Similarity Measure Approaches Applied in Text Document Clustering for Information Retrieval

Naveen Kumar
Computer Science and Information Technology
Sam Higgin Bottom University of Agriculture , Technology and Science,
Allahabad, India
naveen.it23@gmail.com

Sanjay Kumar Yadav
Computer Science and Information Technology
Sam Higgin Bottom University of Agriculture , Technology and Science,
Allahabad, India
yadav_sk@rediffmail.com

Divakar Singh Yadav
Computer Science and Engineering
Institute of Engineering and Technology,
Lucknow, India
dsyadav@ietlucknow.ac.in

2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) | 978-1-7281-7132-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/PDGC50313.2020.9315851

***Abstract*— In today's world with ever increasing amount of text assets overloaded on web with digitized libraries, sorting out these documents got developed into a feasible need. Document clustering is an important procedure which consequently sorts out huge number of articles into a modest number of balanced gatherings. Document clustering is making groups of similar documents into number of clusters such that documents within the same group with high similarity values among one another and dissimilar to documents from other clusters. Common applications of document Clustering includes grouping similar news articles, analysis of customer feedback, text mining, duplicate content detection, finding similar documents, search optimization and many more. This lead to utilization of these documents for finding required information in a competent and efficient manner. Document clustering required a measurement for evaluating how surprising two given information are. This dissimilarity is often estimated by using some distance measures, for example, Cosine Similarity, Euclidean distance, etc. In our work, we evaluated and analyzed how effective these measures are in partitioned clustering for text document datasets. In our experiments we have used standard K-means algorithm and our results details on six text documents datasets and five most commonly used distance or similarity measures in text clustering.**

Keywords*— *Document Clustering, information retrieval, Similarity measures, text clustering, partitioned clustering.

I. INTRODUCTION

We are confronting a regularly expanding amount of text documents. The profuse texts streaming in excess of the Internet, massive collection of text in electronic libraries and stores, and individual data stored digitally. For example, blog articles and text messages are getting accumulated rapidly and consistently. That has brought difficulties for the successful and effective relationship of text documents. Text document clustering alongside its reasonable need has end up being a propelling exploration issue pulling in much work in the area of data recovery [1, 2]. It focuses on blend comparative reports in a single class and isolates this group however much as could reasonably be expected from the ones which contain data on altogether various themes. Internet has enormous utilizations of text clustering. From clustering of outcome for clients on web indexes to gathering of remarks to propose items on online supplies, the method has enormous shortest relevance.

A significant analysis that rapidly emerges within the undertaking is to indicate which parts of documents choose their nearness. Distance measures, like Jaccard coefficient and Cosine similarity have been proposed and broadly utilized for this reason and have end up being proficient. While examining, we have done an experimental investigation for five similarity measures in particular Cosine Similarity, Euclidean Measure, Pearson Correlation Coefficient, Mahalanobis Distance, and Jaccard Coefficient utilizing k-means clustering algorithm. For having steady ends, probed six datasets with various qualities. These incorporate paper articles, inquire about articles, web pages and so on [4, 6]. They accompany marks for classifications that is significant shortly in assessments while we attempt to quantify nature of got clusters dependent on purity and entropy esteems.

II. DOCUMENT REPRESENTATION

To mine large collection of documents it is necessary to pre-process them. As most of the text mining algorithms works on numbers, we used the bucket of terms representation in our job to preprocess the text wherein the documents are model as the bucket of exclusive terms along with a count of total occurrences of these words also called vector of words/term. The order and structure of words is completely discarded. All the documents get converted into a n-dimensional matrix where each row corresponds to a word and each column represents a document vector. This indicates the occurrence of word existence in an exacting text. There is not very complex to observe that the matrix would be a sparse one and we will make utilize of this information for the duration of clustering. The occurrence of each word as its weight which is represent of our assumption that an elevated frequency word is extra expressive of the text. So this though requires a small revision which is bringing in the representation of inverse document frequency value (IDF) [3, 7].

