IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. VI (Jan – Feb. 2015), PP 19-25
www.iosrjournals.org

Correlation Coefficient Based Average Textual Similarity Model for Information Retrieval System in Wide Area Networks

Jaswinder Singh[1], Parvinder Singh[2], Yogesh Chaba[3]
[1,3](Department of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology, Hisar, Haryana , India
[2](Department of Computer Science & Engineering , Deenbandhu Chhotu Ram University of Science & Technology, Murthal, Sonepat, Haryana, India

Abstract: *In wide area networks, retrieving the relevant text is a challenging task for information retrieval because most of the information requests are text based. The focus of paper is on the similarity measurement, performance evaluation and design of information retrieval techniques using the four similarity functions i.e. Jaccard, Cosine, Dice and Overlap. The performance evaluation of these similarity functions has been done for the similarity between the documents retrieved by the search engine for the entered text using the vector space model. The correlation coefficient was applied for evaluating the performance of similarity functions. All the possible combination of similarity functions have been explored and textual similarity model has been proposed for the information retrieval system in wide area networks.*

Keywords: *Information Retrieval System, Similarity Functions, Proposed Model of textual similarity, Wide Area Networks.*

I. Introduction

The large amount of information available from the wide area networks is in the form of text, image, videos and songs i.e. there is variety of data available in the web world [1], [2], [3], [4]. As the major content available from the web world is in the form of text so to retrieve the relevant text is still a challenge for any information retrieval system in wide area networks .The user usually types his or her query as text in the search box of any information retrieval system which is search engine in most of the cases. The search results of the entered keyword in some cases might not display the required documents which might be due to the lack of the search method of the user or due to lack in knowledge of how to use the keyword. The goal of the paper is to design the information retrieval techniques using the four similarity functions i.e. Jaccard, Cosine, Dice and Overlap similarity functions for enhancing the textual similarity between retrieved documents for the entered query as text in the chosen search system. This paper is organized as follows.

The first section of paper describes the brief introduction about the heterogeneity of the data and second section describes the brief introduction about the information retrieval system and about information retrieval techniques used in wide area networks. The third section is about the similarity functions and the related work. The fourth section of the paper describes the steps of the experimentation. The fifth section of the paper describes the results obtained from the experiment. The sixth section of the paper is about the proposed model of the textual similarity in which three approaches are proposed for the similarity scores for the retrieved documents for the entered query and model is represented as a triangle in which the three vertices of triangle represents the results obtained from the three proposed approaches of the information retrieval techniques using the four similarity functions i.e. Jaccard, Cosine, Dice and Overlap similarity functions .The seventh section of paper concludes the results obtained from the three proposed approaches.

II. Information retrieval system and information retrieval techniques in wide area networks

As we know that there is vast amount of information available in the form text in the web world. To retrieve the relevant information from the web world, information retrieval system is used which delivers the relevant information to the user. Any information retrieval system contains three main components i.e. query subsystems, matching mechanism and document database [1], [5]. Fig.1 shows the block diagram of typical information retrieval system. Matching mechanism retrieve those documents that are judged to be relevant to it by the use of similarity functions or similarity measures .Similarity functions or the similarity coefficients or the similarity measures are defined as the functions which measure the degree of similarity between query entered by the user and documents retrieved using the search system [1]. The technique for comparing the query and document is called the retrieval technique and Nicholas J. Belkin et.al [6] described that there are two types of information retrieval techniques i.e. exact match techniques and partial match techniques. Partial match techniques have the advantage over the exact match techniques that these also include those documents that

exactly match with the query in the retrieved documents. Next level of the classification of retrieval techniques distinguishes the techniques that compare the query with the individual document representation and the techniques based on the representation of network of documents. Individual representation based techniques were further classified by Nicholas J. Belkin et.al [6] as the structure based and feature based techniques. In the feature based techniques queries and documents are represented as sets of features such as terms. This category includes the techniques based upon the formal models which include the vector space model, probabilistic model and others.

