ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 6|| Jun – 2018 ||
International Journal of Computational Engineering Research (IJCER)

Impact of Similarity Measures in Information Retrieval

K. Pradeep Reddy1, T. Raghunadha Reddy2, G. Apparao Naidu3, B. Vishnu Vardhan4

[1]*Associate Professor, Dept of CSE, Tirumala Engineering College, Hyderabad*
[2]*Associate Professor, Dept of IT, Vardhaman College of Engineering, Hyderabad*
[3]*Professor, Dept of CSE, JBIET, Hyderabad*
[4]*Professor, Dept of CSE, JNTUHCEJ, Jagtiyal, Karimnagar*
Correspondence Author: K. Pradeep Reddy

ABSTRACT
The World Wide Web is growing exponentially with huge amount of textual content. There is a need of effective techniques to cluster the similar documents into groups and to retrieve most relevant documents to user queries. The information retrieval field mainly deals with the grouping of similar documents to retrieve required information to the user from huge amount of data. The researchers proposed different types of similarity measures and models in information retrieval to determine the similarity between the texts and for document clustering. In this work, the experimentation carried out with various similarity measures on different types of standard datasets. The main objective of this work is analyzing the influence of similarity measures to match a given query against a set of indexed documents.
KEYWORDS: Information Retrieval, Similarity Measures, Evaluation Measures, Standard Datasets

---

Date of Submission: 21-05-2018 Date of acceptance: 05-06-2018

I INTRODUCTION

The basic aim of information retrieval is retrieval of most relevant documents for a given user query. Web searches are the perfect example for this application. Many algorithms was developed for this purpose, which take an input query and match it with the stored documents or text snippets and rank the documents based on their similarity score respective to the given query. Such algorithms rely on matching the indexed documents, which maintain the information concerning term frequencies and positions, against the individual query terms. A score is assigned to each document based on its similarity value.

Different algorithms take different approaches in analyzing this similarity and computing the score. One of the highly rated and used approaches is the Vector Space Model. It trumps the Boolean Model, which takes Boolean queries and matches the document with the query solely based on Boolean logic, whether the required terms are in the document or not.

The Vector Space Model is a simple and the most popular model based on linear algebra allowing documents to be ranked based on their possible relevance. This model represents text objects as vectors in an n-dimensional space, where n represents the number of terms used to represent the document vector. The creation of an index requires the document to be striped and be segregated in the form of its unique terms. The document was further processed to reduce different forms of word into a common stem which helps to increase the efficiency when matching of query with documents. The weights also assigned to the terms of a query to increase their relative importance so as to give better results.

The Vector Space Model is based on linear algebra which was designed to overcome the limitations of the Boolean Model. One of the major advantages of VSM over the Boolean Model is that the weights assigned to the terms are not binary. This allows for better matching by computing over a range of similarity values and thus eliminating the too few or too many results obtained with the Boolean Model. On the other hand, VSM has its own limitations. The major limitation of VSM model was low sensitivity to semantics. For example, the words “car” and “automobile” won‟t give us a match, which is two identical phrases with one using the word “car‟ and the other using the word “automobile‟ will not give a match. Also, it was not distinguish phrases on the

basis of ordering of the constituent terms. For eg. "Mary is faster than John" is indistinguishable from "John is faster than Mary" using this approach.
This paper is organized in 6 sections. Section 2 describes the different types of usages of similarity measures in information retrieval field. The characteristics of dataset used in this experiment and evaluation measures used to evaluate the efficiency of the system is explained in section 3. Section 4 discussed the various similarity measures used in this experiment. The experimental results of similarity measures on standard datasets are analyzed in section 5. The conclusions of this work and future directions were described in section 6.