Consider the documents list where D indicates documents collection set and individual documents named as d_1, d_2, d_3, till d_n, where n is number of documents and term represented by T which has a series of terms like as t_1, t_2, t_3, till t_m, where m is exclusive terms. After this task, document can be expressed as: $t_d = t_{f1}$, t_{f2}, t_{f3}, till t_{fm}, where m is dimensional-vector. It is important that document pre-processing task is required in which we take out high appearance of stop words in all documents like an, for, in, is/are/am etc. We also specify the weight-values of words to terms not simply based on their frequencies but on TF-IDF weight-values. TF-IDF weights discounts the frequency of

terms based on their significance to specific text in the complete text put in concern. This is explain as follows:

$$\text{TF-IDF (d, t)} = \text{TF(d, t)} \times \log(\frac{|D|}{DF(t)})$$

Here DF(t) = quantity of documents (term appears)

A. Vector Space Model

Vector Space Model also called as Term Vector Model or Vector Processing Model that is represent both document and query by term set and compare similarities between them. This vector space model (VSM) [11] a standard technique in Information retrieval, is a way of representing documents through the words they contain. In this each document is N-dimensional vector where N is the number of distinct terms over all the documents. The weight related with every keyword/term find outs the significance of the giving keyword in any document. Therefore a text in vector type can be shown as, Dj = [w1j, w2j, w3j, w4j..., wnj] and wij is represented as weight of keyword i in text j.

B. TF-IDF

In TF-IDF, the term frequency (TF) is the frequency of the term in the current document and find out by following formula:

$$TF = \frac{\text{Number of times of appearance of term in a document}}{\text{Total number of terms in the document}}$$

In the inverse document frequency term weighting the higher weights is assigned to these more discriminative words. The total number of documents in a collection by N and n_i is the no. of documents where term i appears, the inverse document frequency of a term t is defined as by fraction of N and ni by following formula:

$$IDF = \log \frac{N}{n_i}$$

The final calculation of TF-IDF is collective process [11] which is represented as following formula:

$$TF - IDF = Tf * IDF [5]$$

III. SIMILARITY MEASURES

Several methods are available o find out the similarity between user's input as per query and recovered text. Similarity measure must be found out before clustering. This measurement shows the degree of nearness or separation of output object. And it should correspond to its characteristics that are differentiating the cluster which is embedding in data set. All most cases, given characteristic depend on the data and there is no solution that is globally good for all type of clustering difficulties [8]. It is most important to choose appropriate similarity measure for cluster analysis, mainly for a special kind of clustering algorithm. Basically, similarity measure or distance measures.

Generally, the distance between symbolic descriptions of two entities into single numeric value which is map by the similarity/distance measures. And it is depends on two parts, one of them characteristics of the two objects and other is measure itself. Distance measure is necessary for clustering definition which allocates the numeric value to degree of dissimilarity between two documents and on the given dataset, for making different clusters by which clustering algorithm used. The single distance measure is not good for all situations. And selection of distance measure depends on which distance measure is good for concentrate of important differentiating character for given data set. It is very important to know that which similarity is used. Because similarity measure is useful while deal with certain type of clustering algorithm. In this paper, discuss five measures briefly [9].

A. Eaclidean Distance

When we compute the measurement for mathematical issues used of Euclidean distance. This is the conventional space between two focuses and also can be simply calculated with a ruler in a few dimensional spaces. Euclidean-distance is generally implemented in clustering process which is for text clustering. K-means clustering algorithm is an extension with addition of default distance determined. If need to find out the distance between two given text documents, consider the documents d_a and d_b which has their term vectors also i.e $\vec{t}_a$ and $\vec{t}_b$ sequentially, the Euclidean-distance of the two documents represented as:

$$DE(\vec{t}_a, \vec{t}_b) = (\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2)^{1/2}$$

like

When the term position is $T = \{t_1, \ldots, t_m\}$. Now used the *Tf-IDF* for term weights whereas $wt,a = TFIDF(da, t)$.

B. Cosine Similarity

In this method documents represented as term vectors; the comparability of two documents relates to the relationship of the vectors. This is measured as the cosine of the edge between vectors that becomes the cosine similarity. It is most popular method for similarity measure that is used in text documents, as in frequent data recovery applications and clustering [9, 17].

Here two documents $\vec{t}_a$ and $\vec{t}_b$ vectors and cosine similarity as:

$$Similarity_C(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a . \vec{t}_b}{|\vec{t}_a| \times |\vec{t}_b|}$$

Where t_a and t_b are multi-dimensional vectors that used in term set $T = \{t_1, t_2 \text{ till to } t_m\}$. Every dimension represents a term covered by weight in the document that is positive. At the last the cosine similarity will be always positive and bounded between 0 and 1. A significant property of the cosine similarity method is autonomy of records length. When two documents are regarded be the same due to the cosine similarity between *d* and *d'* is 1 where same copies of a document *d* to find a new pseudo document *d'*. For now, given another document *l*, *d* and *d'* will have the same similarity value to *l*, that is,

$similarilty(\vec{t}_d, \vec{t}_l) = similarilty(\vec{t}_{d'}, \vec{t}_l)$.