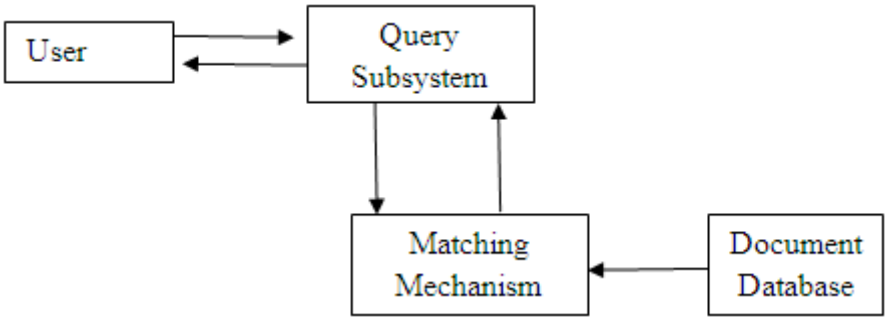


Figure1: Block diagram of typical information retrieval system

III. Similarity functions and related work

In the information retrieval, similarity functions are functions which are used to measure the similarity between user query and documents. To retrieve the documents in response to a user query is the most common text retrieval task. For this reason, most of the text similarity functions have been developed that take input as a query and retrieve the matching documents. Various similarity functions have been developed but how they are best applied in information retrieval and how similarity values or rankings should be interpreted is not answered yet. It is therefore difficult to decide which similarity function should be used for a particular application as wide range of similarity functions were developed which are used in the different fields such as information retrieval [7], image retrieval [8], genetics and molecular biology [9] and chemistry [10]. Several similarity functions were surveyed by McGill et.al [11].Sung-Hyuk Cha [12] classified similarity measures for comparing the nominal type of histograms. The vector space model was used by William P. Jones et.al [7] for the geometric representation of similarity measures i.e. Inner Product, Cosine, Dice and Overlap. The String-based, Corpus-based and knowledge-based are the three categories of textual similarity functions described by Wael H. Gomaa et. al [13]. It was further described that the character-based approach and the term based approach are the two sub categories of the string-based approach The term-based approach includes Jaccard, Cosine, Dice and Overlap similarity functions. Suphakit Niwattanakul et.al [14] concluded that Jaccard similarity coefficient is suitable sufficiently to be employed in the word similarity measurement. Wael Musa Hadi et.al [15] concluded the Cosine similarity measure outperforms Jaccard and Dice similarity functions using the vector space model. From the literature survey of the similarity functions it was found that there are wide range of similarity functions and various authors have used them differently in the different domains and our work is different from their work in view that we have explored all the combinations of four similarity functions i.e. Jaccard, Cosine, Dice and Overlap similarity functions and proposed a model for the design of information retrieval techniques using similarity functions in wide area networks using the vector space model .

IV. Experimentation

In the experiment Google search engine was used as the search tool to retrieve the web pages for the entered keyword and ten queries were considered for the similarity measurement using four similarity functions i.e. Jaccard, Cosine, Dice and Overlap. For the performance evaluation and design of information retrieval techniques with the said similarity functions using the vector space model in wide area networks, binary weights were used for the representation of query and documents which means that the weight of term is '1' if term occurs in the document and '0' if the term does not occurs in the document. The similarity was measured by the four similarity functions i.e. Jaccard, Cosine, Dice and Overlap.

The experiment was divided into the different steps.

Step1: Similarity measurement using the similarity functions.
Step2: Analysis of the similarity functions based upon the similarity scores.
Step3: Correlation coefficient measurement for the similarity scores obtained from step 2.
Step4: Exploring all the combinations of similarity functions.
Step5: Performance evaluation of the similarity functions based upon the correlation coefficient.
Step6: Proposed the model for textual similarity using similarity functions.

V. Results Obtained From Experimentation

Step 1: Similarity measurement using the similarity functions

The similarity between the documents retrieved for the entered query in the search engine was measured and the process was repeated for the ten different queries and similarity scores were obtained by using the Jaccard, Cosine, Dice and Overlap similarity functions and average similarity value was measured for the obtained values of similarity for the different queries [16]. The results obtained are shown in table1.

Table1: Average Similarity for Jaccard, Cosine, Dice and Overlap Similarity Functions for Different Queries.

