© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 01 | January - March 2018
Paper is available at www.jrps.in | Email : info@jrps.in

Measuring Similarity between Documents Using TF-IDF Cosine Similarity Function

Manoj Chahal[1]
Master of Technology (Computer Science and Engineering)
Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India[1]

Abstract: Tremendous amount of information is present in all over the world in the form of document databases. This database is expanding exponentially. Measuring similarity between documents is very difficult task. The field of similarity deals with the problem of documents similarity and for this problem various models and functions have been proposed. Measuring similarity between documents plays a crucial role in various research and application. Plagiarism detection , information retrieval ,text based research, clustering etc methods are possible after measuring similarity between documents . In this paper we would like to compare documents and measure the similarity between them using TF (term frequency) -IDF (inverse document frequency) cosine similarity function.

Keywords: Similarity measures, Term frequency, inverse Document Frequency, Document Database, Cosine, String based Similarity, Vector Space Model.

I. Introduction

As the technology is developing simultaneously information related to various field in the digital world is also increasing on high pace. There are various databases available in digital world which contain huge amount of information or document to retrieve or finding the related documents based on an input query is an interesting problem in this digital world. Many algorithm have been developed for this purpose that take a document or input query and match it with the document databases. This algorithm matches documents using similarity function and also ranks them according to their similarity value.

Similarity measuring method can be divided into three categories : string based method , corpus based method and knowledge based method.

String based method measure similarity between two or more text string by applying appropriate matching function. It is divided in to two type character based and term based .various algorithm for character based matching function are longest common substring ,Daumier-Liechtenstein, jaro, jato-winkle etc. Algorithms used for term based method are block distance m cosine similarity , dice Similarity, Jacquard similarity,overlap coefficients etc. corpus based method measure the similarity between words according to information extract from large corpus. A corpus is a large collection of written or spoken texts that is used for language research. Various corpus based similarity measures are hyperspace Analogue to language , Latent Semantic Analysis ,Generalized latent semantic Analysis etc. Knowledge based similarity is a semantic similarity measure and it based on identifying the degree of similarity between words and it is uses various information derived from semantic networks.

This paper concentration on string based method and measures the similarity between the documents using TF and *IDF cosine similarity function. Its also calculate the percentage of similarity between the documents.*

1. TF- IDF Cosine similarity Function

cosine similarity function determine the similarity between the given document. It measure angle between the given documents . Lower the angle higher the degree of similarity and higher the angle lower the degree of similarity between documents.

Similarity between D_i and D_j is defined as :-

$$COS\ (Di, D_j) = (\sum_{k=1}^{n} W_{ik}\, W_{jk}) \big/ (|D_{i|}|)\ (|D_{j|}|)$$

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 01 | January - March 2018
Paper is available at www.jrps.in | Email : info@jrps.in

Where Di and Dj are the corresponding weighted term vectors and |Di| is the length of the document vector Di

Documents in a collection are assigned terms for a set of n terms.

The term vector space W is defined as:-

if term K does not occur in document D_i then $W_{ik}=0$

if term K occur in document D_i then W_{ik} is greater than zero and W_{ik} is called the weight of term k in document D_i.

2. Weighting Term frequency (tf)

Term frequency is define as the number of time a term appear with in a document. Suppose j appear f_{ij} time in Document D_i then term frequency of j if f_{ij} .

$$Tf = f_{ij} / mi$$

$$\text{where } m_i = \max(f_{ij})$$

m_i is the maximum frequency of any term in document D_i.

3. Weighting Inverse Document Frequency (idf)

It define that a term that occur in a few document is likely to be a better discriminator than a term that appear in most or all documents.

$$idf_j = \log(n / n_j) + 1$$

n=number of documents

n_j=number of time j term appear in document.

