DOCUMENT SIMILARITY WITH AUTHOR USING STYLOMETRY

Nutan K. Bire

ME Computer Engineering, Matoshree college of Engineering and Research Center, Maharashtra, India

ABSTRACT

Similarity measures have been used for document classification and clustering. To calculate similarity measure in two documents with respect to the distinctive attribute the similarity measure takes the following three occurrences: a) The distinctive attribute appears in both documents. b) The distinctive attribute appears in only on e document and c) The distinctive attribute appears in none of the documents. In the first occurrence, similarity increases as the dissimilarity between the two documents associated with a present feature decreases. In the first occurrence, similarity decreases when the number of presence-absence features increases. In the second occurrence, a fixed value is contributed to the similarity. In the third occurrence, an absent feature has no contribution to the similarity. Proposed system utilizes TFIDF and Stylometry features which will increase accuracy and Identification of Author. By utilization of NB, ID3 classifiers proper the calculation .This classifiers also employ to test performance.

Keyword: - *entropy, classifiers, text mining*

1. INTRODUCTION

Text mining are often outlined because it is analysis of information contained in language. It's like text analytics. Text analytics derives prime quality of knowledge from text and it are often derived through the fashioning of patterns and trends through suggests that equivalent to applied mathematics pattern learning. Text mining incorporates method structuring and parsing in conjunction with the addition of some derived linguistic attributes and removal of others and sequent inserting into the information and eventually evaluating and decoding the output. The high quality data contains some combination of relevance, novelty.The subsequent task represent by text mining like clustering, extraction, granular taxonomies, document summarization, and entity relation modeling[2].

Different techniques of clustering are supported by subsequent things corresponding to knowledge illustration model, similarity measure, clustering model, and clustering algorithmic program. Vector Space Document (VSD) is that the basis for document clustering ways. The illustration of document suggests that feature vector of the words in an exceedingly document within the document corpus. The unique word showing within the documents is atomic feature word therefore, words are the basic units in natural languages to represent similar ideas. Especially, the TFIDF (term frequency inversedocument frequency) of the words are contained in every feature vector. The similarity among two documents is evaluated that is predicated on two corresponding feature vectors, e.g., cosine similarity measure, Jaccard coefficients, and euclidean distance[3].

Document clustering is automatically cluster connected documents into clusters this idea is additional helpful in computer science and machine learning. To handle document clustering variety of ways are implementing on numerous distance measures. The euclidian distance live is wide used. The k-means technique is one among the ways that use the euclidian distance that minimizes the total of the square euclidian distance between the info points and corresponding cluster centers. To decrease the computation complexness, its preferred to an occasional dimensional illustration of documents, since the document house is often of high spatiality[5].

2. RELETED WORK

In Hung chim et al[4] found phrase was most informative feature term for improving document clustering.The phrase based similarity calculated by pairwise similarity based on suffix tree documnet.The suffix tree and document similarity was simple but implementation is complicated.

In Taiping Zhang et al [5] found the correlation preserving indexing which is performed in correlation measure space which improves the performance.The correlation method based on one variable one result can be predicted based on another variable. This becomes not suitable for every variable, and hence the efficient clustering cannot be achieved in this method.

In Ms.K.Sruthi [6] found the multiview point similarity measure offers maximum efficiency and performance. It measures similarty and dissimilarity between objects which are present in different clusters.But multiviewpoint used in interpretation of hierarchy is confusing and complex.

In Khaled M. Hammouda et al[7] announced technique is based on multilayered overlay network of peer neighbours.Supernodes are act as representative of neighborhoods are recursively group to form higher level neighborhood.But search scope was quite large.High overhead and nodes come and go after.Search time may be quite long.

In Prof.P.Pradeep Kumar et al[8] found measure is based on semantic similarity between words, which consists of snippets, page counts.Page counts are measure using dice and jaccard constant.But those system have drawbacks like the huge scale of the web and the large no. of documents in the results set,only those snippets for the top ranking results for a query can be processed efficiently.This drawbacks solved by two class support vector machine.The SVM is only directly applicable for two-class tasks. Hence, algorithms that reduce the multi-class task to several binary problems have to be applied.Parameters of a solved model are difficult to interpret.