Query No.	Query Entered in Search Engine	Jaccard Similarity (A)	Cosine Similarity (B)	Dice Similarity (C)	Overlap Similarity (D)
Q1	Terrorist Attack Mumbai	0.3111	0.4280	0.4218	0.4863
Q2	Cloud Burst India	0.2277	0.3112	0.3085	0.3427
Q3	Moist Attack India	0.2443	0.3345	0.3262	0.3960
Q4	Corruption Cricket India	0.2906	0.4093	0.4047	0.4592
Q5	Pollution River Ganga	0.4493	0.5969	0.5914	0.6645
Q6	Power Generation India	0.2800	0.3823	0.3784	0.4269
Q7	Sand Mining India	0.3898	0.5210	0.5176	0.5675
Q8	Mid Day Meal India	0.3111	0.4278	0.4198	0.4949
Q9	Sikh Riots India	0.3536	0.4784	0.4763	0.5141
Q10	Moist Attack Train	0.3760	0.5116	0.5070	0.5627

Step 2: Analysis of the similarity functions based upon similarity scores

From the above table it is clear that the similarity scores of the Overlap similarity function outperforms the similarity scores obtained using the Cosine, Dice and Jaccard similarity functions. The cosine similarity outperforms the Dice and Jaccard similarity.

Step 3: Correlation Coefficient measurement for the similarity scores obtained using similarity functions

The linear associations between the similarity scores obtained using the four similarity functions is obtained using the correlation coefficient .Correlation Coefficient is a measure which measures of the strength of linear association between two variables. Correlation will always between -1.0 and +1.0. If the correlation is positive, a positive relationship is there and if it is negative, the relationship is negative. In this step of experiment the average Jaccard similarity is represented as A, average Cosine similarity is represented as B, average Dice similarity is represented as C and average Overlap similarity is represented as D. The general formula of the Correlation coefficient between the two scores i.e. A and B for N no. of values is given below.

$$\text{Correlation Coefficient} = [N\Sigma AB - (\Sigma A)(\Sigma B) / \text{Sqrt}([N\Sigma A^2 - (\Sigma A)^2][N\Sigma B^2 - (\Sigma B)^2])]$$

Where N = no. of values , A = First score, B= Second score

ΣAB = Sum of product of first and second scores

ΣA = Sum of first scores, ΣB = Sum of second scores

ΣA^2 = Sum of squares of first scores, ΣB^2 = Sum of squares of second scores

In the experiment the evaluation of the similarity scores using the different similarity functions i.e. Jaccard, Cosine, Dice, Overlap have been done by measuring the correlation coefficient [17].The results are summarized in table 2.

Table 2: Correlation Coefficient between Jaccard and Cosine, Jaccard and Dice, Jaccard and Overlap, Cosine and Dice, Cosine and Overlap, Dice and Overlap Similarity Functions

Correlation Between	Correlation Coefficient
A and B(Jaccard and Cosine)	0.974
A and C(Jaccard and Dice)	0.972
A and D(Jaccard and Overlap)	0.963
B and C(Cosine and Dice)	0.999
B and D(Cosine and Overlap)	0.992
C and D(Dice and Overlap)	0.988

Step 4: Exploring all the combinations of similarity functions.

In this step of experimentation all the possible combinations of four similarity functions have been explored .It was found that if two similarity functions are to be combined then six combinations are there i.e. Jaccard Cosine, Jaccard Dice, Jaccard Overlap, Cosine Dice, Cosine Overlap and Dice Overlap. If three similarity functions are to be combined then four combinations are there i.e. Jaccard Cosine Dice, Jaccard Cosine Overlap, Jaccard Dice Overlap and Cosine Dice Overlap. If all the four similarity functions are combined then only one combination is there i.e. Jaccard Cosine Dice Overlap combination.

Step 5: Performance evaluation of the similarity functions based upon the correlation coefficient.

It was proposed in [17] that if two similarity functions are combined then from the possible six combinations which are described in above step if we combine the the similarity scores of Cosine similarity(B), obtained using the Cosine similarity function and similarity scores of Overlap similarity(D), obtained using Overlap similarity function we got the highest average values than the average values of other combinations as shown in table 3. From the table 2 it is clear correlation coefficient between similarity scores of the Cosine and Dice is highest i.e. 0.999 and the correlation coefficient between similarity scores of Cosine and Overlap is 0.992. and correlation between similarity scores between Dice and Overlap is 0.988.In the proposed approach [17] , Cosine Overlap combination was chosen because average of scores of the Cosine and Overlap combination give the results which are in correlation with the other similarity scores using Cosine & Dice simlarity functions and similarity scores is more than Cosine and Dice individually.

Step 6: In this step other possible combimations which are described in step 4 are evaluated on the basis of correlation coefficient and a model for the textual similarity is proposed .