4. Full Weighting: tf-idf

Weight of term j in document i= (term frequency) * (inverse document frequency)

$$t_{ij} = tf_{ij} * idf_j$$

Cosine similarity measure for tf-idf is:

$$COS(Di, D_j) = (\sum_{k=1}^{n} t_{ik} t_{jk}) / (|t_i|)(|t_j|)$$

where t is product of term frequency and inverse document frequency |ti| is the length of the document vector Di

II. Literature Survey

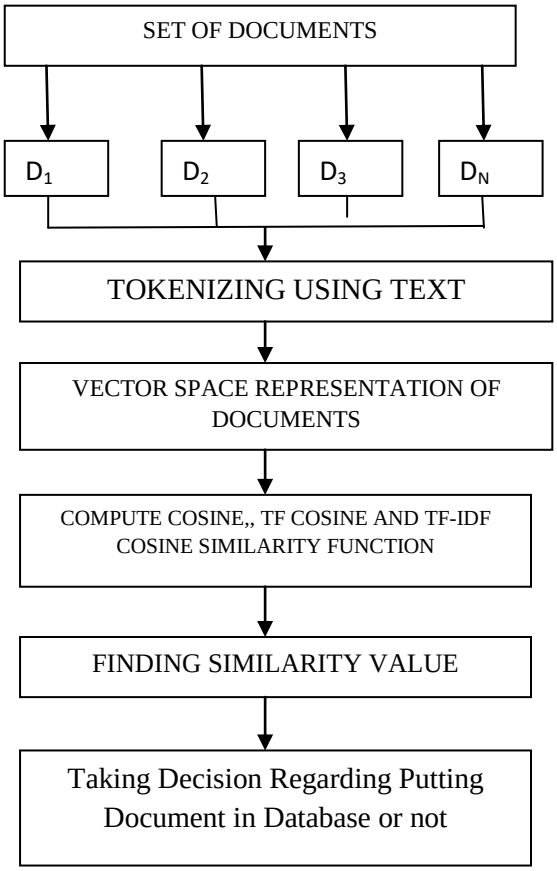
M.K Vijaymeena & K. Kavitha [1] described various clustering algorithm that used for similarity measures in text mining. They also dicussed text similarity by partitioning them into three approaches: string based , knowledge based and corpus based similaritys.A.K. Patidar , J. Agarwal et. al. [2] described the impact of four different similarity measure function on shared nearest neighbor clustering and also compared result with each other. They also explained similarity graph generated bu eucclidean,cosine,jaccard and correlational function.R. Deshpande , K Vaze et al.[3] described sentiment analysis for document similarity and clustering algorithms. They also discussed performance of various clustering algorithm and compared with each other.David Buttler [4] described survey of document structural similarity algorithms that include optimal tree edit distance algorithm and various approximation algorithms. They also explained cluster comparision between path and TED metrics. Tung

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 01 | January - March 2018
Paper is available at www.jrps.in | Email : info@jrps.in

Khuat and Le Thi Hanh [5] described algorithm for compar and measure similarity between two documents. They also explaibned term based algorithm using cosine similarity measure and character based algorithm for fingerprint and winnowing. Abhishek Jain , Aman Jain et al [6] described various model and similarity measure to determine the siomilarity between two object and also explained approaches to match a query text against a set of indexed document.Vikash Thanda and Dr. Vivek Jaglan [7] discussed jaccard , dice and cosine similarity coefficient and compare witgh each other to find the best fitness value after appliying genectic algorithm with different crossover and mutataion value. Z.M.Myint And M.Zinoo [8] described three modified TF-MIDF method to solve the no-relevant problem by modifying the IDE equation and analyses the modified IDF method to ascertain with MIDF method. K. Ramana and A. Venkataramana [9] described method for improving effenciency of clustering using Hadoop Map reduce method and also explained whole process of Map-Reduce procedure. S. Anitha Elavarasi and J. Akilandeswari [10] described frequency and TF-IDF based Categorical data clustering technique based on cosine similarity measure. They also explained performance of system on similarity threshold selected for the clustering process. Pragati Bhatnagar et al. [11] discussed the applications of GA for improving retrieval efficiency of IRS. GA was used to find an optimal set of weights for components of combined similarity measure consisting of different standard similarity measures that are used for ranking the documents. Siti Nurkhadijah Aishah Ibrahim et al. [11] presented a model of hybrid GA-Particle Swarm Optimization (HGAPSO) based query optimization for Web information retrieval. The keywords are used to produce new keywords that are related to the user search.