In S.Kullback [9]found that in terms of similarity measure of information retrieval,it is difference to between the populations.Kullback-Leibler distance measure used for text categorization.It is simple and effient.But,size of textual data is itself a challenge real size corpus,composed of several hundred of thousand texts,may include several thousand of words.

2. PROPOSED SYSTEM

The flow of proposed system is shown in fig 1.1.

Document Selection-The is selected by user from dataset.

Preprocessing Algorithm:Preprocessing algorithm removes special character and stop word.

Calculation of TFIDF:It is weighting scheme in IR.Create the feature vector for finding the similarity.

Ganeration Document similarity matrix :It is gerates after feature vector creation to find relevant document.

Similarity Measures for Text Processing[SMTP]:The SMTP takes into account the following three cases: a) The feature considered appears in both documents, b) the feature considered appears in only one document, and c) the feature considered appears in none of the documents.The SMTP gives more relevant document and display similarity score.

Classification of cluster:It is done by NB and ID3 classifier.NB requires small amount of training data to estimate the parameters nacessary for classification.ID3 is used to generate tree from dataset.

Stylometry Features: It utilizing in increasing accuracy of system and identification of author.

Similar docs with Idnetification of author:The system gives similarity score, similar document and author of document

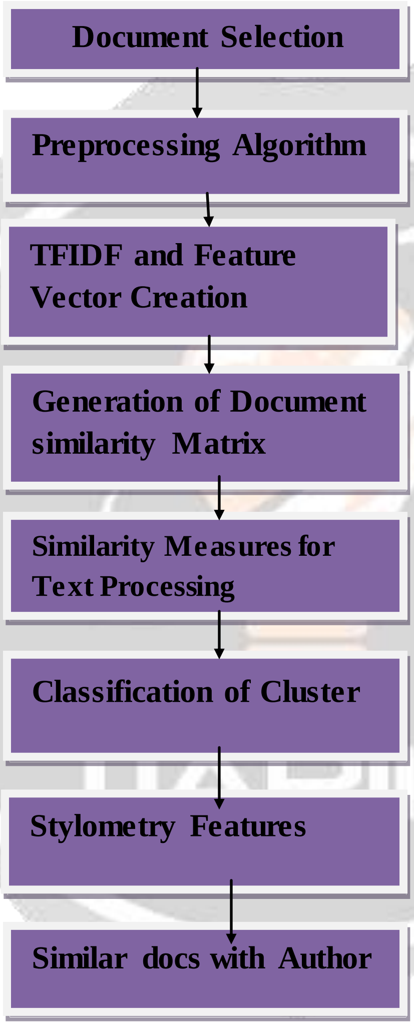
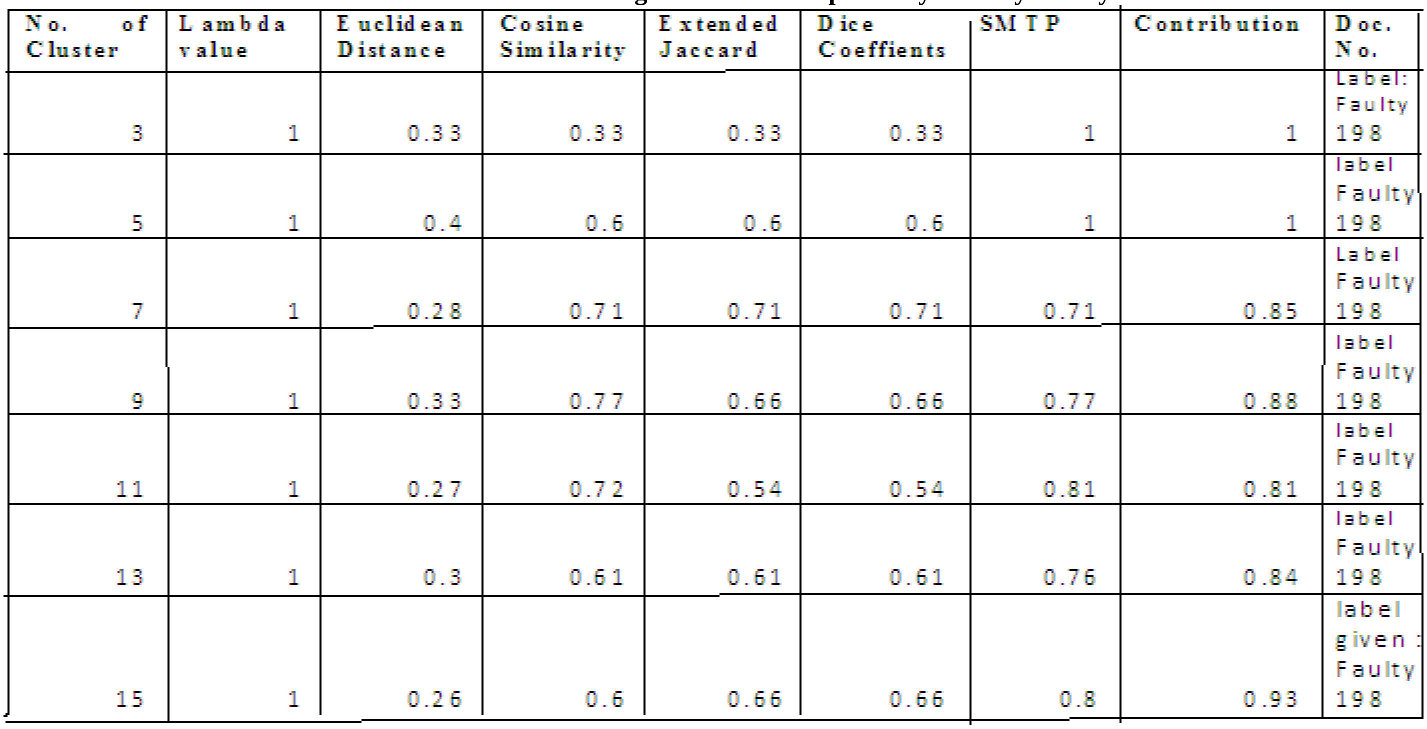

