regular expression inputs to get total years of experience

S. No	Patterns for *Years of Experience*
1	Have 10+ years of professional experience
2	Possess 5 years of experience
3	Have five years of experience
4	Ten years of strong experience
5	Total years of experience-8 years

Sometimes, the total years of experience requires to be calculated in segments, as some resumes doesn't mention it explicitly, as shown in Tab-II. Some simple regular expression is written in Perl, to find the total number of experience and thus created the separate cluster of years of experience.

TABLE II. Experience mentioned in segments in a single resume

S. No	*Years of Experience* mentioned in segments, patterns
1	Worked for PQR industry – from Year 2010 to 2012
	Worked in XYZ Organization – from 2012 to 2014
	Presently associated with 123 firm – 2014 till today/present

G. Association Rule Mining (ARM) to find the Experience in years

The *Education Qualification* is mined through the ARM rules written in R-Scripts. We found thee sufficient threshold for minimum support and minimum confidence for the n-grams (n=8). The Education Qualification is associated with the stream of an Education. Some of the interesting patterns associated with the *Education Qualification* are captured in the "Table-III". The ARM are designed and calibrated for minimum support and minimum confidence thresholds. The successful list of Education Qualifications is listed in Tab-IV

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on July 12,2025 at 04:10:39 UTC from IEEE Xplore. Restrictions apply.

TABLE III. LHS AND RHS FOR ARM RULES FOR EDUCATION QUALIFICATIONS

Education Qualification	
LHS	***RHS***
B.Tech	In Electronics and Communication Engineering from XYZ university
B.E.	In Information Technology from ABC institute of
MBA	From 123 college in International Business
M.S.	In computer Science in 2010 from KLM University

TABLE IV. LIST OF SUCCESSFUL EDUCATION QUALIFICATION USING ARM RULES

Education Qualification		
Graduate	***Masters***	***Doctorate***
B.E.	M.E.	Ph.D.
B.Tech	M.Tech	PGDBM
B.S.	M.S.	PGD
B.Sc.	M.Sc.	PGDCA
B.C.A.	M.C.A.	PGDM
B.B.A.	M.B.A.	PGDIT
B.Com	M.Com	PGDMIT
B.S.C.S	M.S.C.S	PGDOM
B.S.C.E	M.S.C.E.	
B.S.E.E.	M.S.E.E.	
	M.PHIL	

TABLE V. K=31 CLUSTERS FROM K-MEANS ON 1000 PROFILES (SHOWING K1 TO K5 ONLY)

K-Means Clusters				
K1	***K2***	***K3***	***K4***	***K5***
China	C++	Hadoop	networking	lte
Australia	C	Hbase	cisco	wireless
Cyprus	Embedded	Hive	tenderburg	wcdma
India	arm	pig	wireless	gsm
Bangalore	processor	bigdata	tcpip	gprs
Karnataka	freescale	analytics	server	egprs
Hyderabad	linux	clusters	client	edge
Pune	drivers	kmeans	iptable	wimax
Kolkata	xilinx	hana	administrator	wifi
Mumbai	matlab	watson	routers	3gpp
Delhi	labview	nosql	ipv6	2g
Noida	electronics	mongodb	ipaddress	mac
Gurgaon	bsp	cloud	protocol	phy
Faridabad	poky	hdfs	virus	schedulers
Kocchi	toolchain	lucene	conferencing	encoder
Trivandrum	gcc	solr	clearcase	mimo
Canada	ubuntu	iis	ccna	prach

