International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

Improved Similarity Measure For Text Classification And Clustering

Rahul Nalawade[1], Akash Samal[2], Kiran Avhad[3]

[1]*Computer Engineering Department , STES' Sinhgad Academy Of Engineering,Pune*
[2]*Computer Engineering Department , STES' Sinhgad Academy Of Engineering,Pune*
[3]*Professor, Dept. of Computer Engineering, STES' Sinhgad Academy Of Engineering,Pune*

Abstract – *Computing the similarity between documents is an important operation in the text processing. In this paper, a new similarity measure is proposed. To calculate the similarity between two documents with respect to a feature, the proposed measure takes the following three cases in to account I) The same feature appears in both documents, II) The same feature appears in only one document, and III) The same feature appears in none of the documents. For the first case, the similarity will increases as the difference between the two involved feature values decreases. For the second case, a fixed value is involved to the similarity. For the last case, the feature has no appearance to the similarity. The proposed measure is extended to the similarity between the sets of documents. The effectiveness of our measure is computed on the number of data sets for text clustering and classification. The performance obtained by the proposed measure is better than achieved by other measures.*

Key Words: Text Classification , Text Clustering , Clustering Algorithms , Preprocessing , Tokenization , Stemming

1. INTRODUCTION

The progress of computer technology in the few decades has led to large supplies of powerful and affordable computers. Increase in the large electronic documentation it is hard to visualize these documents efficiently by putting manual effort. These have brought challenges for the efficient and effective organization of web page documents automatically. Extracting features from web pages is first task found in mining. On the basis of extracted features similarity between web pages are going to be measure. There is various similarity measures are pointed out for work. Data mining is the technique of mining the previously unknown and potentially useful information from data. Document clustering organizes documents into different clusters. The documents in each cluster share some common properties according to similarity measure. Text clustering algorithms play an important role in helping users to effectively organize, navigate and summarize the information. Due to explosive growth of accessing information from the web, efficient access of information are needed critically. The Text processing plays an important role in information retrieval, web search and data mining. Text mining attempts to discover new, previously unknown information by applying techniques from data mining. Data mining techniques is an unsupervised learning where clustering methods try to identify groupings of the text documents, so that a set of clusters is produced in which clusters exhibit high intra cluster similarity and low inter cluster similarity. Generally, text document clustering methods attempt to keep the documents into groups where each group represents similar topic that is different than those represented by the other groups. The similarity measure reflects the degree of closeness of the target objects and should relate to the characteristics that are believed to distinguish the clusters in the data. Before Clustering a similarity/distance measure must be determined. Choosing an appropriate similarity measure is also difficult for cluster analysis, especially for a particular type of clustering algorithms. Text Categorization is the classification of documents with respect to a set of one or more pre-existing categories. The classification phase consists of generating weighted vector for all categories, then using a similarity measure to the closest category. The similarity measure is used to determine the degree of resemblance between two vectors. For reasonable classification results, a similarity measure should generally respond with larger values to documents which belongs to same class with smaller values. During the last decades, there are a number of methods for text categorization were typically based on the classical Bag-of-Words model where each term is an independent feature. Below the figure shows basic structure of similarity measure. Two documents are taken for similarity measure. After submitting by clicking on submit button. It will show whether two documents are similar or not. If similar then display percentage of similarity.

