IJCTA, 9(27), 2016, pp. 121-126
© International Science Press

ISSN: 0974-5572

Assessment of Clustering Algorithm Using F-measure with Ranking Index

V. Jayaraj *, P. Rajadurai and J. Jegathesh Amalraj*****

ABSTRACT

The advent of internet has facilitated colossal number of resumes received from on-line through emails and through references. We find that so many firms find it difficult and time consuming process to pick up the appropriate resume. Plethora of research work to extract appropriate resume from the large volume of resume data are being carried out and to improve the performance of resume selection process. The clustering algorithm uses the stringent technique to group the resumes. The performance metric to check the cluster quality and to evaluate the accuracy of clustering algorithm used here is F-Measure. The clusters are obtained through Reduction Factor (RF). The reduction factors helps to minimize the resumes by their weight age. The weight ages are given to them as per their specific skills. The weighted sum of the F-measure are calculated by the higher values obtained by the clusters are best cluster and that indicates the higher accuracy. The secondary evaluation measure is Ranking Index (RI) it measure the percentage of decision that are correct. This RI value shows that the 40% - 90% of reduction in the number of features that the recruiters need to browse through select appropriate resumes.

Keywords: Precision, Recall, F-measure, Reduction factor, True Positive, True Negative, False Positive, False Negative.

1. INTRODUCTION

The process of Text Mining is also named as Information Extraction or Information Retrieval. To obtain appropriate knowledge from huge volume of data it is important to extract information by snubbing unwanted data. This helps to achieve correct decision by the admin and facilitates to improve their business to greater extent. Most of the business documents are managed and maintained in the form of documents and records. Hence the documents are in unstructured format. To extract information from these unstructured records it implements more manual effort and process. Hence to eliminate this snag, Text Mining provides as essential technique that is called as “Text Categorization”. In the past years many number of DM techniques have been proposed to perform different tasks on knowledge discovery. In this paper, we focus on the development of a knowledge discovery process that effectively used and update the patterns that is discovered. It is a challenging issue to pick exact features in the text documents and help the users to pick their need. At the initial stage, IR provides many term-based methods to solve the issues, such as Rocchio and probabilistic models, rough set models, BM25 and support vector machine based filtering models[1].

A wide variety of techniques for document analysis that not directly working with document directly, because of their intrinsic difficulty, but with decreased set of feature representation. This transformation is called Feature Extraction. If the features are identified carefully, they are suspected to pick the relevant information from the documents. One important issue is the number of features used. Feature Extraction refers to developing features that comes around there issues while explaining the tough data with required accuracy. We are interested in picking

* Associate Professor, Department of Computer Science and Engineering, Bharathidasan University, Tiruchirappalli–620023, Tamil Nadu, India.

** Research Scholar, Department of Computer Science and Engineering, Bharathidasan University, Tiruchirappalli–620023, Tamil Nadu, India.

*** Assistant Professor, Department of Computer Science, Thiruvalluvar University, Tittagudi – 606106, Tamil Nadu, India, *Email: rajadurai.panju@gmail.com*

up the resumes that are exact matches to a job description, where appropriated means that are a perspective employer would be interested in reading the retrieved resumes. We carry out some series of experiments on datasets contains million resumes, almost a quarter million job description and a number of appropriate judgements that shows which resumes are potentially qualified for a particular job description. Employees or employers usually submit their resumes through online job portal or online forms that contain many free text fields like job title, biography etc. There data's are maintained by relational database engine. The primary approach for this problem is clustering, based on this idea the unstructured documents are retried and plausible values for given field could be inferred from the context available in the record. Usually, resumes contain document level hierarchical structure. Our primary problem is to decrease the huge volume of resumes to few 100's numbers of related resumes to potentially speed up the recruitment process based on qualified Reduction Factor (RF). The extracted resumes are matched to the requirement criteria for a company. The primary goal of the research to develop the enhanced technique those help in picking up the apt resumes by processing the resume data set.

2. RELATED WORK

Many industries are today focussed by technology Jonathan (2008). According to statistics, data's available on internet is about 60% of what we need. The choice of representation depended on information on the internet. The meaningful units of text and natural language rule for combination of their units. In the phrases text representation for web document an argument was suggested in data mining technique it have been used for extracting terms as descriptive phrases from document groups. However, the effects of text mining using phrases show number of significant improvement. Self-organizing feature map proposed by Kohonen, a nervous network outer input receiving to divide the nervous network into different regions and these regions have different feature to the input model[4]. The connecting weights of various neurons have particular distribution, the neighbour neurons excite each, while the farthest neurons in inhibit each and the farthest neurons have weak effect. A kit named as “Learning Pinocchio (LP)” was implemented on resumes to study information extraction rules. The data cached in their work includes a topic structure of name, street, city, E-mail, Mobile, Fax and Zip code. LP is an adaptive and enhanced system for IE, Sumit Maheshwari [5].

