Measuring Similarity between Text Documents for Information Retrieval

Geetanjali Gupta[1] Mr. Kapil Shah[2]
[1]P.G. Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science Engineering
[1,2]JIT Borawan Khargone, India

Abstract— There are several parameters by which similarity can be evaluated. The first categories of similarity evaluation is based on the document size and structure the length of the document, the number of paragraphs, number of sentences, average number of characters per word, average number of words per sentence etc. The second category is based on "style", whether the contents have been written in the first person conversational style or in the third person and so on. Thirdly, similarity can be based on the set of words used in the document. The fourth category of similarity is "content Similarity" which reflects to what extent the contents of the two documents are alike. This category is adopted throughout this thesis wherever similarity is talked of hereafter. The similarity between two documents is computed by any one of the several similarity measures based on the two corresponding feature vectors, e.g. cosine, dice, and jacquard measure. In this paper we measure similarity between texts documents using terms and token. A Document represented in a 3-Dimensional term vector space. There are several similarity coefficient are used to compare similarity. We used Euclidean distance to check the similarity.

Keywords: Document, Text, Similarity, Preprocessing, terms

I. Introduction

The data stored in databases is an example for structured datasets. The examples for semi structured and unstructured data sets include emails, full text documents and HTML files etc. Huge amount of data today are stored in text databases and not in structured databases. Text Mining is defined as the process of discovering hidden, useful and interesting pattern from unstructured text documents. Text Mining is also known as Intelligent Text Analysis or Knowledge Discovery in Text or Text Data Mining. Approximately 80% percent of the corporate data is in unstructured format. The information retrieval from unstructured text is very complex as it contains massive information which requires specific processing methods and algorithms to extract useful patterns. As the most likely form of storing information is text, text mining is considered to have a high value than that of data mining. Text mining is an interdisciplinary field which incorporates data mining, web mining, information retrieval, information extraction, computational linguistics and natural language processing.

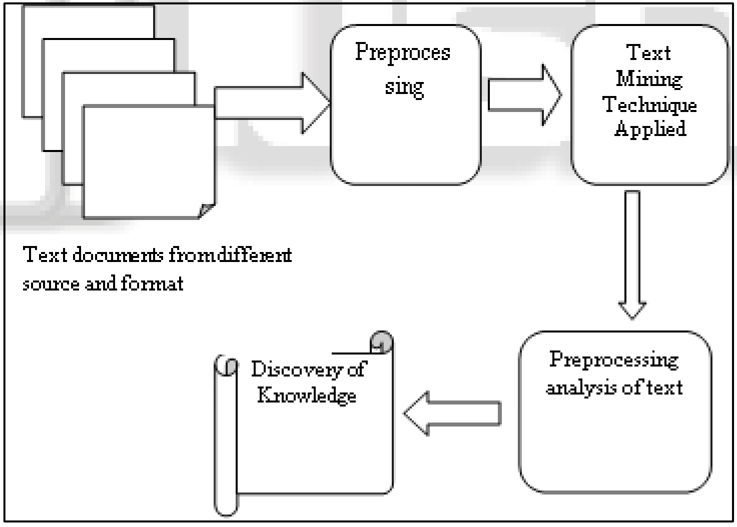
Preprocessing
Text Mining Technique Applied
Text documents from different source and format
Preprocessing analysis of text
Discovery of Knowledge

Fig. 1: Text mining process

Steps used in Text mining process

1) Collecting unstructured data from different sources Text mining interaction with other fields available in different file formats such as plain text, web pages, pdffiles etc.
2) Pre-processing and cleansing operations are performed to detect and remove anomalies. Cleansing process makes sure to capture the real essence of text available and is performed to remove stop words stemming and indexing the data.
3) Processing and controlling operations are applied to audit and further clean the data set by automatic processing.
4) Pattern analysis is implemented by Management Information System (MIS).
5) Information processed in the above steps are used to extract valuable and relevant information for effective and timely decision making and trend analysis.

II. Similarity Measures Coefficient

Utilization of similarity measures is not limited to clustering, but in fact plenty of data mining algorithms use similarity measures to some extent. To reveal the influence of various distance measures on data mining, researchers have done experimental studies in various fields and have compared and evaluated the results generated by different distance measures. Although itis not practical to introduce a "Best" similarity measure or a best performing measure in general, a comparison study could shed a light on the performance and behavior of measures.

