CLASSIFICATION OF TEXT DOCUMENTS USING JACCARD AND EUCLIDEAN DISTANCE SIMILARITY MEASURES BASED ON WORD SCORES GENERATED BY RAKE ALGORITHM

1 L.Shirisha,2 Dr.T.V.Rajinikanth,3 T.Madhuvani
M.Tech 2nd Year, Professor and Dean of R&D,B.Tech 3rd Year 2nd Sem
Department of Computer Science and Engineering
Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, Telangana, India -501301.

ABSTRACT

Day by day the availabilityof text documents are growing in an exponential manner. This growth of text documents throwing challenges to the user and research community in terms of classification. In our paper an attempt was made in finding the keywords in automation way using Rake algorithm on multiple text documents. The Rake algorithm finds more number of Candidate Keywords along with their word scores. In this only stop words was considered. The keywords were sub divided in to High, Middle and low level frequency key words based on word scores and then the similarity measures Jaccard and Euclidean Distance were applied to classify the documents. Comparisons were made among the three approaches High, Middle and low frequency keywords and concluded the results.

KEYWORDS: TEXT CLASSIFICATION, KEYWORD EXTRACTION, RAKE, SIMILARITY MEASURE

1. INTRODUCTION

The text analytics are representing top quality of knowledge from data. Text mining could be most vital analytics of research area to extract the good features and very useful and technical information from large amount of textual information. It is also referred as text data mining there are various types of text analytics techniques such as data mining, information retrieval, natural language processing. Automatic extraction techniques are most helpful. Keyword is that the smallest unit, that expresses that means of entire document that also used for extracting actual information as per user necessities. Key phrases are primarily utilized is find the documents among retrieved system methodologies simple way for defined, revised, remembered. Manual assignment of quality keywords is takes lot of time, unreliable and high priced. Key phrases are selected manually. Assignment key phrases manually are slow method that needs information of subject. Thus, automatic extraction techniques are used. This extraction method to employed for extracting important methods are like text classification, text information retrieval etc.Computing the similarity between documents is a crucial operation within the text information. The similarity measure is the measure of what quantity alike two text documents. The similarity is subjective and is extremely dependent on the domain and Application. Similarity measure is defined as the Distance between varied text documents. The performance of the many algorithms depends upon choosing good features. While, similarity could be a quantity that reflects the strength of relationship between two text documents, dissimilarity deals with the measurement of divergence between two text documents.

2. RELATED WORK

Text Classification is one type of approach which is assigned to predefined categories for text related documents based upon the information. Text Classification is applied to unigram models i.e., Bag Of Words models. In the document level these BOW models doesn't appear to be of co-occurrence of set of words. The author in this work has introduced a new methodology to discover the co-incidence of main feature from another text of Wikipedia pages, incorporate co-occurrence feature to Back Of Words model. Moreover this methodology identifies the two techniques which offer the text classification. The drawback in this

paper is that it normally represents the Wikipedia anchor text in Bunchof Word model which is not represented in text documents and in this work author mainly concentrates on concurrence feature of Wikipedia pages of anchor text [4].

Due to the improvement of web content report on day to day basis, which will have a gradually increment in size with development of World Wide Web. As World Wide Web is taking care of huge number of files so it is time exceptional and tough function will be extracted directly. For this problem automatic method must be used on order to extract main features that are to be printed. Identified a domain keyword extracting system which incorporates a weight methodology that is based on the regular (Conventional) term frequency and inverse document frequency. It is commonly used for precise document term weights; it cannot be replicated by the division of words within the file, degree of terms and distinction into classes. This methodology has introduced another weight in which the weighting scheme could be add the regulate for changes among areas based on first term frequency and inverse document frequency. The Key phrases extraction methodology must be utilized to identify the major information from such type of document set and document categorization to be done. The mail focus is on keyword of news document that are in one of every essential section of text classification, clustering, etc. The problem with in this work is that the author has just performed action on a single domain but not on multiple documents [5]. The authors in this paper have proposed a domain independent technique for instinctively extracting the features from individual document. For this author has proposed an algorithm rule associated degree present an analysis of benchmark corpus and then we have tendency to apply the keyword extraction algorithm. For the current method we will demonstrate the automatic keyword extraction methodology, RAKE which achieves higher accuracy and related call when compared to existing algorithms. In distinction for strategies that depend on natural language process techniques [8].