3. REDUCTION FACTOR

Resumes will differ from people to people, it need rich work to extract the apt resume for their need. Important thing in resume extraction is qualification from the configuration file, the section named as academic qualification or professional qualification will aid the qualification extraction from the resume dataset. The extracted qualification is processed to find the weight age and the qualification weight age is stored for faster clustering process. The performance metric to check the cluster quality is reduction factor (RF) that measures the reduction in number of resumes under consideration. Let ‘X’ be the total number of resume present in dataset and the matching criteria using filtering technique is denoted as ‘Y’.

The reduction factors are evaluated for 500 number of sample resumes are shown in table.

Reduction factor for job requirement,

$$RF(\,jobrequirement) = 1 - y/x \tag{1}$$

Table 1
Extracting skills using RF

Feature Type	*Resumes format*	*| X |*	*| Y |*	*Reduction factor*
Dot net Programmer	.docx	344	109	0.68
Oracle developer	.docx	233	66	0.71
Java Programmer	.pdf	319	111	0.65
Siebel Programmer	.rtf	187	16	0.91

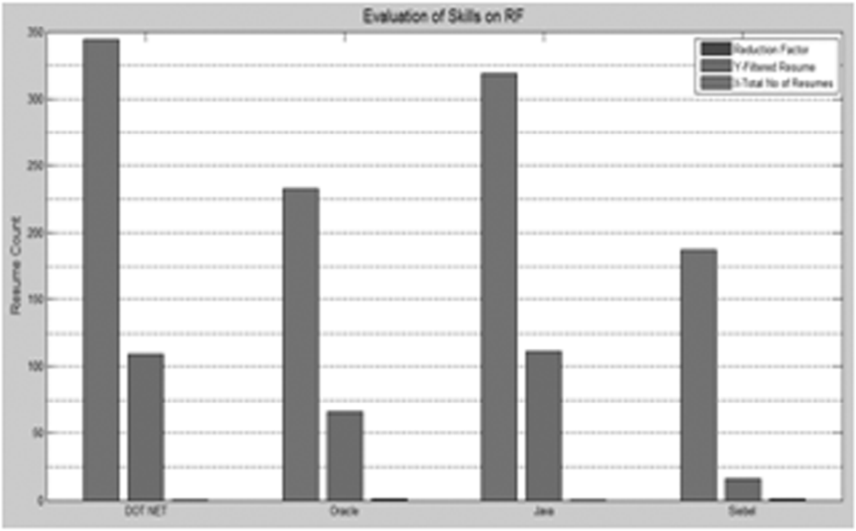

Figure1: Evaluation of skills on RF

4. PERFORMANCE EVALUATION USING F-MEASURE:

To achieve a identical representation in this paper, we call a concept called F-measure. The first step is to find all the resumes in a set of the documents. We use F-measure to evaluate the accuracy of the clustering algorithm. The F-measure evaluates the accuracy of each cluster and shows their quality. It is the combination of Precision and Recall values used in IR. We consider each cluster as the result of a query consideration and every class was the relevant documents for every query. The qualities of the objects extracted from the resumes are measured by two parameters called Precision and Recall. There are useful in evaluating the usefulness and truthfulness of the extracted texts from resumes. Two main criteria for evaluating the proficiency of a system is Precision and Recall which are used for identifying the equality between the objects. The Recall, Precision and F-measure for natural class K_i and cluster C_j are calculated as Formulas[1].

The recall, precision, and F-Measure for natural class K_i and cluster C_j are calculated as follows:

$$Recall = \frac{nij}{|ki|} \tag{2}$$

$$\Pr ecision = \frac{n_{ij}}{|c_j|} \tag{3}$$

Table 2
Calculation of Recall and Precision values

Feature type	*Recall*	*Precision*	*$F(K_i,C_j)$*
C1	0.3	86	0.59
C2	0.28	58.25	0.55
C3	0.34	79.75	0.67
C4	0.08	46.75	0.15

Where, n_{ij} is the number of members of class K_i in cluster C_j, The corresponding F-Measure F (K_i, C_j) is defined as :

$$F(K_i, C_j) = \frac{2 \times \operatorname{Re} call \times \Pr ecision}{\operatorname{Re} call + \Pr ecision} \tag{4}$$

5. EXPERIMENTALANALYSIS

The interesting aim of the experiments is to show how the proposed approach can help improving the effectiveness of identifying the accuracy of clusters using F-measure. Here to give a comprehensive investigation for the proposed work, our experimental analysis involves comparing the performance of different clustering algorithms. In experiments we first select a 500 sample resumes that had annotated matching. We split it into 4 cluster based on their feature type (qualification). The feature types are identified by the query through RF. This process considers only the job specification or category like ‘java programmer’ ‘dot net programmer’. The reduction factors are found used to filtering the resumes, based on weight ages and corresponding to the parameters is taken.