A. *Minkowski*

The Minkowski family includes Euclidean distance and Manhattan distance, which areparticular cases of the Minkowski distance. The Minkowski distance performs well when the dataset clustersare isolated or compacted; if the dataset does not fulfil this condition, then the large-scale attributeswould dominate the others. Another problem with Minkowski metrics is that the

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

B. *Manhattan Distance*

Manhattan distance is a special case of the Minkowski distance at m = 1. Like its parent, Manhattan is sensitive to outliers. When this distance measure is used in clustering algorithms, the shape of clusters is hyper-rectangular.

$$d_{man} = \sum_{i=1}^{n} |x_i - y_i|$$

C. *Euclidean Distance*

The most well-known distance used for numerical data is probably the Euclidean distance. This is a special case of the Minkowski distance when m = 2. Euclidean distance performs well when deployed to datasets that include compact or isolated clusters. Although Euclidean distance is very common in clustering, it has a drawback: if two data vectors have no attribute values in common, they may have a smaller distance than the other pair of data vectors containing the same attribute values.

$$d_{edu} = (\sum_{i=1}^{n} |x_i - y_i|^m)^{1/m} \quad m \geq 1$$

D. *Average distance*

Regarding the above-mentioned drawback of Euclidean distance, average distance is a modified version of the Euclidean distance to improve the results. For two data points x, y in n dimensional space, the average distance.

$$d_{ave} = (\frac{1}{n}\sum_{i=1}^{n} (x_i - y_i)^2)^{\frac{1}{2}}$$

E. *Weighted Euclidean Distance*

If the relative importance according to each attribute is available, then the Weighted Euclidean distance another modification of Euclidean distance can be used. This distance measure is the only measure which is not included in this study for comparison since calculating the weights is closely related to the dataset and the aim of researcher for cluster analysis on the dataset. As an instance of using this measure reader can refer to Jiet. al. research work. They used this measure for proposing a dynamic fuzzy cluster algorithm

$$d_{we} = (\sum_{i=1}^{n} w_i (x_i - y_i)^2)^{\frac{1}{2}}$$

F. *Chord Distance*

Chord distance is one more Euclidean distance modification to overcome the previously mentioned Euclidean distance shortcomings. It can solve problems caused by the scale of measurements as well. Chord distance is defined as the length of the chord joining two normalized points within a hyper sphere of radius one. This distance can be calculated from non-normalized data as well.

$$d_{chord} = (2 - 2\frac{\sum_{i=1}^{n} x_i y_i}{\|x\|_2 \|y\|_2})^{\frac{1}{2}}$$

G. *Mahalanobis Distance*

Mahalanobis distance is a data-driven measure in contrast to Euclidean and Manhattan distances that are independent of the related dataset to which two data points belong. Aregularized Mahalanobis distance can be used for extracting hyper ellipsoidal clusters. On the other hand, Mahalanobis distance can alleviated distortion caused by linear correlation among features by applying a whitening transformation to the data or by using the squared Mahalanobis distance.

$$d_{mah} = \sqrt{(-x - y)S^{-1}(x - y)^T}$$

H. *Cosine Distance*

The Cosine similarity measure is mostly used in document similarity. The Cosine measure is invariant to rotation but is variant to linear transformations. It is also independent of vector length.

$$\text{Cosine}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\|x\|_2 \|y\|_2}$$

III. LITERATURE SURVEY

In 2020 Syed Fawad Hussain "A New Co-similarity Measure: Application to Text Mining and Bioinformatics". They explore two applications of our co-similarity measure. In text mining, document similarity is calculated based on word similarity, which in turn is calculated on the basis of document similarity. They capture the similarity between documents coming from their common words but also the similarity coming from words that are not directly shared by the two documents but that can be considered to be similar. They proposed method to extract gene clusters that show similar expression levels under a given condition from several cancer datasets[1].