In other words, documents with the same composition but different totals will be treating identically. Exactingly speaking, this does not assure the second condition of a metric, for the reason that after all the grouping of two copies is a dissimilar object from the original document [15, 16]. On the other hand, when the term vectors are normalized to a unit length such as 1, and consider this case the representation of *d* and *d'* is the same.

C. Mahalanobis Distance

The Mahalanobis distance is a proportion of the separation between a point P and a circulation D. It is a multiple dimensional speculation of estimating what number of standard deviations (std) away P is from the mean of D (MoD). This separation is 0 if P is at the mean of D (MoD), and develops as P moves from the mean along each principal component axis. If each of these axes is rescaled to have component variation, at that time the Mahalanobis distance corresponds to standard Euclidean distance in the transformed space [13].

The calculation contrasts from Euclidean distance in that it considers the connections of the informational collection and is scale-invariant. The formal appearance as:

$$d^2_{st} = (x_s - x_t)C^{-1}(x_s - x_t)'$$

D. Jaccard Coefficient

The Jaccard coefficient, which is in some cases alluded to as the Tanimoto coefficient, measures comparability as the crossing point isolated by the association of the items. In the text format, the Jaccard coefficient thinks about the entirety weight of shared terms to the whole weight of terms that are available in both of the two documents yet are not the common terms. It is explain by following:

$$SIMILARITY_j(\vec{t_a}, \vec{t_b}) = \frac{\vec{t}_a.\vec{t_b}}{|t_a|^2+|t_b|^2-\vec{t}_a.\vec{t_b}}$$

The is also a similarity measure and limit between [0,1]. 0 means where $\vec{t}_a$ and $\vec{t}_b$ are put out of joint, 1 means when the $\vec{t}_a = \vec{t}_b$.

E. Pearson Correlation Coefficient

Pearson's correlation coefficient is a different evaluate of the amount for used when 2-vectors are associated. There are various types of the Pearson connection coefficient equation. Specified the expression is $T = \{t_1, \ldots, t_m\}$, a frequently equation is:

$$SIMILARITY_P(\vec{t_a}, \vec{t_b}) = \frac{m\sum_{t=1}^{m} w_{t,a} \times w_{t,b} - TF_a \times TF_b}{\sqrt{\left[m\sum_{t=1}^{m} w_{t,a}^2 - TF_a^2\right][m\sum_{t=1}^{m} w_{t,b}^2 - TF_b^2]}}$$

Where $TFa = \sum_{t=1}^{m} w_{t,a}$ and $TF_b = \sum_{t=1}^{m} w_{t,b}$.

However, different the other events, that is ranges between +1 to -1. In successive experiments we use the equivalent distance measure.

IV. CLUSTERING ALGORITHM

For every later experiment, k-means clustering algorithm is used for the clusters forming. The k-means is a partitioning clustering algorithm. It process intends to minimize the least square error principle [15]. The partitioning algorithms handled large amount of document datasets than hierarchical clustering, as earlier told, this algorithm is best suited because of few computational needed[16, 9, 3].

The process of standard k-means algorithm as following- There are used two values as input, one is set of data vector named 'D' and another one is 'k' number of optimal cluster. Each clusters has centroid and then the residual objects assign to the cluster represent by the most related or nearest centroid. Again recomputed new centroid for each cluster and based on new centroid re-assigned the all document [10.11]. This process repeated again and again until find out a fixed solution, where all data object remains in the similar cluster after update of centroid. Steps are following:

Input data:

D: It is a collection of set of 'n' data vectors

k: It is a number of clusters that randomly selected.

