© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 02 | April - June 2018
Paper is available at www.jrps.in | Email : info@jrps.in

A Survey of Document Ranking and Similarity Using Combination of Various Matching Function

Manoj Chahal

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

Abstract: - The Volume of information in this world of digitalization is so vast and present in various forms. The major problem we face related to all these information sets is their organization. To use this information effective and efficiently we categorize or classified them according to their specialization. Without categorizing garbing the relevant information is not an easy task. To make it easy different methods are applied and these methods allow the user to take and put the specific information or document quickly into their respective database. The main objective of this paper is to use combination of cosine-Jaccard ,Jaccard-dice and cosine-dice matching function to find the similarity between documents and ranking them according to their similarity into their respective database and store them into the appropriate classification.

ISSN : 2278-6848 03
9 772278 684800
© International Journal for Research Publication and Seminar

Keywords: Rank, Combined Matching Function, Databases, Documents, Similarity Measure, Classification

I INTRODUCTION

The size of text document in digital repositories is increasing at a very high speed. Digital repositories like digital libraries and Internet are full of text resources and organizing these text resources is our piratical need now-a-day plus a challenge too. For organizing a large number of text resources we have to make small or minimum number of coherent groups based on their content or text. Text clustering is the method which makes it happen. It is methods which organize a large number of unordered document into small number of meaningful clusters. It divide a collection of large document into different specific categories so that it result having a same type of information in an individual category. For better understanding we take an example say a university is having a lot of information related to marks of their students. But to find out hoe the students of computer science performed it can become difficult to analyses for our ease on the website of university we have different links of department which give specific information about their student organizing helps us in easy data retrieval.

For putting the document into appropriate and respective category we use similarity measure between the document and document of category database. Similarity measure is a function which is used to measure the degree of similarity between the documents. This also allows us to rank the documents on the basis of degree.

Various similarity measure technique is use to measure the similarity between documents some of them are cosine matching function, Jaccard matching function, dice matching function etc. In this paper we combine cosine-dice, cosine-jaccard and jaccard-dice to measure the similarity between documents. The basic formula for cosine, jaccard and dice are:-

Cosine similarity measure

It is a measure of similarity between two vectors that measure angle between them. Both document and query is representing in the form of vector.

Cosine formulation as shown below:

$$\cos\theta = \frac{\sum_{i=1}^{t} x_i * y_i}{\sqrt{\sum_{i=1}^{t} x_i^2}\sqrt{\sum_{i=1}^{t} y_i^2}}$$

Where x and y are query and document vectors.

Jaccard similarity measure

Jaccard similarity measure is defined as the size of intersection divided by the size of union of the sample sets. Sample sets mean terms in query and documents.

Jaccard formulation is given as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 02 | April - June 2018
Paper is available at www.jrps.in | Email : info@jrps.in

Where A and B represent as query and document terms. Jaccard similarity measure lies between 0 and 1.If the document and query are more similar than jaccard value lies near 1 but if it is less similar than it lies near 0.

Dice similarity measure

Dice similarity measure defined as multiply by two the size of intersection divided by the sum of length of both sample sets. Here sample sets mean terms in query and documents.

Dice formulation is given as:

$$D(A,B)=\frac{2\times|A\cap B|}{|A|+|B|}$$

Where A and B represent as query and document terms. Dice similarity measure lies between 0 and 1. If the document and query are more similar than Dice value lies near 1 but if it is less similar than it lies near 0.

II PREVIOUS WORKS ON INFORMATION RETRIEVAL

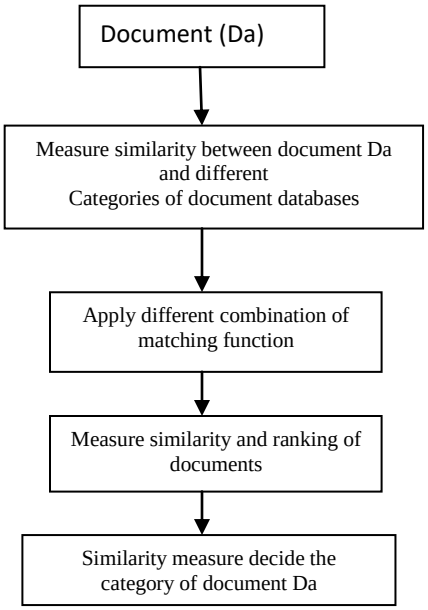
There are several studies that used Similarity function for ranking and measuring similarity between text documents.