In 2011 Wen-tau Yih Kristina Toutanova "Learning Discriminative Projections for Text Similarity Measures". Traditional text similarity measures consider each term similar only to itself and do not model semantic relatedness of terms. They propose a novel discriminative training method that projects the raw term vectors into a common, low-dimensional vector space. Proposed approach operates by finding the optimal matrix to minimize the loss of the pre-selected similarity function (e.g., cosine) of the projected vectors, and is able to efficiently handle a large number of training examples in the high dimensional space[2].

In 2012 Anna Rozeva "Classification of text documents supervised by domain ontologies". The research objective is to establish an approach for supporting the classification of text documents referring to a specified domain. The focus is on the preliminary topic assignment to the documents used for training the model. The method implements domain ontology as background knowledge. The idea consists in extracting the preliminary topics for training the classifier by means of unsupervised machine learning on a text corpus and further alignment of the document vectors to concepts of the ontology[3].

In 2013 Wael H. Gomaa Aly A. Fahmy "A Survey of Text Similarity Approaches". Measuring the similarity between words, sentences, paragraphs and documents is an important component in various tasks such as information retrieval, document clustering, word-sense disambiguation, automatic essay scoring, short answer grading, machine translation and text summarization. They discusses the existing works on text similarity through partitioning them into three approaches; String-based, Corpus-based and Knowledge-based similarities. They survey three text similarity approaches were discussed; String-based, Corpus-based and Knowledge-based similarities [4].

In 2014 Muhammad Shoaib1, Ali Daud2 and Malik Sikandar Hayat Khiyal "An Improved Similarity Measure for Text Documents". In text mining applications such as clustering documents, citation matching and author name disambiguation (AND) similar documents are grouped together by estimating similarity among them in pair wise fashion. They propose an improved similarity measure specially designed for matching terms of two textual documents in pair wise fashion. Proposed similarity measure tries to depict the picture of the proportion of similarity between the documents. It needs not any information about collection of documents as it is required in vector space based similarity functions. It uses simple count of term frequency as term weights [5].

In 2015 Daniel Bär "Composing Measures for Computing Text Similarity" They present a comprehensive study of computing similarity between texts. They start from the observation that while the concept of similarity is well grounded in psychology, text similarity is much less well-defined in the natural language processing community. They define the notion of text similarity and distinguish it from related tasks such as textual entailment and near-duplicate detection. They identify multiple text dimensions, provide empirical evidence. They discuss state-of-the-art text similarity measures previously proposed in the literature, before continuing with a thorough discussion of common evaluation metrics and datasets [6].

In 2016 Sumayia Al-Anazi, Hind AlMahmoud, "Finding similar documents using different clustering techniques" Text clustering is an important application of data mining. It is concerned with grouping similar text documents together. They discussed several models are built to cluster capstone project documents using three clustering techniques: k-means, k-means fast, and k-medoids. Data set is obtained from the library of the College of Computer and Information Sciences, King Saud University, Riyadh. Three similarity measures are tested: cosine similarity, Jaccard similarity, and Correlation Coefficient. The quality of the obtained models is evaluated and compared. The results indicate that the best performance is achieved using k- means and k-medoids combined with cosine similarity[7].

In 2017 Rasmus Hallen "A Study of Gradient-Based Algorithms". Gradient-based algorithms are popular when solving unconstrained optimization problems. By exploiting knowledge of the gradient of the objective function to optimize, each iteration of a gradient-based algorithm aims at approaching the minimize of said function. In the age of web-scale prediction problems, many venerable algorithms may encounter difficulties. They compare the performance of two different gradient-based algorithms; Gradient Descent (GD) and Stochastic Gradient Descent (SGD)[8].

In 2018 Marzieh Oghbaie and Morteza Mohammadi "Pairwise document similarity measure based on present term set". They introduces a novel text document similarity measure based on the term weights and the number of terms appeared in at least one of the two documents. The performance of our measure is compared with that of some popular measures. They present three different methods that not only focus on the text's words but also incorporate semantic information of texts in their feature vector and computes semantic similarities.

In 2019 Pinky Sitikhu, Kritish Pahi "A Comparison of Semantic Similarity Methods for Maximum Human Interpretability". They present three different methods that not only focus on the text's words but also incorporate semantic information of texts in their feature vector and computes semantic similarities. These methods are based on corpus-based and knowledge-based methods, which are: cosine similarity using tf-idf vectors, cosine similarity using word embedding and soft cosine similarity using word embedding. Among these three, cosine similarity using tfidf vectors performed best in finding similarities between short news texts. The similar texts given by the method are easy to interpret and can be used directly in other information retrieval applications.