II LITERATURE REVIEW

Similarity measures define the similarity between two or more documents. The retrieved documents are ranked based on the similarity of content of document to the user query. Abhishek Jain et al., used [1] vector space model to represent the document vector and TFIDF measure was used to compute weights of the terms in the vector. They used jaccard and cosine similarity measures to calculate the similarity between the document vectors. Komal Maher et al., experimented [2] with different similarity measures for text clustering and classification. They experimented with three similarity measures such as Euclidian distance, cosine and similarity measures for text classification. They observed that similarity measure performance was good when compared with other similarity measures.
Moheb Ramzy Girgis et al., proposed [3] Genetic Algorithm based Information Retrieval algorithm which adjust the weights of the terms in a query to generate the optimal query vector. In this algorithm each query was represented as chromosome and these chromosomes were fed into the process of selection, crossover, mutation and similarity measures to find the best relevant documents. They experimented with different similarity measures such as cosine coefficient, inner product, dice coefficient, overlap coefficient and jaccard coefficient.
E man Al Mashagba et al described [4] different similarity measures such as dice, cosine, Jaccard etc in vector space model and compare each similarity measures using genetic algorithms approach based on various fitness functions, different mutations and crossovers were used to find the best solution of the given query. Mohammad Othman Nassaret et al analyzed [5] binary model using genetic algorithm with different fitness function and different mutation strategy to retrieving relevant information and query optimization. Poltak Sihombing et al described [6] Information retrieval system using genetic algorithm and various matching functions to compare the similarity between the user query and document database.
Anna Huang experimented [7] with different similarity measures and distance functions to cluster large amount of unordered text documents. They experimented on seven datasets and reported the results by using five similarity measures and K-means clustering algorithm. Lin Fu et al., used [8] different similarity measures to cluster the similar queries into different categories. They experimented with different types of approaches like content based, feedback based and result based approaches. Donald Metzler et al., exploited [9] various similarity measures to compute the similarity between two short texts in the information retrieval field. They experimented with 363822 queries from web search log. They used different text representations including surface, stemmed and expanded and lexical and probabilistic similarity measures were used to predict the similarity between the queries.

III DATASET CHARACTERISTICS AND EVALUATION MEASURES

Table 1 shows the dataset characteristics of various standard datasets used in information retrieval field.

Table 1. Dataset Characteristics

Dataset Name	Size of Dataset (MB's)	Number of Queries	Number of Documents
LISA	3.4	35	5872
NPL	3.1	93	11429
CACM	2.2	64	3204
CISI	2.2	112	1460
Cranfield	1.6	225	1400
Time	1.5	83	423
Medline	1.1	30	1033

In table 1, LISA stands for Library and Information Science Abstracts. The files in this directory contain the LISA collection as provided by PeterWillett of Sheffield University. The LISA documents contain just the title and abstract fields. The NPL collection is a collection document titles. The CACM collection is a collection of titles and abstracts from the journal CACM. CISI dataset contain a collection of 1460 documents. Cranfield dataset is a collection of abstracts and a set of queries with relevance judgments. The Time collection consists of articles from the magazine Time. Medline dataset contain a collection of articles from a medical journal.
The researchers used various evaluation measures such as Recall, precision, and F-measure to evaluate the performance of the information retrieval systems. Recall is the ratio of the number of documents retrieved

correctly to the total number of relevant documents in the document collection whereas precision is the ratio of the number of documents retrieved correctly to the total number of documents retrieved. F-measure is the standard measure for evaluating IR by combining recall and precision techniques. Precision and recall are used to show how many of the relevant documents are captured and missed by the proposed and the classical IR system for each query whereas the F-measure shows the overall performance of the system for each query by combining the recall and precision values. The harmonic F-measure gives equal weight for recall and precision. In this work, recall measure was used to evaluate our system.

IV SIMILARITY MEASURES

A similarity measure is a function which determines the degree of similarity between a pair of textual objects [10]. Similarity measures are very important for document clustering and Text-Mining [11]. In general, these measures were used to calculate similarity between two queries, two documents and one document and one query [12]. Similarity measures also used to rank the documents based on the similarity scores between the document and query [13]. In this work the experimentation carried out with six similarity measures. The next subsections explain the different similarity measures used in this work.