Document Selection
Preprocessing Algorithm
TFIDF and Feature Vector Creation
Generation of Document similarity Matrix
Similarity Measures for Text Processing
Classification of Cluster
Stylometry Features
Similar docs with Author

Fig1.1.Flow of System

3. Result Analysis

WebKB is CMU text learning group and is webpages gathered from World Wide Knowledge Base i.e. WebKB. The documents classified into several different classes. The documents of this data set randomly as classified into training and testing. The classes of WebKB are i.e. Project, Course, Faculty, and Student. The result shown in below table 1.1 each measure has different similarity value. This table shows the comparison of Stylometry Analysis i.e. contribution with other measures. It gives the better accuracy than other measures. If number of clusters is increase Stylometry analysis i.e. contribution gives better accurate value.

Table 1.1 Performance Measures of Existing SMTP and Proposed System Stylometry



No. of Cluster	Lambda value	Euclidean Distance	Cosine Similarity	Extended Jaccard	Dice Coeffients	SMTP	Contribution	Doc. No.
3	1	0.33	0.33	0.33	0.33	1	1	Label: Faulty 198
5	1	0.4	0.6	0.6	0.6	1	1	label Faulty 198
7	1	0.28	0.71	0.71	0.71	0.71	0.85	Label Faulty 198
9	1	0.33	0.77	0.66	0.66	0.77	0.88	label Faulty 198
11	1	0.27	0.72	0.54	0.54	0.81	0.81	label Faulty 198
13	1	0.3	0.61	0.61	0.61	0.76	0.84	label Faulty 198
15	1	0.26	0.6	0.66	0.66	0.8	0.93	label given : Faulty 198

The performance measure of existing system SMTP and Proposed system Stylometry analysis i.e. contribution is shown in below graph. Stylometry analysis i.e. contribution gives the better accuracy than SMTP. If number of clusters is increase Stylometry analysis i.e. contribution gives better value.