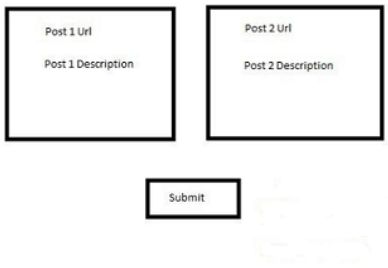

Figure 1: Basic Structure of Similarity Measure

1.1 LITERATURE SURVEY

Gaddam Saidi Reddy and Dr.R.V.Krishnaiah approach in finding similarity between documents while performing clustering is multi-view based similarity. All distance measures such as Pearson, Euclidean, Jaccard, and cosine are compared. They concluded Ecludean and Jaccard are best for web document clustering. Ecludean and Jaccard are selected related attributes for given subject and calculated distance between two values. Both of them used an algorithm known as Hierarchical Agglomerative Clustering in order to perform clustering. Their computational complexity is very high that is the drawback of these approaches. Proposed a similarity measure known as MVS (Multi-Viewpoint based Similarity), when it is compared with cosine similarity, MVS is more useful for finding the similarity of text documents. The empirical results and analysis revealed that the proposed scheme for similarity measure is efficient and it can be used in the real time applications in the text mining domain. It makes use of more than one point of reference as opposed to existing algorithms used for clustering text documents. Shady Shehata, Fakhri Karray and Mohamed S. Kamel mentioned that the most of the common techniques in text mining are based on the statistical analysis of a term, either word or phrase. Text mining model should indicate terms that capture the semantics of text. The mining model can capture terms that present the concepts of the sentence, which leads to discovery of the topic of the document. The mining model that analyzes terms on the sentence, document, and corpus levels are introduced, can effectively discriminate between non important terms with respect to sentence semantics and terms. The term which contributes to the sentence semantics is analyzed on the sentence, document, and corpus levels rather than the traditional analysis of the document only. It is important to note that extracting the relations between verbs and their arguments in the same sentence has the potential for analyzing terms within a sentence. The information about who is doing what to whom clarifies the contribution of each term in a sentence to the meaning of the main topic of that sentence. It is shown that the standard deviation is improved by using the concept-based mining model. Anna Huang declared that before clustering, a similarity/distance measure must be determined. The measure determines the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. It is very difficult to conduct a systematic study comparing the impact of similarity metrics on cluster quality, because objectively evaluation g cluster quality is difficult in itself. The clusters, which are generated in an unsupervised way, are compared to the pre-defined category structure, which is normally created by human experts. This kind of evaluation assumes that the objective of clustering is to replicate human thinking, so a clustering solution is good if the clusters are consistent with the manually created categories. It is found that there is no measure that is universally best for all kinds of clustering problems. The performance of the cosine similarity, Jaccard correlation and Pearson's coefficient are very close, and are significantly better than the Euclidean distance measure experimented with the web page documents. Hung Chim and Xiaotie Deng found that the phrase has been considered as a more informative feature term for improving the effectiveness of document clustering. They proposed a phrase -based document similarity to compute the pairwise similarities of documents based on the Suffix Tree Document (STD) model. By mapping each node in the suffix tree of STD model into a unique feature term in the Vector Space Document (VSD) model, the phrase -based document similarity naturally inherits the term tf-idf weighting scheme in computing the document similarity with phrases. They applied the phrase-based document similarity to the group-average Hierarchical Agglomerative Clustering (HAC) algorithm and developed a new document clustering approach. Their evaluation experiments indicate that the new clustering approach is very effective on clustering the documents. Finally they found that both the traditional VSD model and STD model play important roles in text-based information retrieval. The concept of the suffix tree and the document similarity are quite simple, but the implementation is complicated. Detailed Investigation is required to improve the performance of the document similarity. They conclude that the feature vector of phrase terms in the STD model can be considered as an expanded feature vector of the traditional single-word terms in the VSD model. Yanhong Zhai and Bing Liu studied the problem of extracting data from a Web page that contains several structured data records. They proposed approach to extract structured data from Web pages. Even if the problem has been studied by several researchers, the existing techniques are either not accurate or make many strong assumptions. Inderjit Dhillon, Jacob Kogan & Charles Nicholas found that in particular, when the processing task is to divide a given document collection into clusters of similar documents a choice of good features along with good clustering algorithms is of paramount importance. Feature or term selection along with a number of clustering techniques is important. Syed Masum Emran and Nong Ye said distance metric value is used to find the similarity or dissimilarity of the current observation from the already established normal profile. To find the distance between normal profile and current observation value, one can use many distance metrics. Alexander Strehl, Joydeep Ghosh, and Raymond Mooney studied if clusters are to be meaningful, the similarity measure should not be change to transformations natural to the problem domain. The features have to select carefully. They conducted a number of experiments to assure statistical significance of results. Metric distances such as Euclidean are not appropriate for high dimensional, sparse domains. Cosine, correlation and extended Jaccard measures are successful in capturing the similarities implicitly indicated by manual categorizations as they seen for example in Yahoo. S. Kullback and R. A. Leibler found that in terms of similarity measure for information retrieval, difficult it is to discriminate between the populations. R. A. Fisher introduced the criteria for sufficiency required that the statistic chosen should summarize the whole of the relevant information supplied by the sample. Mei-Ling Shyu, Shu-Ching Chen, Min Chen & Stuart H. Rubin mentioned that In their experimental results they found that the Euclidean distance gives the worst performance, followed by the cosine coefficient. Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee proposed a new measure