IV. Problem Statement

The main problem is to calculate the distance between the different documents. For calculating distance between different documents we used Euclidean distance, Murkowski distance, Manhattan Distance. So we have to discover out which of the distance gives output (that is the recommendation of the items) in least time and efficiently. And what are the advantages and disadvantages of different similarity measure.

V. Objectives

There are several algorithms and methods have been text document clustering. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency. Our major objective are-

1) Apply Euclidean distance measures for distance calculation for one title with another title using terms only.
2) Apply Euclidean distance measures for distance calculation for one title with another title using terms and tokens.
3) Find out which distance method is more accurate.

VI. Proposed Approach

1) Step1: Initially the load number of documents in data base

2) Step2: Preprocess the documents removing the stop words and other words which are keywords
3) Step3: After finding keywords in each document calculates frequency of each word.
4) Step4: Now Calculate total number of words present in the each documents
5) Step5: Compute Euclidean distance between document with terms frequency and with terms frequency and total size of the documents
6) Step6: Finally compare the results of proposed measure with existing measures

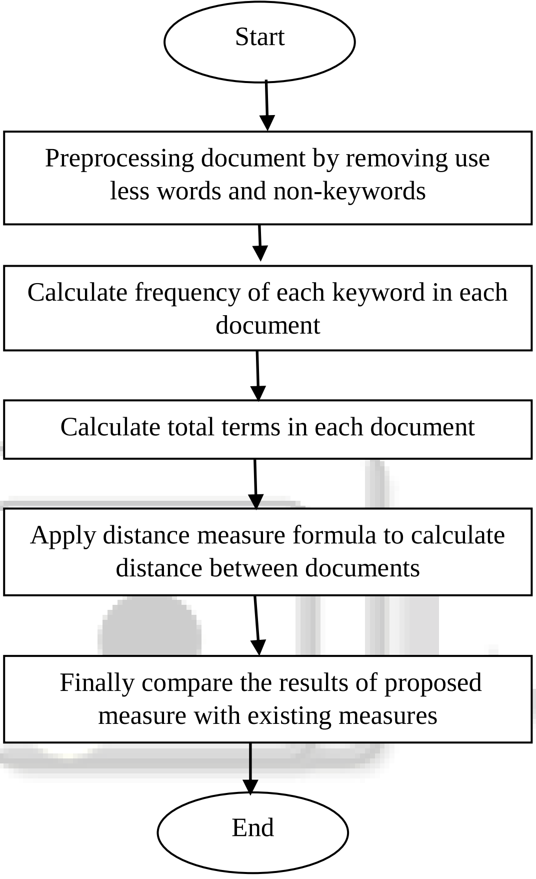
Start
Preprocessing document by removing use less words and non-keywords
Calculate frequency of each keyword in each document
Calculate total terms in each document
Apply distance measure formula to calculate distance between documents
Finally compare the results of proposed measure with existing measures
End

Fig. 2: working process of proposed approach

Distance formula satisfy the following mathematical properties:

1) Non-negativity: d (i, j)≥0: Distance is a non-negative number.
2) Identity of indiscernible: d (i, i)=0: The distance of an object to itself is 0.
3) Symmetry: d (i,j)=d(j, i): Distance is a symmetric function.

1) Generalized form of distance

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i1} - x_{j1}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

This chapter, we evaluate the performance of proposed algorithm and compare it with terms only based approach. The experiments were performed on Intel Core i3processor 1GB main memory and RAM: 4GB Inbuilt HDD: 400GB OS: Windows7. The algorithms are implemented in using Dot Net Framework language version 4.0.1. Synthetic datasets are used to evaluate the performance of the algorithms