Jasmine Irani et al [1] Described the survey of various clustering techniques and similarity measure based on distance based clustering. They also explained the combinated advantage of current system and their limitation and how to overcome the limitation. M.K. Vijaymeena and K. Kavitha [2] described various clustering algorithm that used for similarity measure in text mining. They also dicussed Text similarity by Partitioning them into three approaches: Strng based , Knowledge based and Corpus based similarity. Satya P. Kumar Somayajula et al. [3] described concept based similarity measure into the temporal-semantic clustering model for event detection in mewspaper articales. They also explained hierarchical approach for clustering document based on similarity measure. Manan Mohan Goyal , Neha Agarwal et al. [4] described K-Mean clustering using different similarity measures. They also discussed comparision based on clustering between fuzzy and cosine similarity measure. R. Umamaheswari and K. Rajesh [5] described clustering and importance of clustering in text document database. They also explained hierarchical clustering technique entitled “sub leader algorithm” along with cosine similarity is to cluster the document.

Mirza Ruhi Masuma et al. [6] discussed Text document classification and clustering with the help of similarity function. They also explained various component of information Reterival system and their uses in retrieving relevant information. Pragati Bhatnagar et al. [7] discussed the applications of GA for improving retrieval efficiency of IRS. GA was used to find an optimal set of weight for components of combined similarity measure consisting of different standard similarity measure that are used for ranking the documents. D.Renukadevi and S.Sumathi [8] described the importance of classification and clustering in document similarity. They also explained Term based, Inverse Document frequency based similarity function and fuzzy c-mean algorithm that used this similarity function for classification and clustering. P.Sowmya Lakshmi et al. [9] described process of building multi-classifier model for textual data. They also attempt to explore different similarity measure, different feature selection techniques in the process of desiging textual multi-classification. S. Brin and L. Page [10] described the working of search engine. They also explain the architecture of search engine.

III. EXPERIMENT

In this section we discuss how the experiment is conducted and result occur during experiment. In our experiment we take three category of database (Cat 1, Cat 2 and Cat 3). Each contains 5 documents. Da is the document which is inserted into one of these three categories of databases. The categories are decided on the basis of similarity measures. The similarity is measure using combined matching function (cosine-dice, cosine-jaccard and dice-jaccard).



IV. RESULT

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 02 | April - June 2018
Paper is available at www.jrps.in | Email : info@jrps.in

D1, D2, D3, D4, D5 represent Documents present in different categories of databases. In our experiment Category 1 database contain five documents (D11 , D12 , D13, D14 , D15) , Category 2 database contain five documents (D21 , D22 , D23, D24 , D25) and

Category 3 database contain five documents (D31 , D32 , D33, D34 , D35)

	Cosine-dice maching function	Cosine-jaccard matching function	Jaccard-dice matching function
(Da , D11)	0.580177	0.4771475	0.4527315
(Da , D12)	0.3589275	0.2755965	0.259417
(Da , D13)	0.14693	0.1147005	0.1116935
(Da , D14)	0.372727	0.2789635	0.2405425
(Da , D15)	0.1342995	0.1238895	0.085366

Table 1.1 Comparing Document Da with the Documents in Category (Cat1)

	Cosine-dice maching function	Cosine-jaccard matching function	Jaccard-dice matching function
(Da , D21)	0.1742215	0.133577	0.1324765
(Da , D22)	0.1709965	0.1319195	0.130594
(Da , D23)	0.2454675	0.192625	0.1905195
(Da , D24)	0.2048665	0.1485625	0.152496
(Da , D25)	0.812315	0.6860275	0.67716

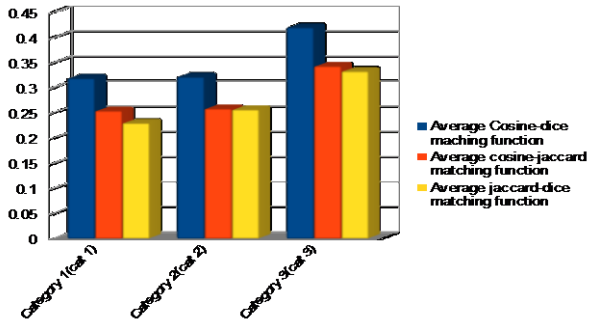
Table 1.2 Comparing Document Da with the Documents in Category (Cat2)