III. Experiment

- Extract Documents from Database
- Apply text Analyzer Tool to find the frequency of the term used in Documents.
- Encoded them into weight Matrix.
- Encoded Matrix is given as input to the TF-IDF cosine similarity function.
- Process the given input by using TF-IDF cosine similarity function to get output.
- Output data define the similarity of document with respect to database.



© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 01 | January - March 2018
Paper is available at www.jrps.in | Email : info@jrps.in

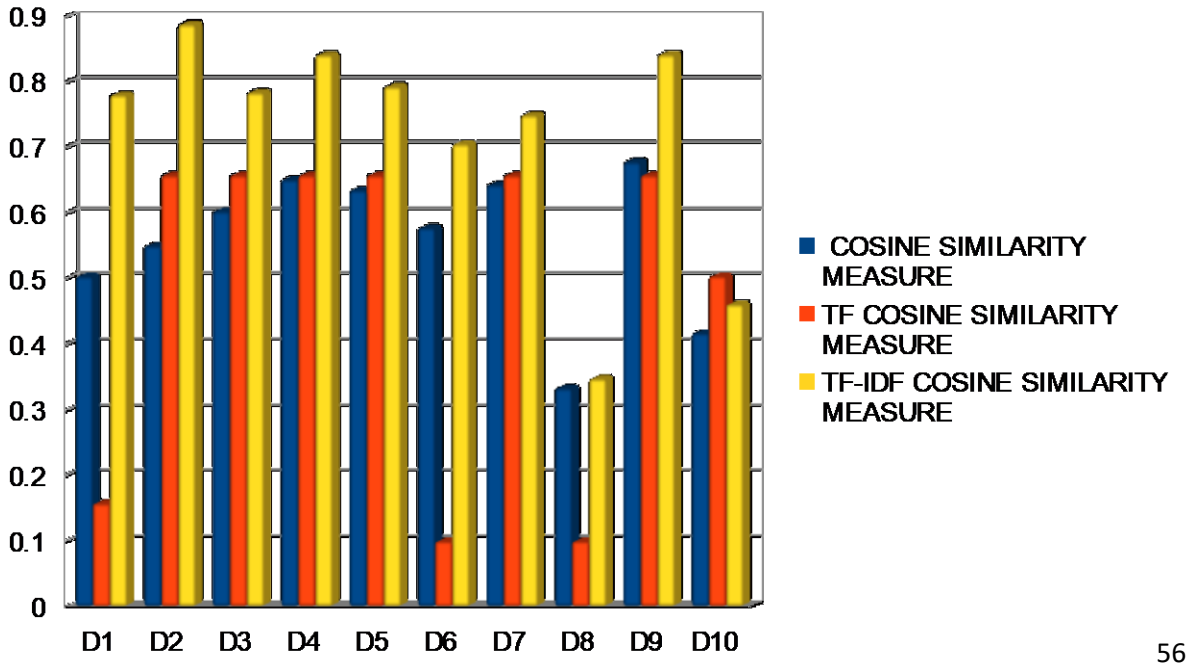
IV. Result

Cosine, TF cosine, TF-IDF cosine matching measure the similarity of document with respect to database. With the help of this we can find whether the document belong to the given database or not.

Database = (D_1, D2, D3, D4, D5, D6, D7, D8, D9, D10)

Table 1.1: Average similarity measure using cosine ,TF cosine and TF-IDF cosine similarity function.