Text analytics is one of the important tools for search knowledge that is based on vast data. For information related to text analytics, key phrases are utilized. In this paper author have introduced keyword algorithm which is programmed for automatic extracted key phrases from text file. Moreover this technique supports the quantity of grouping text files automatically and determined by exploitation extract to key phrases which are used in categorization. Key phrases are being chosen manually. Thus automatic extraction techniques are advantages in many ways but the drawback within this that the total numbers of clusters are pre-specified in advance which gives economical ways for extracting of text documents from huge quantity of resources with least performance when it is compared with keyword extraction mean algorithm[6].

In this paper the author is designed a two documents and applying text preprocessing and find out the term frequency and inverse document frequency and apply k-nearest neighbor classifier then the document compute the similarity measure text preprocessing method will be used this paper drawback is we used only two documents not taking multiple documents [11].

3. KEYWORD EXTRACTION

To compare and looking out document's content, the simple and easy approach is use keyword. This is the way we will increase performance of text mining. Extracting keywords is one of the very important tasks once operating with text. Readers have the benefit of keywords as a result of they will choose additional quickly whether the text is worth reading. Website creators have the benefit of keywords as a result they will cluster similar content by its topics. Algorithmic programmer has the benefit of keywords because of dimension reduction of text to the very important topics. If consistency of keywords across several documents is very important, I continuously suggest that you simply used to synonym finder and otherwise depend upon the interested text documents using text classification.

4. PROPOSED WORK

Rapid automatic keyword extraction algorithm simplicity and economical modify its used to several applications. RAKE is generally uses Natural language process. This method depends on different topics. This process mainly based on key phrases wherever it can be leveraged. Rake is finds Keypharses from text files. RAKE designed for the observed to key phrases that of frequently used in standard punctuation for ex. a, as, are, the etc.. These are expelled from indexes with retrieved of information system and these are not considered in data analytics because there are premeditated to be in significant. The main reason beyond

this method using is to stop words give as accuracy information in research task. Words that are carrying with accurate information within document are called as content words. And the terms will be co inside of the terms in these candidate key phrases and allow correct sentences in text files.

4.1 SYSTEM ARCHITECTURE

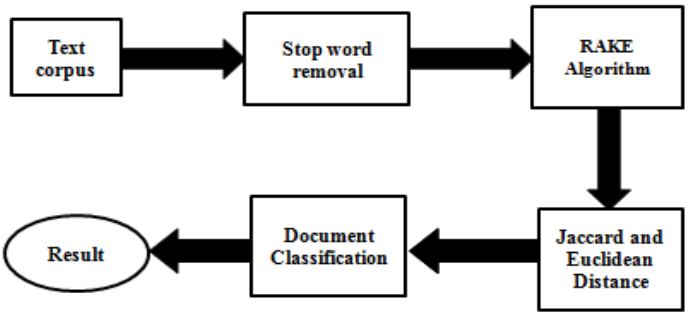

Fig 1: System Architecture

In this system framework we have a tendency to introduced to (RAKE) algorithm first we have a tendency to take the text corpus, then it removed to prevent words and applying the rake algorithm in this to search out candidate keywords the verify word score. Supported word score we have a tendency to apply Jaccard similarity live and Euclidean distance supported 2 similarities live to classify the documents and compare the results.