	Cosine-dice maching function	Cosine-jaccard matching function	Jaccard-dice matching function
(Da , D31)	0.377124	0.31332	0.313311
(Da , D32)	0.286838	0.2232445	0.2198145
(Da , D33)	0.230164	0.195032	0.185567
(Da , D34)			
(Da , D35)	0.4514505	0.3413665	0.341286

Table 1.3 Comparing Document Da with the Documents in Category (Cat3)

Document Da similarity measure	Average Cosine-dice maching function	Average cosine-jaccard matching function	Average jaccard-dice matching function
Category 1(cat 1)	0.3186122	0.2540595	0.2299501
Category 2(cat 2)	0.3215825	0.2585423	0.2566492
Category 3(cat 3)	0.4197424	0.3427809	0.3325855

Table 1.4 Average Similarity Measure between Document Da and Cat1 ,



Cat2 , Cat3 Document Databases

Graph 1.1 show average similarity of document Da with three categories of document databases (Cat1 , Cat2 , Cat3

Above graph and table show that document Da measure highest similarity with documents of category 3 (Cat3) database. It decides that document Da belong to category 3 database and we insert document Da in category 3 databases.

Ranking of Document Da in Category 3 (Cat3) database:-

	Cosine-Dice similarity measure	Cosine-Jaccard similarity Measure	Jaccard-Dice Similarity Measure	Ranking of Documents

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 09 Issue: 02 | April - June 2018
Paper is available at www.jrps.in | Email : info@jrps.in

Da	0.516452	0.451729	0.4437715	5
D31	0.590194	0.5093725	0.486063	2
D32	0.538771	0.467003	0.455167	3
D33	0.4816895	0.4386545	0.42054	6
D34	0.59782	0.518681	0.5006235	1
D35	0.534817	0.4522695	0.445022	4

Table 1.5 Ranking of documents based on combination of similarity measure

V CONCLUSION

Large amount of database is present in Internet or web world. Each database is categories according to the information store in databases. This information is updated every day as new document is created or inserted in Internet. Manually insertion and ranking of document into particular database is difficult task. To solve this combined similarity measure function is use. This combined similarity measure function calculates similarity value between documents with the documents present in different categories of databases. By comparing our document with all those categorized databases and we put our document in that category where we get highest similarity measure value and also giving ranking them according to similarity value.

VI REFERENCES

[1] Jasmine Irani et al. , "clustering Techniques and Similarity Measures used in Clustering : A Survey" , *International Journal of Computer Application*(0975-8887),volume 134 ,No.7 ,pp 19-28 , January 2016.

[2] 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.

[3] Satya P Kumar Somayajula et al. , " Application of the concept-based similarity measure in topic detection" , *International Journal of computer science and Information Technology* , ISSN 0975-9646 , Vol 2(4) , pp 1743-1746 , 2011.

4] Manan Mohan Goyal , Neha Agrawal et al., "Comparision Clustering Using Cosine anf Fuzzy Set based Similarity Measures of Text Documents" *International Conference on Computing and Communication Systems* 2015 (I3CS'15), ISBM: 978-1-4799-5857-01, 2015

[5] R.Umamaheswari and K. Rajesh , "Text Clustering Using Cosine similarity and Matrix Factorization" , *International Journal of Research in Computer and Communication Technology* , ISSN(0) 2278-5841 , ISSN(P) 2320-5156 , Vol 3 , Issue 10 ,pp 1343-1347,October 2014.

[6] Mirza ruhi Masuma et al. , "Text Classification and Clustering through Similarity Measures" , *International Journal of Latest Technology in Engineering, Management & Applied Science* , ISSN 2278-2540 , Volume V , Issue III,pp 91-94 , March 2016.

[7] 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.

[8] D.Renukadevi and S.Sumathi , " Term Based Similarity Measure For Text Classification and Clustering Using Fuzzy C-Means algorithm" , *International Journal of Science And Technology Research* , ISSN 2278-7798 , Volume 3 , Issue 4 , pp 1093-1096 , April 2014.

[9] P.Sowmya Lakshmi et al. , "Different Similarity Measure For Text Classification Using Knn" , *IOSR Journal of Computer Engineering , ISSN 2278-0661* , ISBN 2278-8727 , Volume 5 , Issue 6 , pp 30-36 ,Sep-Oct 2012.

[10] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. Seventh Int'l Conf. World Wide Web* (WWW '98), pp. 107-117, 1998.

.