4.1 Euclidian Distance (ED)

Euclidian distance measure is an ordinary distance measure to compute the distance between two points in two and three dimensional space. This measure is used in document clustering to group the documents into similar clusters based on the distance between the documents [2]. Euclidian Distance measure is represented in equation (1).

$$ED(d_x, d_y) = \sqrt{\sum_{i=1}^{m} \left(w(t_i, \mathrm{d_x}) - w(t_i, \mathrm{d_y}) \right)^2} \quad (1)$$

ED(dx, dy) is the Euclidian Distance between dx and dy documents. (t1, t2, … , tm) is the set of terms, w(ti,dx), w(ti,dy) is the weights of term ti in document x and y respectively.

4.2 Cosine Similarity Measure (CSM)

In VSM, the sets of documents and queries are viewed as vectors. Cosine similarity measure is a popular method for calculating the similarity value between the vectors[3]. With document and queries being represented as vectors, similarity signifies the proximity between the two vectors. Cosine similarity measure computes similarity as a function of the angle made by the vectors. If two vectors are close, the angle formed between them would be small and if the two vectors are distant, the angle formed between them would be large. The cosine value varies from +1 to -1 for angles ranging from 0 to 180 degrees respectively, making it the ideal choice for these requirements. A score of 1 evaluates to the angle being 0o, which means the document are similar. While a score of 0 evaluates to the angle being 90o, which means the documents are entirely dissimilar.

The cosine weighting measure is implemented on length normalized vectors for making their weights comparable. Equation (2) gives the formula for Cosine Similarity.

$$CSIM(q, d_j) = \frac{\sum_{i=1}^{m} w(t_i, q) \times w(t_i, \mathrm{d_j})}{\sqrt{\sum_{i=1}^{m} w(ti, q)^2} \times \sqrt{\sum_{i=1}^{m} w(ti, \mathrm{d_j})^2}} \quad (2)$$

Where, w(ti,q), w(ti,dj) are the weights of the term ti in query q and document dj respectively.

4.3 Jaccard Similarity Measure (JSM)

Jaccard Similarity measure is another measure for calculating the similarity between the queries and documents [1]. In this measure, the index starts with a minimum value of 0 (completely dissimilar) and goes to a maximum value of 1 (completely similar).

Jaccard similarity measure measures similarity between the two documents. The value is between 0 and 1. 0 show that documents are dissimilar and 1 shows those documents are identical with each other. Value between 0 and 1 show the probability of similarity between the documents. Equation (3) represents the Jaccard Similarity measure.

$$JSIM(q,d_j) = \frac{\sum_{i=1}^{m} w(t_i,q) \times w(t_i,\mathrm{d_j})}{\sum_{i=1}^{m} w(ti,q)^2 + \sum_{i=1}^{m} w(ti,\mathrm{d_j})^2 - \sum_{i=1}^{m} w(t_i,q) \times w(t_i,\mathrm{d_j})} \quad (3)$$

Where, w(ti,q), w(ti,dj) are the weights of the term ti in query q and document dj respectively.

4.4 Dice Coefficient Measure (DCM)

Dice Coefficient measure is used to compare the similarity between two samples of text [4]. Equation (4) shows the Dice Coefficient measure.

$$DSIM(q,d_j) = \frac{\sum_{i=1}^{m} w(t_i,q) \times w(t_i,\mathrm{d_j})}{\alpha \sum_{i=1}^{m} w(ti,q)^2 + (1-\alpha) \sum_{i=1}^{m} w(ti,\mathrm{d_j})^2} \quad (4)$$

Where, w(ti,q), w(ti,dj) are the weights of the term ti in query q and document dj respectively. α Parameter range is from 0 to 1. α control the magnitude of penalties of false negative versus false positive errors. In general Alpha value is 0.5. if α> 0.5, DCM measure gives more significance to precision and if α< 0.5, this measure gives more significance to recall.