4.2 WORKFLOW

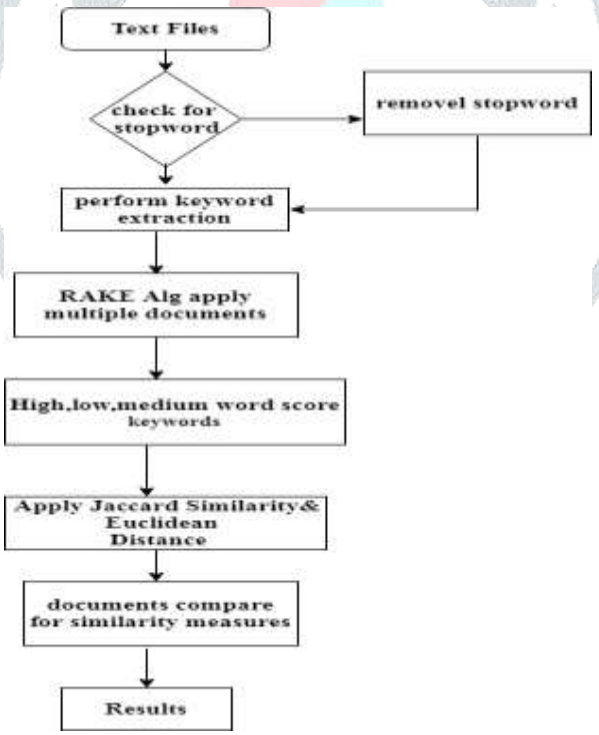

Fig2: Workflow of the proposed system

The work flow of the system indicates step by step method of enforced work.

- **Step1**: 1st we tend to take the text files.
- **Step2**: Then, preprocess the files by removing the stop words and by keyword extraction this keyword extraction algorithm is employed for extract the main features.
- **Step3**: then we tend to apply rake algorithmic program of multiple documents.

- **Step4**: currently apply rake algorithm and we should take the word score Key phrases
- **Step5**: we tend to taken high, low, medium word score Keyphrases then, we tend to apply Jaccard similarity live and Euclidean distance.
- **Step6**: based on word score we tend to compare the Jaccard similarity and Euclidean distance and classify the documents
- **Step7**: we tend to conclude which word score Key phrases give higher performance.

4.3 RAKE ALGORITHM

FILE: Keypharses techniques for in computer science and it will be used to machine learning algorithms in computer scientists.

- Given file we would like to extract keywords, Split the file into associate degree array of words, breaking it at words delimiters.
- Keypharses extracted to a text documents parse the text file to sets for candidate Keypharses.
- To tokenize the terms. Split token groups by punctuation and stop words.
- Identify co occurrences of sequences of unfiltered words. Splits the terms into sequence of continuous terms break the every sequence at a stop word. Every sequences of term is currently a "candidate keyword".
- Candidate keywords: Keypharses techniques – computer science – machine learning algorithms – computer scientists.
- Each and every candidate keyword is identifying table of word co incidence.
- To establish the candidate keyword word score.
- Word score = degree/frequency (word).

Table.4.3.1 Concurrence graph result

	Keypharses	Techniques	Computer	Science	Used	Machine	Learning	Algorithms	Scientists
Keypharses	1	1							
Techniques	1	1							
Computer			2	1					1
Science			1	1					
Used					1				
Machine						1			
Learning							1		
Algorithms								1	
Scientists			1						1

- Then we take words in co-occurrence graph and find out the candidate keywords. Then we should find out the word score of each and every candidate keyword

Table 4.3.2 Word score calculated co occurrence:

	Keypharses	Techniques	Computer	Science	Used	Machine	Learning	Algorithms	Scientists
Degree(w)	2	2	4	2	1	3	3	3	2
Frequency(w)	1	1	2	1	1	1	1	1	1
Word score	1	2	2	1	1	3	3	3	2

The candidate keywords with scores:

We add the Word scores

Keypharses Techniques: 4, Computer Science: 4, Used: 1, Machine learning algorithms: 9, Computer scientists: 4

5. SIMILARITY MEASURES

The similarity measure is a measure of what quantity alike two documents. During this measurement text analytics technique could be a distance with dimensions representing options of the text documents in that distance is little, it will be the high degree of similarity wherever large distance are going to be the low degree of similarity measurement. This is very subjectively and high depends to the area. The values of every part should be normalized. Similarity measure is 1 the documents are similar and the measure is 0 the documents are not similar.

5.1 JACCARD SIMILARITY MEASURE

The Jaccard coefficient commonly used to measure for the overlap between two sets of boolean values P and Q and simply take the number of items in the intersection of P and Q and you divide it number of items in the union of P and Q and so we take the Jaccard coefficient of sets with itself and the set has same size will be also the size of intersection and union. This method measures similarity between finite samples sets.

$$\mathbf{JC(P,Q)} = \frac{\mathbf{P \sqcap Q}}{\mathbf{P \sqcup Q}}$$

In that the jaccard similarity measure value is 1 the document is exact similar and the value is 0 the document is not similar the value 0.5 the jaccard similarity is partially similar

5.2 EUCLIDEAN DISTANCE

Euclidean distance is that the common and easy use of distance. It is also called as a simply distance. Once information is continuous this is can be simplest proximity measurement. This distance between two documents is that the length of the path connects them. This distance is measured by the mix of the each two documents based on vectors.