VI. Proposed Model of Textual Similarity Using Similarity Functions

Model of textual similarity is proposed for the information retrieval system in which all the possibilities of the combinations of four similarity functions have been explored.

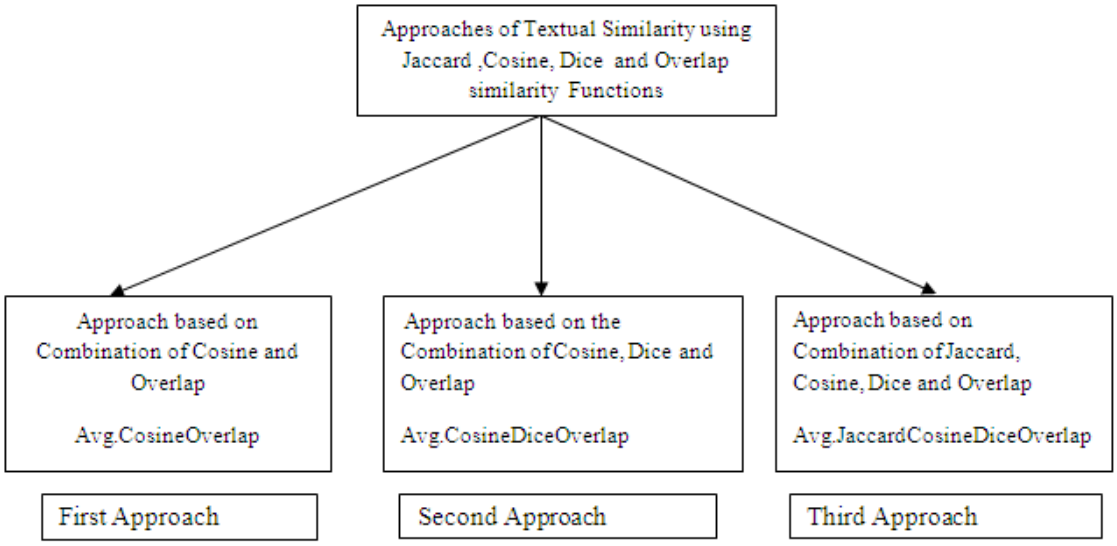


Figure 2 Three approaches for the textual similarity using Jaccard, Cosine, Dice and Overlap Similarity functions.

From the possible six combinations of two similarity functions it i.e. JaccardCosine, JaccardDice, JaccardOverlap, CosineDice, CosineOverlap and DiceOverlap, the best one is Avg.CosineOverlap combination. From the possible four combinations of three similarity functions i.e. JaccardCosineDice, JaccardCosineOverlap, JaccardDiceOverlap and CosineDiceOverlap the best one is Avg.CosineDiceOverlap.The last possible combination is of combination of four similarity functions i.e. Avg. JaccardCosineDiceOverlap. In the proposed model all the three approaches are explored.

(1) First approach based on the combination of Cosine and Overlap similarity functions (Avg. CosineOverlap):

It was proposed in [17] that on combining the similarity scores of Cosine similarity(B) and similarity scores of Overlap similarity(D) which is obtained using the Cosine and Overlap similarity functions , the highest average values was obtained than the average values of other combinations as shown in table 3 and figure 2.The results obtained are highly correlated with ths similarity scores of Cosine, Dice and Overlap similarity.