for computing the similarity between two documents. Several characteristics are embedded in this measure. The difference between presence and absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity decreases when the number of presence-absence features increases. An absent feature has no appearance to the similarity. The similarity increases as the difference between the two values associated with a present features decreases. This work mainly focuses on textural features. To improve the efficiency, they have provided an approximation to reduce the complexity involved in the computation. The results have shown that the performance measure by the proposed measure is better than other measures.

1.2 EXISTING SYSTEM

The K value of the initial cluster centres are determine the impact throughout the clustering process and the clustering results, the K value in practical applications is very difficult to direct . If the amount of data tends to infinity which is, the K value of the K-means algorithm is difficult to determine. At present, there are two algorithms of clustering to determine the K value is effective which is the cost function based on the distance and propagation clustering algorithm based on the nearest neighbours. Thus to obtain the corresponding K value. The latter using nearest neighbour clustering algorithm to calculate the number of cluster center, the number of cluster centre provides for the maximum K value to get the optimal value of K. Second, about the assumption of initial cluster centers. K-means using the iterative method to solve the problem, except the first step, the clustering results of each step are improved otherwise terminate the process of iteration. The K-means clustering algorithm takes the cluster squares error and the criterion function value change or not change as iterative termination conditions. But the clustering results obtained from this criterion function easily fall into local minimum solution, the result is the clustering results of search are moving toward the direction of diminishing the criterion function value.

A. Procedure of K-means Algorithm

1. Distribute all objects to K number of different cluster at random;
2. Calculate the mean value of each cluster, and use this mean value to represent the cluster;
3. Re-distribute the objects to the closest cluster according to its distance to the cluster center;
4. Update the mean value of the cluster. That is to say, calculate the mean value of the objects in each cluster;
5. Calculate the criterion function E, until the criterion function converges
   $E = \Sigma kj = 1 \Sigma ni = 1 xi = cj \|xi - mj\| 2$

In which, E is total square error of all the objects in the data cluster, xi bellows to data object set, mi is mean value of cluster Ci (x and m are both multi-dimensional). The function of this criterion is to make the generated cluster be as compacted and independent as possible.

B. Limitations

1. **Empty Clusters Handling:** The problems with the K-means algorithm given earlier is that empty clusters can be obtained if points are not allocated to a cluster . If this happens, then a strategy is needed to choose a other centroid, since otherwise, the squared error will be larger than necessary.

2. **Difficult to measure the no of clusters:** The user has to first choose the value of K, the number of clusters. Although for 2D data this choice can easily be made by visual inspection, it is not for higher dimension data, and there are no clues as to what number of clusters might be appropriate.

2. PROPOSED SYSTEM

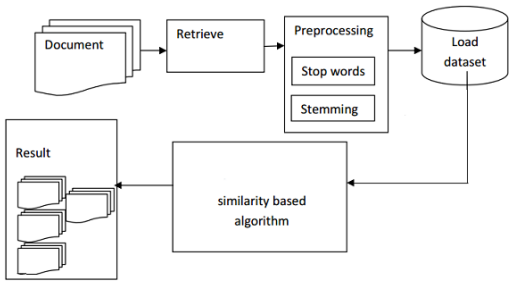

Figure 2: System Architecture

The Number of documents is crawl from different sources and Preprocessing is done on it. Following is the three steps of preprocessing the results obtained by conventional search engine are preprocessed. For that we have to follow some preprocessing steps.