6. F-MEASURE RANKING INDEX

The second evaluation we do is to calculate the Ranking Index (RI) measures the percentage of decision that are correct. We used the F-measure to evaluate the accuracy of the clustering algorithms. The F-measure is a combination of *precision* and *recall* values are applied in information retrieval. Every cluster obtained must be considered as the identified by query, whereas each pre-classified documents can be considered as a desired set of documents for that findings. We consider every cluster as the result of a query and each class as if it was the relevant set of documents for that query.

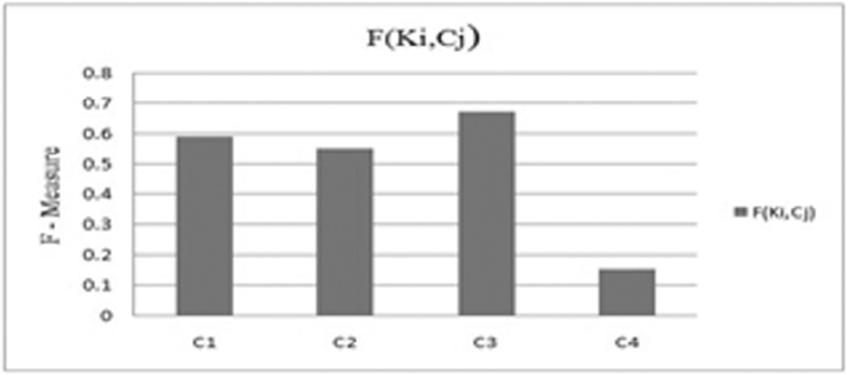

Figure 2: Evaluation of Recall and Precision with $F(k_i,C_j)$

Table 3
Assessment of TP, FP, TN, FN on clusters

Feature Type	*TP*	*FP*	*TN*	*FN*
C1	109	235	344	156
C2	66	167	233	267
C3	16	171	187	413
C4	111	138	319	181

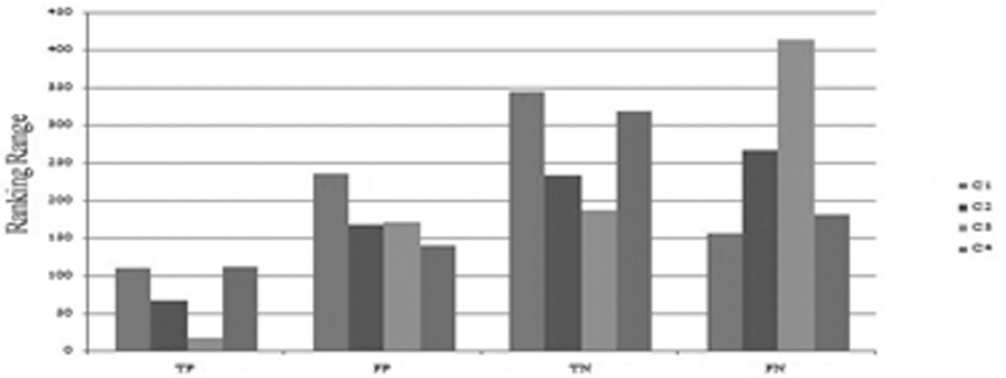

Figure 3: Evaluation TP, FP, TN, FN on clusters

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Table 4
Assessment of Ranking Index on Clusters

Clusters Evaluation	*Ranking Index (RI)*
Cluster 1	0.54
Cluster 2	0.41
Cluster 3	0.57
Cluster 4	0.26