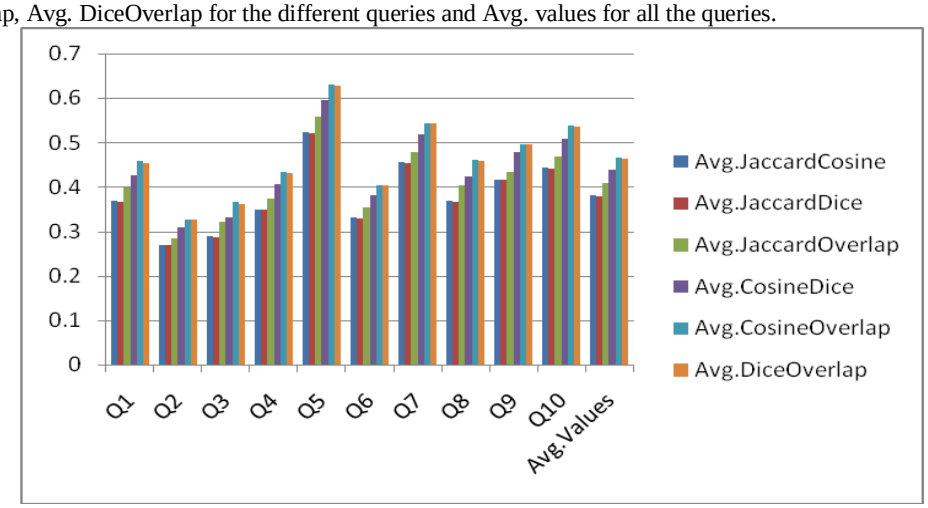
Evaluation of First Approach: On evaluation of scores of Jaccard similarity(A), scores of Cosine similarity(B), scores of Dice similarity(C) and scores of Overlap similarity(D) using Jaccard,Cosine, Dice and Overlap similarity functions respectively it was found from table 1 that the Overlap similarity outperforms the Cosine similarity, Dice similarity and Jaccard similarity but from table 2 it was found that the correlation coefficient between scores of Cosine similarity(B) and scores of Dice similarity(C) is highest i.e. 0.999 So we proposed our approach that on taking the average of similarity scores of Cosine similarity(B) and similarity

scores of Overlap similarity(D) which is obtained using the Cosine and Overlap similarity functions and obtained results shows that the highest average values for the said combinations than the average values of other combinations as shown in table 3[17] and fig. 3[17].The results obtained are correlated with ths similarity scores of Cosine, Dice and Overlap similarity.

Table 3: Average of JaccardCosine, JaccardDice, JaccardOverlap, CosineDice, CosineOverlap, DiceOverlap.

Query	**JaccardCosine (Avg. AB)**	**JaccardDice (Avg. AC)**	**JaccardOverlap (Avg.AD)**	**Cosine Dice (Avg. BC)**	**CosineOverlap (Avg. BD)**	**DiceOverlap (Avg.CD)**
Q1	0.36955	0.36645	0.3987	0.4249	0.45715	0.45405
Q2	0.26945	0.2681	0.2852	0.30985	0.32695	0.3256
Q3	0.2894	0.28525	0.32015	0.33035	0.36525	0.3611
Q4	0.34995	0.34765	0.3749	0.407	0.43425	0.43195
Q5	0.5231	0.52035	0.5569	0.59415	0.6307	0.62795
Q6	0.33115	0.3292	0.35345	0.38035	0.4046	0.40265
Q7	0.4554	0.4537	0.47865	0.5193	0.54425	0.54255
Q8	0.36945	0.36545	0.403	0.4238	0.46135	0.45735
Q9	0.416	0.41495	0.43385	0.47735	0.49625	0.4952
Q10	0.4438	0.4415	0.46935	0.5093	0.53715	0.53485
Avg. Value	0.381725	0.37926	0.407415	0.437635	0.46579	0.463325

Figure3. Values of Similarity for Avg. JaccardCosine, Avg. JaccardDice, Avg.JaccardOverlap, Avg. CosineDice, Avg. CosineOverlap, Avg. DiceOverlap for the different queries and Avg. values for all the queries.



(2) Second approach based on the combination of Cosine, Dice and Overlap similarity functions i.e. Avg. CosineDiceOverlap

From the table 2 it was found that the correlation coefficient is maximum between Cosine and Dice similarity scores i.e. 0.999 and it is 0.992 for the Cosine and Overlap and it 0.988 for Dice and Overlap. So from this evaluation of correlation coefficient we here proposed another approach that if we combine Cosine Dice Overlap then the results obtained are optimum. The results of the combination are shown in table 4. We have ignored the Jaccard Similarity function because from the table 2 it was found that the correlation coefficient between the Jaccard and Cosine Similarity scores was 0.974 and correlation coefficient between Jaccard and Dice similarity scores was 0.972 and it was 0.963 for the Jaccard and Overlap.