TABLE VI. MAPPING OF K-MEANS CLUSTERS TO DOMAINS - MANUALLY

Domains	K-Means clusters of Skills
Embedded Systems	C++ C Embedded arm processor freescale linux drivers xilinx matlab labview electronics programming poky toolchain gcc microcontroller dsp processing platform
Big Data Analytics	Hadoop Hbase Hive pig bigdata analytics clusters kmeans hana watson nosql mongodb cloud hdfs lucene solr iis kibana database svm
Networking	networking cisco tenderburg wireless tcpip server client iptable administrator routers ipv6 ipaddress protocol virus conferencing clearcase ccna ccnp security remote
Wireless	lte wireless wcdma gsm gprs egprs edge wimax wifi 3gpp 2g mac phy schedulers encoder mimo prach vswr signalling aeroflex
Telecommunication	pdcch pdcp pdh pdp rlc rnc rrc rtcp rtp sip son wireshark lte umts hspa mimo wimax femtocell bsc gprs gsm cdma wcdma
Embedded Systems	vxworks scilab tornado assembly labview usb bluetooth pi chipsets pcie porting analog microcontroller microprocessor embedded
Automotive	automotive diagnostics autosar exhaust calibration braking emission ercos powertrain gasoline injection canalyzer valve combustion brake simulink torque cylinder ecus
IC Design	ovm uvm vmm verilog systemverilog vhdl asic fpga soc testbench floorplanning modelsim questasim amba axi cadence synopsys xilinx vlsi
Web Developer	dreamweaver drupal joomla jquery netbeans photoshop ruby sharepoint html flash eclipse wordpress css php javascript ajax angular angularjs animation

IV. HIRING PATTERNS

Hiring patterns are the weighted attributes and mandatory, which plays an important role in ranking and re-ranking of the resumes based on the attributes in Tab-VI. The parameters are categorized as the basic search elements e.g. *Current Location, Years of Experience, Technical Domain, Technical Skills and Education Qualification.* We have created the individual clusters of such attributes. We have solved this problem by taking the approach of the Venn Diagram as shown in Fig-1. Say, we have the following criteria from the company XYZ, which from the given JD. Which is understood and placed in the below matrix shown in Tab-VII. The possible best indexed outcomes can be seen in Tab-VIII.

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on July 12,2025 at 04:10:39 UTC from IEEE Xplore. Restrictions apply.

TABLE VII. HIRING PATTERN OF COMPANY XYZ

Hiring pattern of company XYZ		
Attribute	***Weightage***	***Label***
Education	40%	A
Experience in years	30%	B
Technical Domain	20%	C
Current Location	10%	D

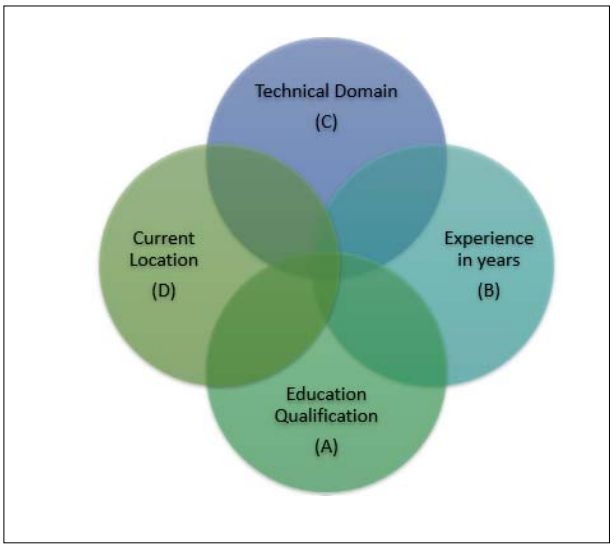

Fig. 1. Venn diagram of the Hiring Pattern Tab-VI

TABLE VIII. INDEXING OF THE SEARCH OUTCOMES

Hiring pattern of company XYZ	
Indexing Order (Not Rank)	***Venn Segment Result***
1	A∩B∩C∩D
2	A∩B∩C
3	A∩B∩D
4	A∩B
5	A∩C∩D
6	A∩C
7	A∩D
8	A
9	B∩C∩D
10	B∩C
11	B∩D
12	B
13	C∩D
14	C
15	D

V. RANKING OF INDEXED DATA

The individual Venn Segment obtained in "Tab. VII" will undergoes the cosine distance. To measure the cosine distance between the indexed document and the input query, the R-Package Latent Semantic Analysis (LSA) package is used. Minimum the cosine distance, maximum is the matching index and thus higher the rank is assigned. By this way, combining all the segments together gives the composite ranked list of the final resumes. Once the weighted attributes are assigned a new weightage, the Venn Segment results are shuffled, and hence the final ranking of the resumes shuffles.