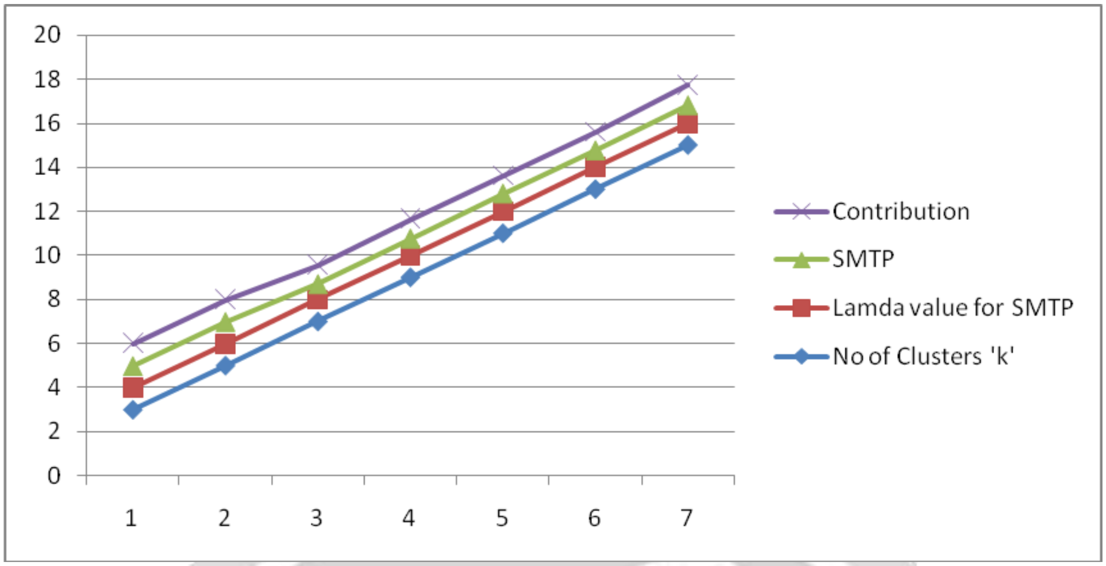


Fig 1.2 Graph of performance measure for Existing system SMTP and Proposed system Stylometry analysis i.e. contribution

4. CONCLUSIONS

The existence and nonexistence of a distinctive attribute are important than the difference between the values associated with current distinctive attribute. The similarity increases when the number of existence-nonexistence distinctive attributes decreases. Two documents are minimum similar to each other if none of the features have non-zero values in both documents. The Euclidean distance, dice coefficients, Jaccard coefficients, Cosine similarity and SMTP [similarity for text processing] as compared to this Stylometry features gives better accuracy and similarity score, Author identification with similar documents.

5. REFERENCES

1.Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering" In IEEE Transaction on Knowledge and Data Engineering, Vol. 26, No.7, July 2014.
2.Nutan Bire``A survey on Classification and Clustering Techiniques using Similarity Measures`` International Journal of Advanced Research in Computer Science and Software Engineering, Jan 2016.
Umajancy.S, Dr. Antony Selvadoss Thanamanii``An Analysis on Text Mining-Text Retrieval and Text Extraction``, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8,August 2013
3.Hung Chim and Xiaotie Deng, "Efficient Phrase-Based Document Similarity for Clustering," In IEEE Transaction on Knowledge and Data Engineering, Vol. 20, No.9, Sept 2008.
4.Taiping Zhang, Yuan Yan Tang,Bin Fang, and Yong Xiang, "Document Clustering in Correlation Similarity Measure Space," In IEEE Transaction on Knowledge and Data Engineering, Vol. 24, No.6, June 2012.
5. Ms.K.Sruthi, "Document Clustering on Various Similarity Measures," International Journal of Advanced Research in Computer Science and Software Engineering, Apr 2013.
6.Khaled M. Hammouda and Mohamed S. Kamel, "Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization," In IEEE Transaction on Knowledge and Data Engineering, Vol. 21, No.5, May 2009.

7.KProf.P.Pradeep Kumar, Naini.Shekhar Reddy, R.Sai Krishna, Ch.Kishor Kumar, M.Ramesh, ``Measuring Of Semantic Similarity Between Words Using WebSearch Engine Approach,`` In International Journal of Engineering Research and Applications Vol. 2, Issue 1, Jan-Feb 2012, pp. 401-404
8.S. Kullback, R.A.Leibler, "On information and sufficiency", Annu. Math.Statist.,Vol.22,No.1,pp.79-86.
9. Robert Goodman, Matthew Hahn, Madhuri Marella, Christina Ojar, Sandy Westcott, “The Use of Stylometry for Email Author Identification: A Feasibility Study” CSIS, Pace University 2007.