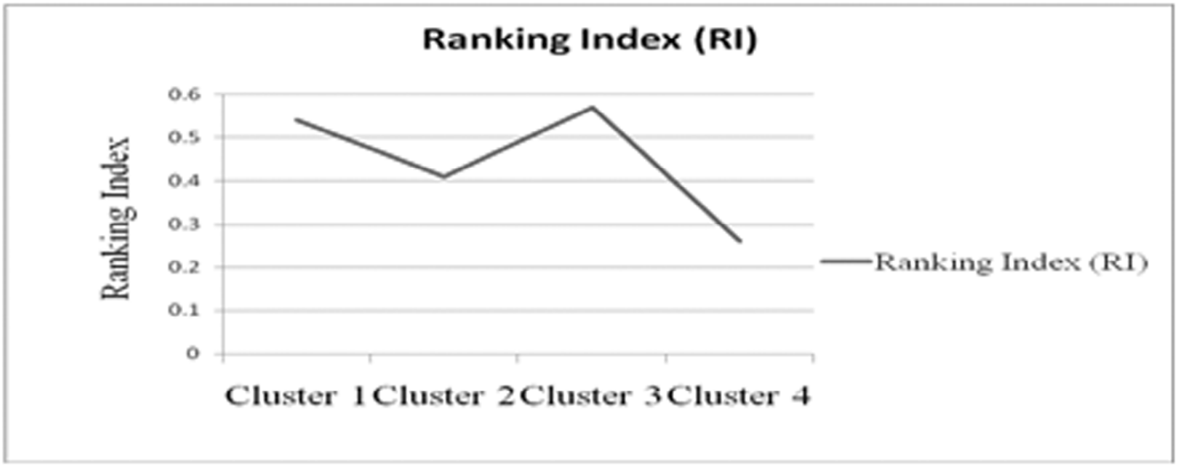

Figure 4: Evaluation of Ranking Index on clusters

CONCLUSION

The process of identifying apt resumes from large collection of resumes is the major problem often focused by the so many recruiter companies. We have extended the notion of special features to extract the special skills and experience from given set of resumes. The F-measures experimental result shows that 40% - 90% of reduction in the number of features that the recruiter needs to surf through to select appropriate resumes. The experimental result of Ranking Index shows that the decisions are correct. As each resume contains different sections with each section containing different types of text, an integrated approach has to be developed by considering information in each section. In future, we have to develop an approach to overcome this problem of extracting the important features present in the resume and cluster the similar resumes with accuracy

REFERENCES

[1] V. Jayaraj, P.Rajadurai, V.Mahalakshmi. "An empirical study of feature extraction Approaches *International Journal of Advance Research in Science and Engineering (IJARSE)*, **4(11)**, 140- 145, 2015.

[2] Afolabi I.T., Musa G.A., Ayo C.K. and Sofoluwe A. B (2008). "Knowledge discovery in online repositories: a text mining approach". *European Journal of Scientific Research*, **22 (2),** 241-25, 2008.

[3] Feldman R., Regev Y and Gorodetsky M. "A modular information extraction system. Intelligent Data Analysis", **12(1)**, 51-71, 2008.

[4] Feldman Ronen, and Ido Dagan. "Knowledge Discovery in Textual Databases (KDT)", KDD-95, 112-117, 1995.

[5] Sumit Maheshwari., Sainani A and Reddy P.K. "An approach to extract special skills to improve the performance of resume selection". *6th International Workshop on Databases in Networked Information Systems (DNIS2010)*, 256-273, March 2010.

[6] Abhishek Sainani. "Extracting Special Information to Improve the efficiency of Resume Selection Process", *Springer Berlin Heidelberg*, **5999,** 256-273, June 2011.

[7] Hawkins B.L., Rudy J. A and Wallace W. H. Recruiting, Retaining, and Reskilling Campus IT Professionals. "Technology Everywhere: A Campus Agenda for Educating and Managing Workers in the Digital Age". Dolan, A. F., *Jossey-Bass*: 75-91, 2004.

[8] Kumari and Sneha. "Automated Resume Extraction and Candidate Selection System". *IJRET: International Journal of Research in Engineering and Technology* , **3(1)** , 206-208, January 2014.

[9] Norm Schneider. Job Hunters: Resume Filters May Help or Hinder Your Job Search, Employment http://bizcovering.com/employment/job-hunters-résumé- filters-may-help-or- hinder-your-job-search/#ixzz1fajVNEJ2, Published on August 23, 2011

[10] Un Yong Nahm and Raymond J.Mooney , "Text Mining with Information Extraction", *AAAI Technical Report SS-02-06*. 60-68, 2002.

[11] Li Gao, Elizabeth Chang, and Song Han, "Powerful Tool to Expand Business Intelligence: Text Mining", *World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering*, **1(8)** , 2007.

[12] Divya Nasa, "Text Mining Techniques – A Survey", *International Journal of Advanced Research in Computer Science and Software Engineering*, **2(4)**, April 2012.

[13] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", *IEEE Transactions on Knowledge and Data Engineering*, **24(1)**, January 2012.