To validate the obtained results, please observe the Experiment section.

VI. EXPERIMENTS AND RESULTS

The results for the machine learning models for ARM, K-Means and SVM are collected in the subsequent tables from Table IX to Table XVIII. We have written the scripts and programs in R-Analytics tool [4] to create the data, which support the results mentioned in the below tables from Tab-IX to Tab-XVIII.

TABLE IX. RESULTS ACHIEVED FOR ARM

ARM Results	
Learning Data (90%)	***Testing Data (10%)***
900 profiles	100 profiles

TABLE X. RESULTS ACHIEVED BY ARM

ARM Metrics					
				Accuracy	***Precision***
TP	***FP***	***FN***	***TN***	***(TP+TN)/(P+N)***	***TP/(TP+FP)***
2	0	0	0	100	100

TABLE XI. RESULTS ACHIEVED FOR K-MEANS FOR LOCATION EXTRACTION

Hiring pattern of company XYZ		
S. N0	***Learning Data***	***Testing Data***
1	900 profiles (90%)	100 profiles (10%)
2	800 profiles (80%)	200 profiles (20%)
3	700 profiles (70%)	300 profiles (30%)

TABLE XII. K-MEANS RESULT FOR 90-10 METRICS

K-Means Metrics					
				Accuracy	***Precision***
TP	***FP***	***FN***	***TN***	***(TP+TN)/(P+N)***	***TP/(TP+FP)***
30	0	0	0	100	100

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on July 12,2025 at 04:10:39 UTC from IEEE Xplore. Restrictions apply.

TABLE XIII. K-MEANS RESULT FOR 80-20 METRICS

K-Means Metrics					
				Accuracy	*Precision*
TP	***FP***	***FN***	***TN***	***(TP+TN)/(P+N)***	***TP/(TP+FP)***
30	0	0	0	100	100

TABLE XIV. K-MEANS RESULT FOR 70-30 METRICS

ARM Metrics					
				Accuracy	*Precision*
TP	***FP***	***FN***	***TN***	***(TP+TN)/(P+N)***	***TP/(TP+FP)***
0	22	0	0	26.6666	26.666

TABLE XV. RESULTS ACHIEVED FOR SVM FOR DOMAIN EXTRACTION

Hiring pattern of company XYZ		
S. N0	***Learning Data***	***Testing Data***
1	900 profiles (90%)	100 profiles (10%)
2	800 profiles (80%)	200 profiles (20%)
3	700 profiles (70%)	300 profiles (30%)

TABLE XVI. SVM RESULT FOR 90-10 METRICS

K-Means Metrics					
				Accuracy	*Precision*
TP	***FP***	***FN***	***TN***	***(TP+TN)/(P+N)***	***TP/(TP+FP)***
28	12	0	0	88	88

TABLE XVII. SVM RESULT FOR 80-20 METRICS

K-Means Metrics					
				Accuracy	*Precision*
TP	***FP***	***FN***	***TN***	***(TP+TN)/(P+N)***	***TP/(TP+FP)***
137	63	0	0	68.5	68.5

TABLE XVIII. SVM RESULT FOR 70-30 METRICS

ARM Metrics					
				Accuracy	*Precision*
TP	***FP***	***FN***	***TN***	***(TP+TN)/(P+N)***	***TP/(TP+FP)***
171	129	0	0	57	57

VII. CONCLUSION

A benchmarking is done to observe the results obtained by the manual process and the resumes recommended by the proposed method through Machine Learning. The results obtained by the proposed method are mathematically and practically much better than traditional methods. Ranking and Re-ranking based on the Hiring Pattern are very useful for next generation Head Hunting solution. To overcome the residual error, the ANN and BPN is preferred. In future, ANN and BPN neural layers will be accommodated in the present design to make the system robust and self-error corrective.

REFERENCES

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

[2] Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, Mining of Massive Datasets, Stanford Infolab, 2014.

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

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

Authorized licensed use limited to: VIT University- Chennai Campus. Downloaded on July 12,2025 at 04:10:39 UTC from IEEE Xplore. Restrictions apply.