IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-12 E-ISSN: 2347-2693

An Experimental Study of Recall and Precision Rates in Retrieval of Text Documents Using Different Distance Measures

U.S. Patki[1*], A.B. Kurhe[2] and P.G. Khot[3]

[1]*Dept. of Computer Science, NES Science College, Nanded, India
[2]Dept. of Computer Science, SGBS College Purna(Jn), India
[3]Dept. of Statistics, RSTM University, Nagpur, India

[*]*Corresponding Author: patkiulhas@gmail.com, +91 9860056449*

Available online at: www.ijcseonline.org

Received: 17/Nov/2017, Revised: 29/Nov/2017, Accepted: 15/Dec/2017, Published: 31/Dec/2017

***Abstract*—**Searching is the most important process in an information retrieval from available large databases. Many times we search for a set of documents which is relevant to the given search document. Text mining helps us to mine the information from a given set of documents and it is most popular technique in Information retrieval. In this research paper we have applied distinct distance measures for retrieval of most similar documents to the queried document from a set of given document. For obtaining optimality for required search, we have gone through pre-processing of documents, creating vector space model and used distance measure techniques. Precision and recall are the basic measures used in evaluating search strategies. We have presented five distance measure technique applied on hundred text documents from standard database 20NewsGroup and calculated Recall and precision rate for text documents retrieval. We have used MatLab 10a as a development tool for our experiment.

***Keywords*—** Text Mining, Information retrieval, distance measure, recall rate, precession rate, document.

I. INTRODUCTION

An explosive growth of Knowledge in the form of textual documents in almost every area of digital era needs an extensive demand for new powerful tools to filter the text documents and extract required knowledge from it. After finishing a search the nagging question in every searcher's mind is: "Have I found the most relevant material or am I missing important items?".Text mining is a research technology to discover useful knowledge from enormous collection of text documents and develop a system to provide this knowledge to support the user's decision. The text miner program gathers the relevant textual documents together, mines the information and converts this unstructured data into structured database. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases, but text mining can also work with unstructured or semi-structured data sets such as emails, text documents and HTML files etc. As a result, text mining is a best solution for discovering knowledge [1] .

The Text mining task will become easier if the documents containing information on similar topic are grouped together into a single class.

Automatically organizing text documents into meaningful clusters or groups is called Document Clustering [2]. For this grouping, we need to measure the similarity among these documents. The similarity between two objects might be calculated by comparing its attributes. For ex. Objects with green color can be grouped into a single class. Here we have considered color as an attribute. Documents can be grouped together by retrieving and matching their contents or key words which we further refer as features.

In our experiment, we have attempted to retrieve most similar documents to the query document. In this research paper we have presented our experimental result using six different distance measures and applied Recall and Precision technique for calculating accuracy. To represent the research work, we have split it in six sections. The first section is an introduction. The second section explains the pre-processing steps for extracting features of document. The third section includes the details of recall and precision ratio for document retrieval. The algorithm for vector space model and distance measures is discussed in the fourth section. The results obtained in the experiment are discussed in fifth section. Finally the sixth section summarize conclusion.

II. PRE-PROCESSING

Form the standard database 20newsgroup, we have used 100 documents. 20Newsgroup contains total 17 distinct folders which contain numbers of sample text documents, out of which we have used 100 documents, 50 from folder windows group and remaining 50 from folder religious. We have

chosen one document out of these 100 documents as a query document and find out recall and precision ratio.

Preprocessing involves following steps [3].

1. Removal of Stop words: Initially we have removed all the stop words from the documents. We have assumed more than 320 Stop words. Stop words are grammatical words that do not have much more importance in the document. Stop words are for ex. 'was', 'the', 'about' etc [3]. This representation of documents is called "Bag of Words" method.

2. Feature generation: These 100 documents are converted in the form of vector space model (VSM). Each row indicates a document number while each column indicates unique words and their frequency in each document. In this way we have obtained a matrix of size 100 X 4432 and called it as c_table

3. Feature Selection: Feature selection is performed by reducing size of VSM and this process is completed by applying two criteria as

a) Keeping minimum length of each word at least three characters.

b) The feature word must be occurred in at least two documents

Feature selection causes reduction in size of VSM and we are able to reduce it up to 100 X 3162. This matrix is referred as c1_table.

