© 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:

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