Table 4: Similarity scores of JaccardCosineDice, JaccardCosineOverlap, JaccardDiceOverlap and CosineDiceOverlap

Query	**Avg. JaccardCosineDice (Avg. ABC)**	**Avg. JaccardCosineOverlap (Avg. ABD)**	**(Avg. JaccardDiceOverlap (Avg. ACD)**	**(Avg. CosineDiceOverlap (Avg. BCD)**
Q1	0.386967	0.408467	0.4064	0.445367
Q2	0.282467	0.293867	0.292967	0.3208
Q3	0.301667	0.324933	0.322167	0.352233
Q4	0.3682	0.386367	0.384833	0.4244
Q5	0.545867	0.570233	0.5684	0.6176
Q6	0.3469	0.363067	0.361767	0.395867
Q7	0.476133	0.492767	0.491633	0.535367
Q8	0.386233	0.411267	0.4086	0.4475
Q9	0.4361	0.4487	0.448	0.4896
Q10	0.464867	0.483433	0.4819	0.5271
Avg. Value	0.3994	0.41831	0.4166667	0.455583

(3) Third approach based on the combination of Jaccard, Dice, Cosine and Overlap similarity functions i.e. Avg. JaccardCosineDiceOverlap:

In the last proposed approach the similarity scores of all the four similarity functions are combined using the four similarity functions i.e. Jaccard, Cosine, Dice and Overlap similarity functions and average is taken which is represented as Avg. JaccardCosineDiceOverlap and results obtained are shown in the table 5.

Table 5: Similarity scores using Avg. JaccardCosineDiceOverlap approach

Query	**Avg. JaccardCosineDiceOverlap (Avg. ABCD)**
Q1	0.4118
Q2	0.297525
Q3	0.32525
Q4	0.39095
Q5	0.575525
Q6	0.3669
Q7	0.498975
Q8	0.4134
Q9	0.4556
Q10	0.489325
Avg. Value	0.422525

Comparative analysis of the proposed design approaches of information retrieval techniques using four similarity functions:

Based on the above three proposed design approaches the experiment is repeated with the different queries i.e. Q1,Q2,Q3,Q4,Q5,Q6,Q7,Q8,Q9,Q10 and average of all the scores were taken to get the average values for the ten entered queries as shown in table 6. Three average values have been obtained from the three proposed design approaches from the different combinations of similarity functions.

Table 6: Similarity scores using the three proposed approaches

Avg. Values for ten queries(Q1,Q2.....Q10) using	**Results**
Avg. CosineOverlap approach	0.46579
Avg. CosineDiceOverlap approach	0.455583
Avg. JaccardCosineDiceOverlap approach	0.422525

Representation of proposed model:

These three avg. values represent the three vertices of a triangle in the proposed model for the textual similarity as shown in figure 4. In the proposed model R1, R2 and R3 are the vertices of triangle where R1 is result1 and it is the avg. value of CosineOverlap combination which is first approach in the proposed model, R2 which is result 2 and it is the avg. value of CosineDiceOverlap combination which is the second approach in the proposed model and R3 which is result 3 and it is the avg. value of JaccardCosineDiceOverlap combination which is the third approach in proposed model..

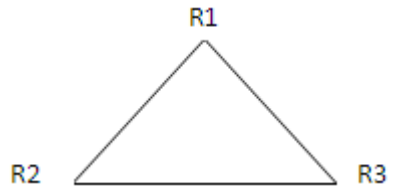


Figure 4: The proposed model of textual similarity using similarity functions.

VII. Conclusions

The model is proposed for the textual similarity between the documents retrieved for the entered query in the information retrieval system using the similarity functions in wide area networks.The model is based upon the correlation coefficient. While proposing the model for the matching mechanism for the information retrieval system for the textual similarity all the posible combinations of similarity functions were explored and it was found that there are sixteen possible combinations including empty set. On evaluation of Jaccard, Cosine, Dice and Overlap similarity functions it was found from the table 2 that correlation oefficient between the scores of similarity of Cosine & Dice is highest i.e 0.999 than the others. But from table 1 it is clear that the scores of similarity of Overlap similarity function outperforms the similarity scores of Cosine, Dice and Jaccard similarity function. From the table 3 it is concluded that first proposed approach of taking

the average of similarity scores of Cosine & Overlap combination using Cosine and Overlap similarity functions outperforms the avg. of other combinations and from the fig. 3 it is clear that the Avg. CosineOverlap combination give better results than average of other combinations of two similarity functions i.e. JaccardCosine, JaccardDice, JaccardOverlap, CosineDice, Dice Overlap. It is also concluded from the second proposed approach that avg. CosineDiceOverlap give the results better than the avg. of other combinations of three similarity functions i.e. JaccardCosineDice, JaccardCosineOverlap and JaccardDiceOverlap. The last approach combines the similarity scores of Jaccard, Cosine, Dice and Overlap similarity functions and average is taken. In the proposed model R1, R2 and R3 are the results of average value for all the said queries and are the results of three proposed approaches and represented by a triangle as shown in figure 4.