This can be used to extremely dimensionality area. The equation is:

$$d(g,h) = \sqrt{(g_1 - h_{1)}{}^2} + (g_2 - h_2\ )^2\ + \ldots\ldots + (g_{n-h_n)}{}^2 = \sqrt{\sum_{i=1}^{n}\ {}^{(g_{i-h_i)}}\,2}$$

Here n is that the number of dimensions. that is employed to measures numerical distinction between corresponding variables of documents g and document h.In that the Euclidean distance the value is 1 the document is not similar and the value is 0 the document is exact similar and also here the value below 0.5 the Euclidean distance is similar.

6. RESULTS

6.1 Comparisons of multiple documents based on Rake word score we classify the documents using Jaccard similarity and Euclidean distance

Table 6.1.1 RAKE high level word scores applied to Jaccard similarity & Euclidean distance

Sim measure\doc	td1,td2	td1,td3	td1,td4	td1,td5	td2,td3	td2,td4	td2,td5	td3,td4	td3,td5	td4,td5
JACCARD SIMIALRITY	0	0	0	0	0.25	0.25	0	1	0	0
EUCLIDEAN DISTANCE	4.472	2.44	2.44	1.0	3.46	3.462	4.35	0.0	2.23	2.23

In Fig.3 the RAKE algorithm we have selected to high score keywords then we apply jaccard similarity for two text documents (td3, td4) are exact similar and remaining documents are not similar. In Euclidean distance (td3,td4) documents are exact similar and remaining documents are not similar. This highest score keywords gives better results.

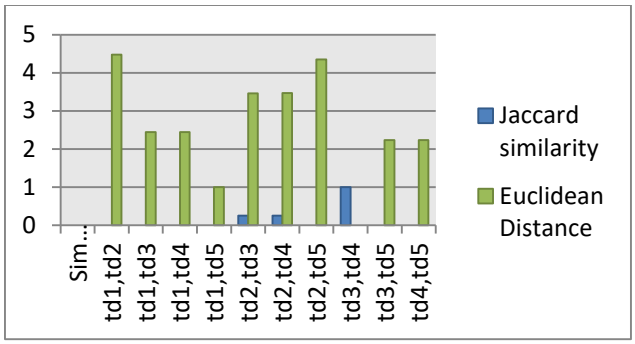

Fig.3 Similarity measures jaccard and euclidean based on high level wordscore

Table 6.1.2 RAKE medium level word scores applied to Jaccard similarity & Euclidean distance

Sim measure\doc	td1,td2	td1,td3	td1,td4	td1,td5	td2,td3	td2,td4	td2,td5	td3,td4	td3,td5	td4,td5
JACCARD SIMIALRITY	0.5	0	0	0	0.25	0.25	0.2	1	0	0
EUCLIDEAN DISTANCE	3.87	4.12	4.12	4.24	2.0	2.0	3.60	0.0	3.0	3.0

InFig.4 the RAKE algorithm we have selected to middle score keywords then we apply jaccard similarity for two text documents (td3,td4) are exact similar (td1,td2) are partially similar and remaining documents are not similar. In Euclidean distance of these two documents are (td3,td4) are exact similar and remaining documents are not similar. The medium score of rake is gives accurate result.