4.5 Pearson Correlation Coefficient (PCC)

Pearson Correlation Coefficient is used to compute the linear correlation among two documents dx and dy [7]. PCC measure gives a similarity value between -1 and +1, where -1 is total negative linear correlation, 0 is no linear correlation, and +1 is total positive linear correlation. Equation (5) is used to compute the Pearson Correlation Coefficient.

$$PCSIM(d_x,d_y) = \frac{m\sum_{i=1}^{m} w(ti,dx) \times w(ti,dy) - \sum_{i=1}^{m} w(ti,dx) \times \sum_{i=1}^{m} w(ti,dy)}{\sqrt{\left(m\sum_{i=1}^{m} w(t_i,d_x)^2 - \left(\sum_{i=1}^{m} w(t_i,d_x)\right)^2\right) \times \left(m\sum_{i=1}^{m} w(t_i,d_y)^2 - \left(\sum_{i=1}^{m} w(t_i,d_y)\right)^2\right)}} \quad (5)$$

Where, w(ti,dx), w(ti,dx) are the weights of the term ti in documents dx and dy respectively.

4.6 Averaged Kullback-Leibler Divergence (AKLD)

The Kullback–Leibler divergence measure computes how one probability distribution of terms in a document diverges from a second document probability distribution of terms [7]. Kullback–Leibler divergence measure value of 0 indicates that two documents distributions are similar, while a Kullback–Leibler divergence measure value of 1 indicates that the two probability distributions were different. Equation (7) is used to compute the Kullback–Leibler divergence measure

$$KLD(dx \,||\, dy) = \sum_{i=1}^{m} w(t_i,d_x) \times \log\left(\frac{w(t_i,d_x)}{w(t_i,d_y)}\right) \quad (6)$$

Equation (7) is used to compute the Averaged Kullback–Leibler divergence measure.

$$AVGKLD(d_x, d_y) = \sum_{i=1}^{m} (\pi_1 \times KLD(dx \,||\, M) + \pi_2 \times KLD(\mathrm{M} \,||\, d_y)) \quad (7)$$

Where,

$$\pi_1 = \sum_{i=1}^{m} \frac{w(t_i, d_x)}{w(t_i, d_x) + w(t_i, d_y)}$$

$$\pi_2 = \sum_{i=1}^{m} \frac{w(t_i, d_y)}{w(t_i, d_x) + w(t_i, d_y)}$$

V EXPERIMENTAL RESULTS

Table 2 shows the experimental results of similarity metrics on standard datasets by using recall evaluation measure.

Table 2. The accuracies of Recall measure

SIMILARITY MEASURE /DATASET	ED	CSM	JSM	DCM	PCC	AKLD
LISA	0.7060	0.7315	0.7680	0.7516	0.7789	0.8280
NPL	0.7170	0.7335	0.7715	0.7571	0.7732	0.8115
CACM	0.7105	0.7320	0.7705	0.7539	0.7833	0.8305
CISI	0.7075	0.7275	0.7700	0.7687	0.7885	0.8200
Cranfield	0.7070	0.7190	0.7710	0.7701	0.7944	0.8410
Time	0.7060	0.7195	0.7725	0.7795	0.8031	0.8315
Medline	0.7060	0.7235	0.7725	0.7841	0.8187	0.8425

In this work, the experimentation carried on seven standard datasets by using six similarity measures. Averaged Kullback–Leibler divergence measure obtained a good recall value of 0.8425 on Medline dataset. The performance of Averaged Kullback–Leibler divergence measure was good when compared with other similarity measures.