References

[1]. R. Baeza-Yates, B. Ribiero-Neto, Modern Information Retrieval, Pearson Education, 1999.

[2]. M.P.S. Bhatia, Akshi Kumar Khalid, "A Primer on the Web Information Retrieval Paradigm", Journal of Theoretical and Applied Information Technology, Vol.4, No.2, pp.657-662, 2008.

[3]. Nicholas J. Belkin, W. Bruce Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" Communications of the ACM, Vol.35, No.12, p29 (10), 1992.

[4]. Jaswinder Singh, Parvinder Singh, Yogesh Chaba, "A Study of Similarity Functions Used in Textual Information Retrieval in Wide Area Networks", International Journal of Computer Science and Information Technologies, Vol. 5, Issue 6, pp. 7880-7884,2014.

[5]. G. Salton and M.H. McGill, Introduction to Modern Information Retrieval, McGraw Hill, 1983.

[6]. Nicholas J. Belkin, W. Bruce Croft, "Retrieval Techniques," Annual Review of Information Science & Technology, M.E Williams, Ed. Chapter. 4, pp.109-145 Elsevier, 1987.

[7]. William P. Jones, George W.furnas, "Picture of Relevance: A Geometric Analysis of Similarity Measures," Journal of the American Society for Information Science, Vol.38, No.6, pp.420-442, 1987.

[8]. Siti Salwa Salleh, Noor Aznimah Abdul Aziz, Daud Mohamad and Megawati Omar, "Combining Mahalanobis and Jaccard Distance to Overcome Similarity Measurement Constriction on Geometrical Shapes," International Journal of Computer Science Issues, Vol. 9, Issue 4, pp. 124-132, 2012.

[9]. Jair Moura Duarte, Joao Bosco dos Santos and Leonardo Cunha Melo, "Comparison of Similarity Coefficients Based on RAPD Markers in the Common Bean," Genetics and Molecular Biology, Vol. 22, Issue 3, pp. 427-432, 1999.

[10]. P. Wallet, J. M. Barnard and G.M. Downs, "Chemical Similarity Searching," Journal of Chemical and Information and Computer Sciences, Vol. 38, No. 6, pp. 983-996, 1998.

[11]. McGill, Koll and Noreault, "An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems," Project report, Syracuse University, 1979.

[12]. Sung-Hyuk Cha, "Comprehensive Survey on the Distance/Similarity Measures between Probability Density Functions," International Journal of Mathematical Models and Methods in Applied Sciences, Vol. 1, Issue 4, pp. 300-307, 2007.

[13]. Wael H. Gomaa, Aly A. Fahmy, "A Survey of Text Similarity Approaches," International Journal of Computer Applications, Vol. 68, No. 13, pp. 13-18, 2013.

[14]. Suphakit Niwattanakul, Jatsada Singhthongchai, Ekkachai Naenudorn and Supachanun Wanapu, "Using of Jaccard Coefficient for Keywords Similarity," Proc. of International Multi Conference of Engineers and Computer Scientists, IMECS 2013, 2013, Hong Kong.

[15]. Wa'el Musa Hadi, Fadi Thabtah, Hussein Abdel-jaber, "A Comparative Study Using Vector Space Model with K-Nearest Neighbor on Text Categorization Data," Proc. of the World Congress on Engineering, WCE2007,2007,Vol.1, London, U.K.

[16]. Jaswinder Singh, Parvinder Singh, Yogesh Chaba, "Performance Modeling of Information Retrieval Techniques Using Similarity Functions in Wide Area Networks", International Journal of Advanced Research in Computer Science and Software Engineering,Vol.4, Issue12, pp.786-793, 2014.

[17]. Jaswinder Singh, Parvinder Singh,Yogesh Chaba, "Performance Evaluation and Design of Optimized Information Retrieval Techniques Using Similarity Functions in Wide Area Networks" , International Journal of Advanced Research in Computer Science and Software Engineering,Vol.5, Issue1, pp.461-469, 2015.