1. Stop words Removal
2. Tokenization
3. Stemming

1. Stop words removal

Many times, it makes sense do not index "stop words" during the indexing process. The Stop words are words which have very little informational content. The stop words such as: and, the, of, it, as, may, that, a, an, of, off,etc. Studies have shown that by removing stop words from the index, you may get reduced index size without significantly affecting to the accuracy of a query ie user's query. By eliminating stop words , the index size is reduced by about

33 percent for a word level index. After stop words removal tokenization and stemming is performed.

2. Tokenization

Tokenization is the process of making sensitive data with unique identification symbols which retain all the essential information without compromising its security.

3. Stemming

The concept stemming has been applied to information systems from their initial automation in 1916; s. the original goal of stemming was to increase performance and having less system resources by reducing the number of unique words that a system has to contain. Stemming algorithms are used to improve the efficiency of the information system and to improve recall. Stemming is the process for reducing words to their stem; base or root forms generally a written word form. There are many stemming algorithms are available. After preprocessing a similarity measure algorithm is applied and cluster is formed. Assign the documents to cluster and result is display to user.

A. Improved K-means Clustering Algorithm

Optimize the initial cluster centers, to find a set of data to reflect the characteristics of data distribution as the initial cluster centers, to support the division of the data to the greatest extent. Optimize the calculation of cluster centers and data points to the cluster center distance, and make it more match with the goal of clustering.

Algorithm: Improved K-means Algorithm

Input: data set x contains n data points; the number of cluster is k.

Output: k clusters of meet the criterion function convergence. Program process-

1. Initialize the cluster center.

1.1. Select a data point xi from data set X, set the identified as statistics and compute the distance between xi and other data point in the data set X. If it meet the distance threshold, then identify the data points as statistics, the density value of the data point xi add 1.

1.2. Select the data point which is not identified as statistics, set the identified as statistics and compute its density value. Repeat Step 1.2 until all the data points in the data set X have been identified as statistics.

1.3. Select data point from data set which the density value is greater then the threshold and add it to the corresponding high-density area set D.

1.4. Filter the data point from the corresponding high-density area set D that the density of data points relatively high, added it to the initial cluster center set. Followed to find the k-1 data points, making the distance among k initial cluster centers are the largest.

2. Assigned the n data points from data set X to the closet cluster.
3. Adjust each cluster center K.
4. Calculate the distance of various data objects from each cluster center by formula and redistribute the n data points to corresponding cluster.
5. Adjust each cluster center K.
6. Calculate the criterion function E, to determine whether the convergence, if convergence, then continue else go to the Step 4

B. Solutions For Existing Systems Limitations

Solution 1: One approach is to choose the point that is farthest away from any current centroid. If nothing else, this eliminates the point that currently contributes most to the total squared error. Another approach is to choose the replacement centroid from the cluster that has the highest SSE (Sum of Squared Error). This will typically split the cluster and reduce the overall SSE of the clustering. If there are several empty clusters, then this process can be repeated several times.

Solution 2: To calculate the no of clusters we have to make the possible no of groups of observations and for each no of group we have to assign the cluster. From this we have calculate the no of clusters

C. Overcome To K-Means Disadvantages

To overcome to K-Means Disadvantages we should use optimization algorithms to minimize the cross-validation error.

3. Cosine Similarity

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. cosine similarity is most commonly used in high-dimensional positive spaces. For example, in information retrieval and text mining, each term is notionally assigned a different dimension and a document is characterised by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter.One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors, as only the non-zero dimensions need to be considered.

The cosine of two vectors can be derived by using the following formula:

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

Wher,
qi is the tf-idf weight of term i in the query.
di is the tf-idf weight of term i in the document.
cos(q,d) is the cosine similarity of q and d … or, equivalently, the cosine of the angle between q and d.

Example :Cosine Similarity

Following table is tf-idf values of documents.on which Cosine similarity will apply.