Output:

Obtain a set of data that contain ' k' clusters

Steps:

- Randomly picked 'k' data vector from n data set after that calculate first cluster center and assign to cj ($1 \leqslant j \leqslant k$)
- Determining the distance of all data vectors in 'D' from every clusters center in 'C'
- Now that data vector has minimum distance i.e. Most similar data vector is put into the cluster
- Again find out the updating cluster center by calculating the average of data vector assigned to a cluster
- Repeat step 2 to 4 until get same mean and stop

A. Stopping criteria:

Basically K-means algorithm calculates the distance measure. Main aims to minimize the distance within clusters. As because of this, algorithm does not directly adjust into the similarity measure, because low values show dissimilarity [12]. Here distance measure and similarity both are calculated separate. The Euclidean distance, averaged KL divergence and cosine similarity calculate distance measure and Jaccard and Pearson coefficient calculate similarity measure. Conversion of the similarity measure to distance values here used a simple transformation. Because the range of cosine similarity and Jaccard coefficient are [0, 1] and both are bounded with monotonic. For Pearson coefficient, take corresponds distance values is ($D_p = 1-SIM_p$). It limits from -1 to +1. Here –

If $SIM_p \geq 0$ the select $D_p = 1 - SIM_p$ and

if $SIM_p < 0$ then take $D_p = |SIM_p|$.

V. EXPERIMENTS

It very hard to study to compare the impact of similarity metrics on cluster quality, because of very difficult to calculate cluster quality in itself. For evaluating clusters, generally used as a baseline criterion when humanly assigned category labels. As a result, compared with clusters and pre-defined category structure. The cluster, which is generate in an unsupervised way, and created intelligently predefined category structure of the cluster.

The usage of clustering is just to reproduce human thinking. So the idea of clustering is better if it consists humanly created categories. Often, dataset comes without any humanly created categories, when it is used in practice and clustering can help for that right point.

In this situation, there are two criteria can be used for evaluation [13]

1. Measures- cluster coherence in terms of the within cluster distances.
2. Well separateness between clusters in terms of the between cluster distances

Here we want to outcome of this investigation to compare with result of previously researches. But for this, a special type of datasets are selected which have been frequently used for evaluating clustering as the testing datasets.
And an appropriate category labels have been already assigned for all datasets. Remaining part of this datasets describes the characteristics, and then gives the explanation and performance evaluation on measures and simulation analyzes the simulation outcomes [14].
In some cases the classification of document is needed. Classification of document as would have done by human, sometime we may want same. We want to say that if cluster work on replica of human thinking then clustering is good. But often, we do not have supervised document i.e. called labeled data. That's why is the reason, clustering repeatedly again and again. In this matter, distances the within cluster and inter clusters separateness on the basis of distances between cluster can be used for evaluation [6].
Here we are using the same dataset for evaluation that were used in previous work [4]. The dataset are:

Table 1: Summary of Datasets Used

Data	Documents	Classes	Terms	Source	Description
20news	18828	20	28553	20news-18828	Newsgroup posts
Classic	7089	4	12009	CACMCISICRANFIELDMEDLINE	Academic papers
Hitech	2301	6	13170	San Jose Mercury (TREC, TIPSTER)	Newspaper articles
re0	1504	13	2886	Reuters-21578	Newsgroup posts
tr41	878	10	7454	TREC5 and TREC6 (TREC 1999)	Newspaper articles
wap	1560	20	8460	WebACE	Web pages

A. Evaluation

The above dataset in table1, we have find out clustering outcome from the k-mean algorithm. In the data set, the number of assigned different categories as same with set of numbers of cluster. We have used two evaluation measure-purity and entropy that given quality of a clustering result. The mostly used of purity and entropy, for examine the unsupervised learning algorithm's performance. In this technique the process begins with-

a) Every cluster is marked or named with groups of group of members that visible in that cluster.

b) If a category label has been assigned or allocated to particular cluster, it still can assign to other cluster, if it is key class of that cluster. We have calculated purity and entropy measure, based on the cluster labels [18, 20].

B. Results

In this study, the methodology of cluster analysis is partial. For this, similarity measure or distance measure is needed. We have observed there are three basic components, effects clustering algorithm, distance or similarity measures, representation of the objects. In table 2 and 3 represent the 30 experiment. Here we can see, the Euclidean is very simple way to measure distance between. That cluster made by using two technique pearson correlation and cosine similarity is closely nearest to human made categories which is makes by using replica of human thinking and are the best used of its if the final task is little bit alike. There are two methods, Pearson and Jaccard that discover the a large amount coherent cluster with better purity values. And it representing document from a single group dominate each clusters.