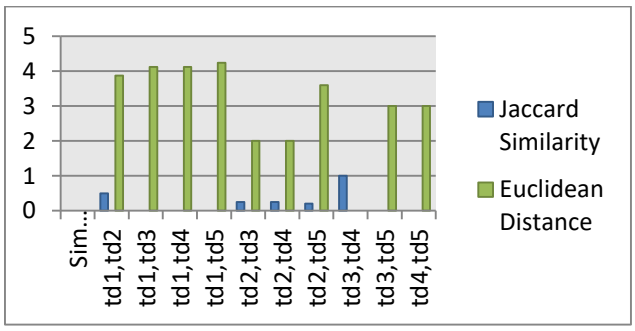

Fig.4 Similarity measures jaccard and euclidean based on medium level word score

Table6.1.3 RAKE low level word scores applied to Jaccard similarity & Euclidean distance

Sim measure\doc	td1,td2	td1,td3	td1,td4	td1,td5	td2,td3	td2,td4	td2,td5	td3,td4	td3,td5	td4,td5
JACCARD SIMIALRITY	0	0	0	0	0	0	0	0.5	0.0	0
EUCLIDEAN DISTANCE	1.41	2.44	1.41	1.41	2.44	1.41	1.41	1.41	2.44	1.41

In Fig.5 the RAKE algorithm we have selected to lowest score keywords then we apply jaccard similarity for two text documents (td3, td4) are partially similar and remaining documents are not similar. In Euclidean distance all documents are not similar. In all above documents result the low score of rake is gives worst result.

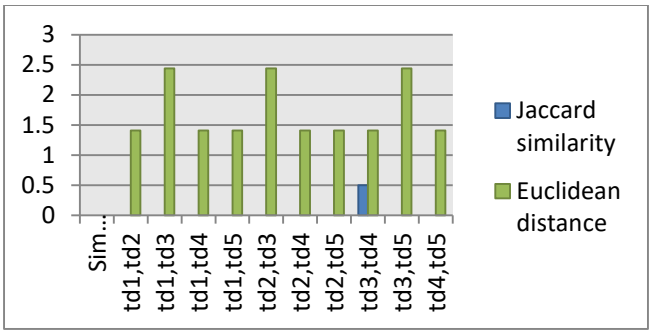

Fig.5 Similarity measures jaccard and euclidean based on low level word score

The comparison tables and graphs the all documents applied Jaccard similarity measure and Euclidean distance the highest, medium is given better results than lower result.

7. CONCLUSIONS

It was observed from the results that among the highest, medium, low level word scores used for evaluation of similarity based on Jaccard, and Euclidean measures the higher, medium word score approaches are giving better results than low level word score. This approach was found to be more suitable for finding word scores from multiple documents when considered. Classification of multiple documents will be done based on word scores using distance measures for measuring similarity among those documents.

8.REFERENCES

[1] Han.J, Kamer.M. .Data Mining Concepts and Techniques.BeiJing:Higher education press, 2001. 285-295.

[2] R. Feldman and I. Dagan.Kdt - knowledge discovery in texts. In Proc.of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117,1995.

[3] Robertson, S. E., "Term specificity [letter to the editor]", Journal of Documentation, Vol. 28, 1972, pp. 164-165./

[4] Soumya George, K, Shibily Joseph,"Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature," 2014.

[5] Rakhi Chakrabory, "Domain Keyword Extraction Technique: A New Weighting Method Based On Frequency Analysis".2013.

[6] Shobha S. Raskar, D. M. Thakore,"Text Mining Using Keypharse Extraction,"Vol 1 No 2, 82-85.

[7] Rahul Nalawade1, Akash Samal2, Kiran Avhad3,"Improved Similarity Measure For Text Classification And Clustering,".2016

[8] Stuart Rose, Dave Engel, Nick Cramer,wendy Cowley, "Automatic Keyword Extraction From Individual Documents ,".2010

[9] "http://hackage.haskell.org/package/rake".2014

[10]"https://dataconomy.com/2015/04/implementing-the-five-most-popular-similarity-measures-in-python/"2015

[11]Radhamothukuri,Nagaraju,Divya chilukuri,"Similarity Measure For Text Classification".2016

[12] Sifatullah siddiqi*,Aditi Sharan,"Keyword extraction from single documents using mean word intermediate distance".2016

[13] Lima Subramanian,R.S Karthik,"Keyword extraction : A comparative study using graph based model and rake".2017

[14] Gali Suresh Reddy,T.V.Rajinikanth,"A TEXT SIMILARITY MEASURE FOR DOCUMENT CLASSIFICATION".2015

[15]G.Sureshreddy,Dr.T.V.Rajinikanth,Dr.A.Aananda Rao," A Frequent Term Based Text Clustering Approach Using Novel Similarity".2014