After these pre processing steps we have performed our experiment on this vector space model for calculating recall and precision ratio of a query document.

III. RECALL AND PRECISION RATIO

RECALL is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the database. It is usually computed using following equation [4].

$$Recall = \frac{Number\ of\ Relevant\ Documents\ Retrieved}{Number\ of\ Relevant\ Documents\ in\ Database} \quad \text{-- (1)}$$

PRECISION is the ratio of the number of relevant documents retrieved to the total number of irrelevant and relevant documents retrieved. It is usually computed using following equation.

$$precision = \frac{Number\ of\ \text{Re}levant\ Documents\ Retrieved}{Total\ Number\ of\ Documents\ Retrieved} \quad \text{-- (2)}$$

Distance Measures:

In order to measure the similarity among the data-items, a technique Distance metrics plays very significant role. It is necessary to identify, in what manner the data are inter-related, how various data dissimilar or similar with each other and what measures are considered for their comparison. The main purpose of metric calculation is to obtain an appropriate distance similarity function. A metric function or distance function is a function which defines a distance between objects of a set.

This distance metric technique also plays a very important role in clustering.[4] Literature review pointed that there are different distance measures techniques available which can be used for computing inter relationship or similarity among different objects.

In the present study, we have applied six different distance measure techniques namely Fuzzy c-means distance, Euclidian Distance, Cityblock, Minkowski, Cosine distance and correlation distance.

i) Fuzzy c-means distance:

In this module we measure the distance in fuzzy c-means clustering. For this distance we have used *distfcm* function of MatLab. This function calculates the Euclidean distance between each row in CENTER and each row in DATA, and returns a distance matrix and the membership value of each row.

ii) Euclidian Distance:

Euclidean distance is widely used in clustering problems, including clustering text. It is also the default distance measure used in the K-means algorithm.

To measure the distance between two text documents, following expression is used.

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m} (W_{ik} - W_{jk}} \quad \text{--- (3)}$$

Here x & y are set of terms in documents and W_{ik} and W_{jk} are weight terms.

iii) Minkowski Distance: [5]

Minkowski Distance is the generalized metric distance.

$$Dist_{xy} = \left(\sum_{k=1}^{d} \left(\left| X_{ik} - X_{kj} \right| \right)^{1/p} \right)^{p} \quad \text{--- (4)}$$

Note that when p=2, the distance becomes the Euclidean distance

iv) CityBlock Distance:

When the value of p is 1 in expression (4), it works as city block distance measure.

v) Cosine distance: [6]

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, which is nothing but the

cosine similarity. Cosine similarity is one of the most popular similarity measure applied to text documents which can be calculated with the expression 5.

$$Sim(t_a, t_b) = \frac{t_a . t}{|t_a| . |t_b|} \quad \text{--- (5)}$$

Where t_a and t_b are m-dimensional vectors over the term set $T = \{t_1, \ldots, t_m\}$. Each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between [0,1].

vi) Correlation distance:

One minus the sample linear correlation between observations (treated as sequences of values) is known as Correlation distance.

We have used 'pdist2' function in MatLab to calculate the above distances measures except FCM.

IV. ALGORITHM FOR VSM & DISTANCE MEASURES

For finding recall and precision values we have designed following algorithm.

Recall_Precision(N,Q)

//We have given N documents. Following algorithm builds a vector space model for these N Documents. Then it computes distance of query document Q from each of these N documents by applying different distance measures.//

1. Remove stop-words from all 100 documents N=100
2. Select feature words with character length > 2 // To obtain matrix known as *c_table*
3. Remove columns with single non-zero entry to form *c1_table* also known as Vector Space Matrix.
4. Select one document from these N documents as a query documents to search similar documents from document database.
5. Subtract the frequency of all words of query document from all documents similar words frequency.
6. Apply distance measures technique on Vector Space Matrix.
7. Derive the threshold value by calculating median of distances in step 6.
8. Retrieve all those documents which have less distance than threshold.
9. Calculate Precision and Recall by using equation (1) and (2).
10. Stop.