	Arjun Kapoor	John Abraham	Akshay Kumar	Imran Khan	Alia Bhatt	Katrina Kiaf	Kangana Ranauth	Amir Khan	Hrithik Roshan	Ajay Devgan
d1	0	0	0.17	0	0.54	1.61	0	0.54	0.17	0.54
d2	0.61	0.61	0.51	0	0	0	0.17	0	0	0
d3	0	0	0	0.92	0	0	0.51	0	0.51	0
d4	0.46	0	0.46	0	0	0	0.51	0	0	0
d5	0	0.92	0	0.42	0	0	0	0	0.51	0

Table-1:Example tf-idf values of documents

cos(d3,d1)≈0*0.61+0*0.61+0.51∗0.17+0*0.54+0.17*0+0*0.54+0*0.54 ≈ 0.08

cos(d3,d2)≈0*0.61+0*0.61+0∗0.51+0*0.92+0.17*0.51+0*0.51+0*0 ≈ 0.08

cos(d3,d4)≈0∗0.46+0∗0+0∗0+0.92*0.92+0*0+0*0+0.51*0.51 ≈ 0.26

cos(d3,d5)≈0∗0+0∗0.92+0∗0+0.92*0.42+0*0+0*0+0.51*0.52 ≈ **0.64**

Answer: d3 and d5 are more similar to each other.

4. Results

Performance Comparison by improved *k*-means with Different Measures on Testing Data of Reuters-8.here we used cosine similarity measure . cosine gives better performance than euclidean and euclidean-jaccard. Cosine performs as well as runs much faster.

AC				
	$k=8$	$k=16$	$k=24$	$k=32$
Euclidean	0.5117	0.5139	0.5161	0.5241
EJ	0.6073	0.6107	0.6039	0.5952
Cosine	**0.6395**	**0.6649**	**0.6495**	**0.6421**
En				
	$k=8$	$k=16$	$k=24$	$k=32$
Euclidean	0.6423	0.6389	0.6351	0.6209
EJ	0.4698	0.4632	0.4578	0.4608
Cosine	**0.4411**	**0.4273**	**0.4348**	**0.4392**

Table-2: Performance Comparison by improved *k*-means

5 CONCLUSIONS

We have presented a novel similarity measure between documents. Several desirable properties are embedded in this measure. The similarity degree increases when the number of presence-absence features pair's decreases. We use optimization algorithm to minimize cross validation error. The main article is at top and others at bottom so duplication is avoided. User will get all the related data at one site .The proposed system has also been extended to measure the similarity between sets of documents.

REFERENCES

[1] Gaddam Saidi Reddy and Dr.R.V.Krishnaiah," Clustering Algorithm with a Novel Similarity Measure", IOSR Journal of Computer Engineering (IOSRJCE), Vol. 4, No. 6, pp. 37-42, Sep-Oct. 2012.

[2] Shady Shehata, Fakhri Karray, and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering , Vol. 22, No. 10, October 2010.

[3] Anna Huang, Department of Computer Science, The University of Waikato, Hamilton, New Zealand," Similarity Measures for Text Document Clustering", New Zealand Computer Science Research Student Conference (NZCSRSC), Christchurch, New Zealand, April 2008.

[4] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 9, pp. 1217 – 1229, 2008.

[5] Yanhong Zhai and Bing Liu, "Web Data Extraction Based on Partial Tree Alignment", International World Wide Web Conference Committee (IW3C2), ACM 1-59593-046, 9/05/2005.

[6] I. S. Dhillon, J. Kogan and C. Nicholas, " Feature Selection and Document Clustering", In Berry MW Ed. A Comprehensive Survey of Text Mining, 2003.

[7] Syed Masum Emran and Nong Ye, "Robustness of Canberra Metric in Computer Intrusion Detection", IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY, 5-6 June, 2001.

[8] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney," Impact of Similarity Measures on Web-page Clustering", Workshop of Artificial Intelligence for Web Search, July 2000.

[9] S. Kullback and R. A. Leibler, "On information and sufficiency", Annuals of Mathematical Statistics, Vol. 22, No. 1, pp. 79–86, March 1951.

[10] Mei-Ling Shyu, Shu-Ching Chen, Min Chen and Stuart H. Rubin," Affinity-Based Similarity Measure for Web Document Clustering", Distributed Multimedia Information System Laboratory, School of Computer Science Florida International University Miami, FL 33199, USA.

[11] Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee," A Similarity Measure for Text Classification and Clustering", IEEE Transactions On Knowledge And Data Engineering, 2014.

[12] J. Han and M. Kamber,Data Mining: Concepts and Techniques,2nded. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA,USA: Elsevier, 2006.