VI CONCLUSIONS AND FUTURE SCOPE

This paper gives a brief overview of a basic Information Retrieval model, VSM, with different similarity measures. This field has seen a lot of research in the past decade. This research was focused on the influence of similarity measures for finding more relevant documents for the given query. In this work, six similarity measures such as Euclidian distance, cosine similarity, jaccard similarity, dice coefficient, Pearson Correlation Coefficient and Averaged Kullback–Leibler divergence measures were used to compute the similarity between the queries and the documents. The experimentation was carried out on seven standard datasets in information retrieval field. The Averaged Kullback–Leibler divergence measure performance was good when compared with other similarity measures.

Thus, future work in this field should be focused on developing new models, weighting schemes and similarity measures that can perform effectively on large data sets utilizing semantic information on the same.

REFERENCES

[1]. Abhishek Jain, Aman Jain, Nihal Chauhan, Vikrant Singh, Narina Thakur , "Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model", International Journal of Computer Applications, Volume 164, No 6, PP.28-30, 2017.

[2]. Komal Maher, Madhuri S. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering", International Journal of Computer Science and Information Technologies, Vol. 7, No.4, pp.1715-1720, 2016.

[3]. Moheb Ramzy Girgis, Abdelmgeid Amin Aly & Fatima Mohy Eldin Azzam, "The Effect Of Similarity Measures On Genetic Algorithm-Based Information Retrieval", International Journal of Computer Science Engineering and Information Technology Research, Vol. 4, Issue 5, pp. 91-100, Oct 2014.

[4]. E man Al Mashagba , Feras Al Mashagba and Mohammad Othman Nassar, "Query optimization using genetic algorithm in the vector space model", International Journal of Computer Science, vol. 8, no. 3, pp.450-457, Sept. 2011.

[5]. Mohammad Othman Nassar, Feras Al Mashagba and Eman Al Mashagba, "Improving the user query for the Boolean model using genetic algorithm", International Journal of Computer Science, vol. 8, no. 1, pp. 66-70, Sept. 2011.

[6]. Poltak Sihombing, Abdullah Embong, Putra Sumari, "Comparison of document similarity in information retrieval system by different formulation", Proceedings of 2nd IMT-GT Regional Conference on Mathematics Statics and Application, Malaysia, Jun. 2006.

[7]. Anna Huang , "Similarity Measures for Text Document Clustering", proceedings of the New Zealand Computer Science Research Student Conference, Christchurch, New Zealand,pp.49-56, 2008.

[8]. Lin Fu, Dion Hoe-Lian Goh, Schubert Shou-Boon Foo, Jin-Cheon Na, "The Effect of Similarity Measures on The Quality of Query Clusters", ournal of Information Science, Vol.30, No.5, pp.396-407, 2004.
[9]. Donald Metzler, Susan Dumais, Christopher Meek, "Similarity Measures for Short Segments of Text", European Conference on Information Retrieval, Lecture Notes in Computer Science book series, volume 4425, pp 16-27, 2007.W.F. Ames. Numerical Methods for Partial Differential Equations. Academic Press, 1977.
[10]. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Survey on Author Profiling Techniques", International Journal of Applied Engineering Research, March 2016, Volume-11, Issue-5, pp. 3092-3102.
[11]. Raghunadha Reddy T, Vishnu Vardhan B, GopiChand M, Karunakar K, "Gender prediction in Author Profiling using ReliefF Feature Se-lection Algorithm", Proceedings in Advances in Intelligent Systems and Computing, Volume 695, PP. 169-176, 2018.
[12]. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Pro-file specific Document Weighted approach using a New Term Weighting Measure for Author Profiling ", International Journal of Intelligent Engineering and Systems, Nov 2016, 9 (4), pp. 136 - 146.
[13]. K. Pradeep Reddy, T. Raghunadha Reddy,G. Apparao Naidu, B. Vishnu Vardhan, " Term weight measures influence in information retrieval", International Journal of Engineering and Technology, Vol.7, Issue-2, pp. 832-836, May 2018.

K. Pradeep Reddy." Impact of Similarity Measures in Information Retrieval." International Journal of Computational Engineering Research (IJCER), vol. 08, no. 06, 2018, pp. 54-59.