DOCUMENTS	COSINE SIMILARITY MEASURE	TF COSINE SIMILARITY MEASURE	TF-IDF COSINE SIMILARITY MEASURE
D_1	0.503645	0.157735	0.780169
D_2	0.550201	0.657735	0.887904
D_3	0.603121	0.657735	0.783909
D_4	0.651355	0.657735	0.841108
D_5	0.635102	0.657735	0.792810
D_6	0.577603	0.100000	0.705488
D_7	0.643220	0.657735	0.749400
D_8	0.333350	0.100000	0.347762
D_9	0.678181	0.657735	0.841644
D_{10}	0.417153	0.504145	0.461305



V. Conclusion and Future Works

In this digital world, there are a huge number of databases which contains different type of documents. Datebases are categorized according to the data stored in them. Manullay it is very difficult to add a new document in respective database. In order to do this cosine similarity , TF cosine similarity and TF-IDF cosine similarity is use .If the similarity measure is less than 0.5 than document does not belong to the particular database but if similarity measure is more than 0.5 than document is belong to the particular database. So we store document in database if similarity measure is more than 0.5 . It is observe that documents D_8 and D_{10} similarity measure is smaller than 0.5 .So that document D_8 and D_{10} does not store in database and all other document store in given database.

VI. References

[1] M.K.Vijaymeena and K.Kavitha, "A Survey On Similarity Measures In Text Mining", Machine Learning and Applications: An International Journal (MLAIJ), vol 3,no.1,pp 19-28, march 2016.

[2] A.K.patidar, Jitender Agrawal and N.Mishra " Analysis of different Similarity Measure Functions and Their Impact On Shared Nearest Neighbour Clustering Approach",International Journal of Computer Application, ISSN 0975-8887, vol. 40, no.16,pp. 1-5, Feb. 2012.

[3] Rugved Deshpande , Ketan Vaze et all., "Comparative Study of Document Similarity Algorithms and Clustering Algorithms for Sentiment Analysis",International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),ISSN 2278-6856,vol.3,Issue 5,Sep-Oct 2014.

4] David Buttler, "A Short Survey of Document Structure Similarity Algorithms",Lawrence Livermore National Laboratory ,Livermore, CA 94550.

[5] Tung Khuat and Le Thi Hanh, "A Comparison of Algorithms used to measure the Similarity between two documents", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),ISSN 2278-1323,Volume 4, Issue 4, pp 1117-1121,April 2015

[6] Abhishek Jain,Aman Jain er all. "Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model", International Journal of Computer Applications (0975 – 8887) ,Volume 164 ,No 6, pp 28-30, April 2017

[7] Vikas Thada, "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm", International Journal of Innovations in Engineering and Technology (IJIET),ISSN 2319-1058, vol. 2, Issue 4, pp.202-205, Aug. 2014.

[8] Zun May Myint and May Zin oo, "Ananysis Of Modified Inverse Document Frequency Variants For Word Sense Disambiguation", International Journal Of advanced Computational Engineering and Networking,ISSN 2320-2106,Vol. 4,Issue 8 pp.46-50, Aug 2016.

[9] K. Ramana and A. Venkataramana ," Enhance the Efficiency of Clustering by Minimizing the Processing Time using Hadoop MapReduce", International Journal of Advanced Research in computer science and software Engineering ,ISSN 2277-128X, vol 5 Issue 9 ,pp- 841-845 ,Sep 2015.

[10] S.Anitha Elavarasi and J.Akilandeswari, "Categorical Data Clustering using Frequency and TF-IDF based cosine similarity", Proceedings of the Intl. Disciplinary Research in Engineering and Technology,ISBN 9778-81-929742-0-0,pp 39-43, 2014.

[11] Pragati Bhatnagar and N.K. Pareek, " A combined matching function based evolutionary approach for development of adaptive information retrieval system",International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, vol. 2, no. 6,pp. 249-256, Jun. 2012.

[12] Nurkhadijah Aishah Ibrahim, Ali Selamat, Mohd Hafiz Selamat, "Query optimization in relevance feedback using hybrid GA-PSO for effective webinformation retrieval", IEEE Transaction DOI 10.1109, pp. 91-96, 2009.