Table 2: Purity Results

DATA	EUCLIDEAN	COSINE	MAHALANOBIS	JACCARD	PEARSON
20news	.11	.46	.44	**.47**	**.47**
Classic	.58	.85	.70	**.97**	.82
Hitech	.24	**.60**	.51	**.60**	.51
re0	.41	.71	.62	.77	**.81**
tr41	.69	.70	.71	.77	**.80**
Wap	.31	.59	.52	.63	**.66**

Table 3: Entropy Results

DATA	EUCLIDEAN	COSINE	MAHALANOBIS	JACCARD	PEARSON
20news	.89	.48	.47	.49	**.44**
Classic	.75	.26	.32	**.06**	.28
Hitech	.91	.65	.70	.64	**.61**
re0	.64	.24	.39	.31	**.21**
tr41	.59	.35	.40	.31	**.27**
wap	.70	.35	.36	**.34**	.37

As shown in table 4, when the number of documents is increase in experiment before reaching a final terminal value then the result of purity and entropy is gets better.

Shown graphically in figure1, the blue cure show the purity and orange curve represent the entropy. As the result, if machine quality is better and could use the entire document then here we can hope get the better outcome and outperform all previous results.

NUMBER OF DOCUMENTS	100	200	500	1000	2000	4000
Purity	.40	.41	.44	.44	.47	.48
Entropy	.62	.59	.55	.47	.45	.44

Table 4: Disparity table of purity and Entropy with Number of Documents

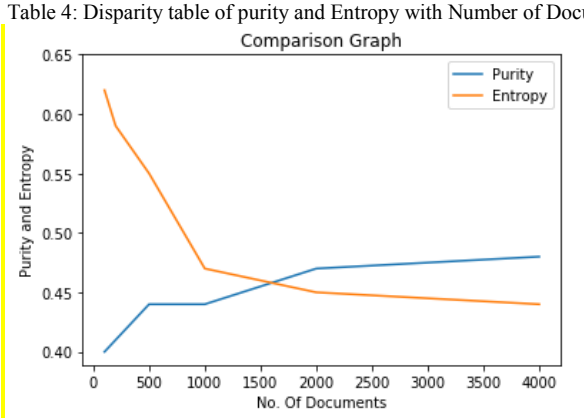

Figure1: Graph representation of Purity and Entropy with No. of Documents

VI. CONCLUSION

The conclusion of this research paper is that, the study of Euclidean distance measure is more effective for document clustering, apart this , here study other measure that have similar effective for the Partitioned text document clustering assignment. Pearson and cosine method is much unbiased to obtain the clusters and it is very nearest to human made categories. If final task quite similar then can be the best used of replicating human thinking.

Pearson and Jaccard method more suitable for finding rational clusters with high clarity value that represent by documents from a single group that control every cluster. The future work of this paper to provide better accuracy and venture for the text mining similarity measure [19]. This paper focuses on improvement in 'bag of word' approach and gives some new ideas and semantics.

REFERENCES

1] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. ACM Computing Surveys (CSUR), 31(3):264–323, 1999.

[2] P. Willett. Recent trends in hierarchic document clustering: a critical review. Information Processing and Management: an International Journal, 24(5):577–597, 1988.

[3] G. Salton. Automatic Text Processing. Addison-Wesley, New York, 1989.

[4] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In AAAI-2000: Workshop on Artificial Intelligence for Web Search, July 2000.

[5] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning, 55(3), 2004.

[6] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In Proceedings of the International Conference on Information and Knowledge Management, 2002.

[7] J. M. Neuhaus and J. D. Kalbfleisch. Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics, 54(2):638–645, Jun. 1998.

[8] R. B. Yates and B. R. Neto. Modern Information Retrieval. ADDISONWESLEY, New York, 1999.

[9] Martin Ester, Hans peter Kriegel, Jrg S, and Xiaowei Xu. A densitybased algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[10] Gerard Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[11] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In In KDD Workshop on Text Mining, 2000.

[12] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '99, pages 16–22, New York, NY, USA, 1999. ACM.

[13] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '92, pages 318–329, New York, NY, USA, 1992. ACM.

[14] J. M. Neuhaus and J. D. Kalbfleisch. Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics, 54(2):pp. 638–645, 1998.

[15] Anna Huang. Similarity Measures for Text Document Clustering. In Jay Holland, Amanda Nicholas, and Delio Brignoli, editors, New Zealand Computer Science Research Student Conference, pages 49–56, April 2008.

[16] Andreas Hotho, Steffen Staab, and Gerd Stumme. Wordnet improves text document clustering. In In Proc. of the SIGIR 2003 Semantic Web Workshop, pages 541–544, 2003.

[17] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

[18] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.

[19] D. R. Cutting, J. O. Pedersen, D. R. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, 1992.

[20] D. Arthur and S. Vassilvitskii. k-means++ the advantages of careful seeding. In Symposium on Discrete Algorithms, 2007.