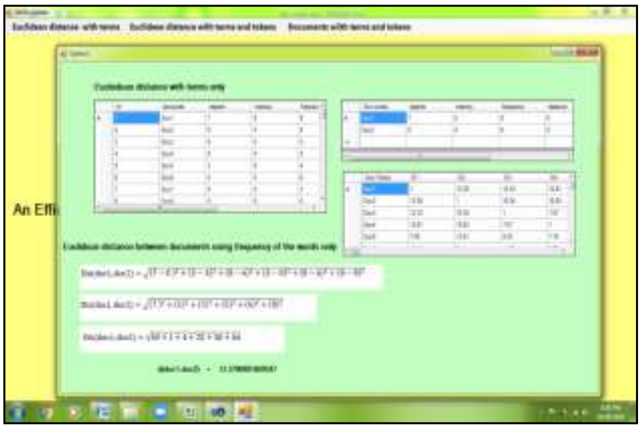

Fig. 3: implementation of proposed approach

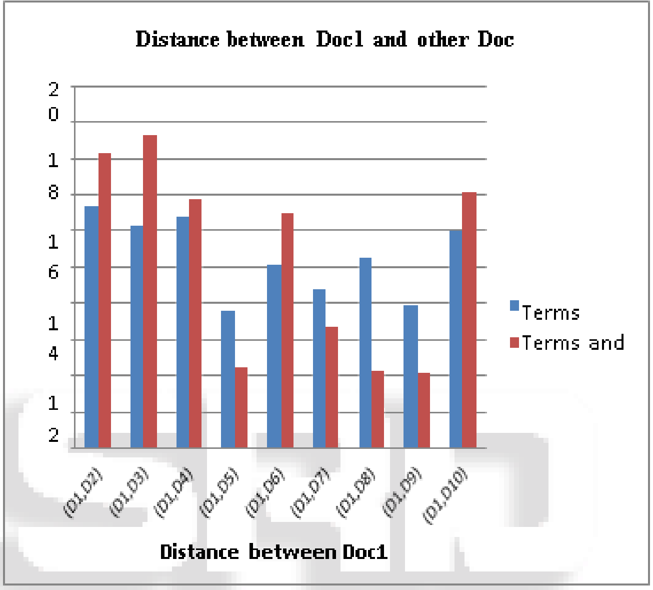

Fig. 4: Comparison between document similarities

VII. Conclusion and Future Work

Text mining is a burgeoning new field that attempts to glean meaningful information from natural language text. Distance between different texts documents can be calculated using Euclidean distance with terms only and Euclidean distance with terms and tokens. So discover which of the distance gives more correct output is difficult. Euclidean distance with terms only is more correct as compare to Euclidean distance with terms and token. In the paper we compare these two approaches to find the correct distance between two text documents. By the experimental analysis we calculate the distance between 10 documents.

References

[1] Syed Fawad Hussain A New Co-similarity Measure: Application to Text Mining and Bioinformatics HAL Id: tel-00525366 https://tel.archives-ouvertes.fr/tel-00525366 Submitted on 11 Oct 2010.

[2] Wen-tau Yih Kristina Toutanova Learning Discriminative Projections for Text Similarity Measures Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 247–256.

[3] Portland, Oregon, USA, 23–24 June 2011. c 2011 Association for Computational Linguistics.

[4] Anna Rozeva "Classification of text documents supervised by domain ontologies" ATI - Applied Technologies & Innovations Volume 8 | Issue 3 | November 2012.

[5] Wael H. Gomaa Aly A. Fahmy "A Survey of Text Similarity Approaches International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013.

[6] Muhammad Shoaib1, Ali Daud2 "An Improved Similarity Measure for Text Documents" Journal of Basic and Applied Scientific Research www.textroad.com.

[7] Daniel Bär "Composing Measures for Computing Text Similarity" Technical Report TUD-CS-2015-0017 January, 2015

[8] Sumayia Al-Anazi, Hind Al Mahmoud, Isra "Al-Turaiki Finding similar documents using different clustering techniques". Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

[9] Rasmus Hallen A Study of Gradient-Based Algorithms IBM Journal of Research and Development.

[10] Marzieh Oghbaie and Morteza Mohammadi "Pairwise document similarity measure based on present term set Oghbaie and Mohammadi Zanjireh J Big Data (2018) https://doi.org/10.1186/s40537-018-0163-2

[11] Pinky Sitikhu, Kritish Pahi and Pujan Thapa "A Comparison of Semantic Similarity Methods" for Maximum Human Interpretability arXiv:1910.09129v2 [cs.IR] 31 Oct 2019.