V. EXPERIMENT & RESULTS

By pre-processing, removal of stop-words, feature generation and feature selection steps, we calculated a matrix in which rows represents the number of documents and columns represents the selected features. These features are the words which exist in minimum two documents. By considering any one document from our database of hundred documents for searching its similarity with other documents, we calculate the distance measures by using all six methods and shown in table 1.1. The Graphs shows the distance of 100 different documents from threshold for each method

In this way we tried to reduce the semantic gap.

Results for Recall and Precision Ratio

DIST. MEASURE	THRESHOLD	RECALL	PRECISION
DISTFCM	34.65545	0.9	0.9
SEUCLIDIAN	57.7487	0.88	0.88
MINKOWSKI	34.65545	0.9	0.9
CITYBLOCK	517.5	0.9	0.9
COSINE	0.7912	0.76	0.76
CORELATION	0.81775	0.76	0.76

Table 1.1

The figures 1-6 show graphs along with details of each distances result. Horizontal blue line shows threshold value. We have used median of data as a threshold. We have also applied mode and average values for obtaining threshold but do not found better results.

i) Graphical representation of Distance Measure using FCM :

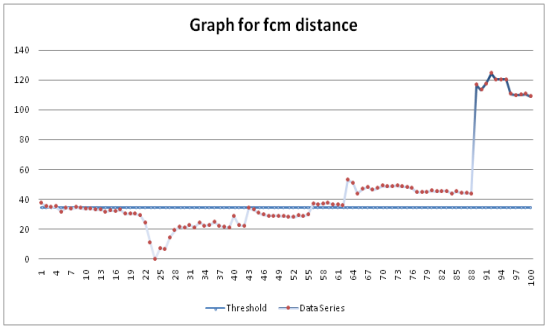

Figure. 1 FCM Distance

ii) Graphical representation of Distance Measure using Standard Euclidian Distance

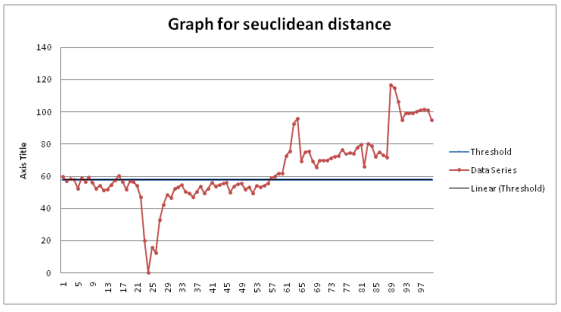

Figure. 2 Standard Euclidian Distance

iii) Graphical representation of Distance Measure using CityBlock :

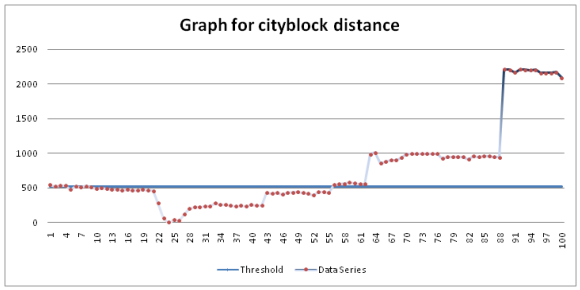

Figure. 3 City Block Distance

iv) Graphical representation of Distance Measure using Minkowski :

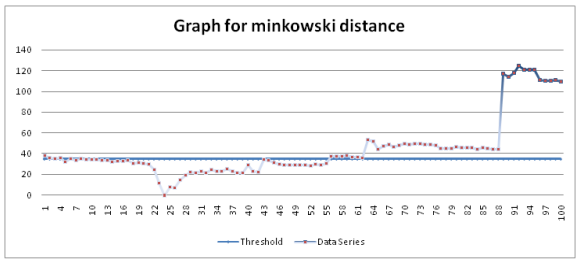

Figure. 4 Minkowski Distance

v) Graphical representation of Distance Measure using Cosine Distance :

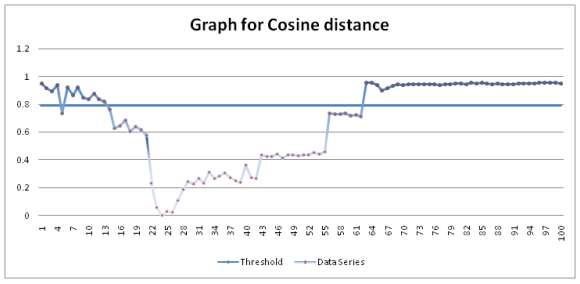

Figure. 5 Cosine Distance

vi) Graphical representation of Distance Measure using Correlation Distance:

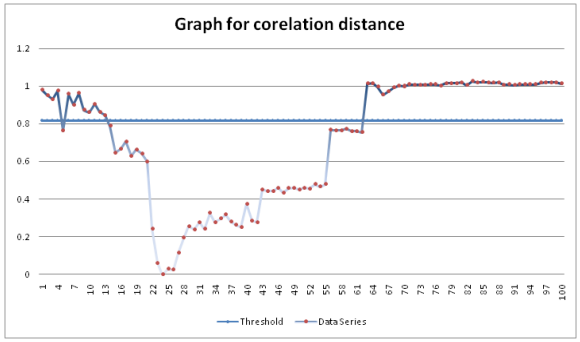

Figure. 6 Correlation Distance

Results:

The devised results of experiment shows that three methods distfcm, cityblock and Minkowski distances are best for recall and precision ratio of document retrieval as compare to the remaining methods standard Euclidean, cosine and correlation distances.

We found 0.9 as a precision-recall value for distfcm, cityblock and Minkowski distances, and which is the best as compare to remaining three methods. The table and graph shows 0.88 for standard Euclidean (seuclidian). If we see the distance measure values for cosine and correlation distances, which is 0.76 for both methods, that is this value is not better as compared to above distances.

VI. CONCLUSION

In the present study, we attempt to reduce the semantic gap to retrieve similar documents. The experimental result shown that out of six distance methods, The result of FCM, CITYYYBLOCK AND MINKOWSKI is 0.9 for both precision and recall rate. This is the optimum result for experiment. Standard Euclidean distance shows the precision and recall rate of 0.88. COSINE and CORRELATION shows the result 0.76 which is lower success rate as compare to previous results.

REFERENCES

[1] L. Kumar, et.al ,*"Text Mining: Concept, Process and Applications"* JGRCS,Vol-4,No.3, March 2013

[2] K Mugunthadevi, et.al, *"Survey on Feature Selection in Document Clustering"* IJCSE, Vol-3,3 March 2011,

[3] Sowmya P, et.al , *"Survey On Algorithms Used for Text Document Clustering",* IJAEC Special Issue September 2016

[4]A. Sudha Ramkumar et. al, *"Text Document Clustering Using Dimension Reduction Technique",* IJAER Vol -11, November 7, 2016,

[5] A. Singh, et.al ,"*K-means with Three different Distance Metrics",* IJCA, (0975 – 8887) Volume 67– No.10, April 2013

[6] A.Huang, *," Similarity Measures for Text Document Clustering",* NZCSRSC 2008, April 2008, Christchurch, New Zealand

[7] S. Goswami et.al, *"A Fuzzy Based Approach To Text Mining And Document Clustering"*2013,

[8] A Text Book *" Text Mining and Application Programming"* Manu Konchady ,Ed. 3 Indian Edition

Authors Profile

Mr. U.S. Patki Pursued Bachlor of Science and Master of Computer Science from Dr. B.A.M.U. Aurangabad. He has also completed his M.Phil in Computer Science from YCMOU Nashik (MS) He is working as a Asst. Professor in the department of Computer Science & IT in Science College Nanded (MS). He is Pursuing his Ph.D from Gondwana University Gadhchiroli. His Main research area is Text Mining, a branch of Data Mining

Dr. A.B. Kurhe Pursued Bachlor of Science and from Dr. B.A.M.U. Aurangabad and Master of Computer Science from SRTMU Nanded. He is awarded with Ph.D by SRTMU Nanded. He is working as a Asst. Professor in the department of Computer Science in SGBS College Purna (Jn) (MS). His Main research area is Image processing and recognition.

Dr. P.G. Khot is Ex-Professor and Head of department of Statistics of RSTM University Nagpur. He is a research guide in Computer Scinece at Gondwana University Gadhchiroli. Under his guidence 28 researchers have awarded Ph.D. from RSTM University Nagpur.

