2016 IEEE Second International Conference on Big Data Computing Service and Applications

A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering

Arash Heidarian
Department of Computer Science
University of Auckland
Auckland, New Zealand
Email: ahei844@aucklanduni.ac.nz

Michael J. Dinneen
Department of Computer Science
University of Auckland
Auckland, New Zealand
Email: mjd@cs.auckland.ac.nz

***Abstract*—The increasing numbers of textual documents from diverse sources such as different websites (e.g. social networks, news, magazines, blogs and medical recommendation websites), publications and articles and medical prescriptions leads to massive amounts of daily complex data. This phenomenon has caused many researchers to focus on analysing the content and measuring the similarities among the documents and texts to cluster them. One popular method to measure the similarity between documents is to represent the documents as vectors and measure the similarity among them based on the angle or Euclidean distance between each pair. By only considering these two criteria for similarity measurement, we may miss important underlying similarities in this area. We propose a new method, TS-SS, to measure the similarity level among documents, in such a way that one hopes to better understand which documents are more (or less) similar. This similarity level can be used as a handy measure for clustering and recommendation systems for documents. It also can be used to show top n similar documents to a particular document or a search query. Our study gives insights on the drawbacks of geometrical and non-geometrical similarity measures and provides a novel method to combine the other geometric criteria into a method to measure the similarity level among documents from new prospective. We apply Euclidean distance, Cosine similarity and our new method on four labelled datasets. Finally we report how these three geometrical similarity measures perform in terms of similarity level and clustering purity using four evaluation techniques. The evaluations' results show that our new model outperforms the other measures.**

***Keywords*-Document Clustering; Similarity Level; Document Similarity; Geometric Similarity; VSM**

I. Introduction

There are several algorithms used for measuring similarity between documents such as Pearson (Kornbrot 2005), Spearman (Zar 1998), Kullback-Leibler divergence (Kullback & Leibler 1951), Jaccard coefficient (Jaccard 1912), Shannon (Wartena & Brussee 2008), Euclidean Distance (ED) and Cosine similarity (Salton & Buckley 1988). One popular method to measure the similarity between documents is to use the terms within the documents to represent them as vectors and measure the similarity among them based on the angle or Euclidean distance between each pair. By only considering these two criteria for similarity measurement, we may miss important underlying semantic similarities in this area. We propose a new method, TS-SS, to measure the similarity level among documents, in such a way that one hopes to better understand which documents are more (or less) similar. Though the results of some studies (Salton & Buckley 1988, Nelson et al. 2004) show the geometric and VSM based models are more robust in measuring similarities among documents compared to non-geometric models, there are not many works focusing on geometric methods to boost these similarity measures for better results. That is why in this study we focus on geometric methods.

The main purpose of this study is to measure the similarities among documents with high accuracy in order to have a better understanding of which documents are more similar (or less similar). We call this concept *similarity level* which is focused in this study. In this work, *accuracy* refers to the power of a measure in differentiating the similarity level among documents in such a way that one can understand which documents are least, less, more and most similar. This power of differentiation can be significantly useful for recommendation systems and clusterings. It also can be used to find top n similar documents to a particular document or a search query.

As mentioned earlier, this study focuses on geometric similarity measures which are popular for document clustering but there are not much works on improvement of the current geometric models in such a way that can be used for measuring a concept called *similarity level*. In some research and surveys (Nelson et al. 2004, Cross et al. 2002), diverse similarity measures used for IR have been evaluated and results show Cosine similarity outperforms other measures. This study gives insights on the drawbacks of geometrical and some non-geometrical similarity measures and provides a novel method to combine the other geometric criteria into a method to measure the similarity level among documents from new prospective.

This study contains the following chapters: in Chapter II some studies related to recommendation systems, semantic concept for measuring similarity and geometrical space to measure the similarities are reviewed in brief.

978-1-5090-2251-9/16 $31.00 © 2016 IEEE
DOI 10.1109/BigDataService.2016.14

In Chapter III firstly, text pre-processing techniques to prepare the documents are described. Secondly we explain how Vector Space Model forms vectors from body of texts. Then we explain how traditional geometric similarity measures namely Cosine similarity and Euclidean distance use vectors associated to each document to identify the similar between documents/texts. Finally, clustering techniques, clustering tools and methods for evaluating clusters are explained in details.

In Chapter IV drawbacks of Cosine similarity and Euclidean distance have been scrutinized from different views and it is explained why existing geometric measures are not robust enough to measure the similarity level accurately. In Chapter V a new method called TS-SS proposed, which covers the mentioned drawbacks in Chapter IV to measure the similarity level among documents. In Chapter VI, four datasets used for experiments and four evaluation methods for making comparison among the measures are described in detail. In Chapter VII the proposed method and other similarity measures are applied on four datasets to cluster the documents and then the results of clustering and other evaluations methods of TS-SS and other similarity measures are compared using four evaluation methods. The evaluations' results show the proposed model, TS-SS outperforms the other measures. Finally in Chapter VIII the time complexity of Cosine similarity, Euclidean distance, TS-SS and K-Means algorithm which is used for clustering are calculated.

II. Related works

In this study, the literature briefly reviews some works around document similarity measures which are related to recommendation systems (Nelson et al. 2004), semantic concept for measuring similarity (Hammouda & Kamel 2002, Chim & Deng 2008, Lakkaraju et al. 2008, Elsayed et al. 2008, Wan & Peng 2005) and geometrical space to measure the similarities (Lebanon 2006, Zhang et al. 2012, Sahami & Heilman 2006, Becker & Kuropka 2003). However, to the best of our knowledge no previous work has used the idea of measuring similarity level, so this approach is a significant contribution. Hamuda and Kamel present a semi-structured phrase-based document similarity model which indexes web documents based on phrases rather than single term only. This model identifies potential phrases which match between documents and it indicates strong similarity between the documents (Hammouda & Kamel 2002). Chim and Deng also present phrase-based document similarity based on Suffix Tree Model in an efficient way (Chim & Deng 2008). The earth mover's distance (EDM) method measures similarity by computing semantic distance between words based on electronic lexical database WordNet (Wan & Peng 2005). Lebanon uses Riemannian geometry associated with differentiable manifold and set of points to measure the distance between documents based on the provided data. In general the approach is related to maximum likelihood under a model which assigns probabilities inversely proportional to Riemannian volume element (Lebanon 2006). Zhang et al. present a similarity measure space model for document clustering. The model derives low dimensional semantic subspace of documents corresponding to the same semantic by maximizing and minimizing the correlation between the documents in the local patches and outside these patches respectively (Zhang et al. 2012). Sahami and Heilman proposed a kernel function to measure the similarity between short texts. First x is issued as a query. Then for the retrieved documents using query x, TF-IDF is used to weight the words in each document, and finally the similar documents are identified based on the weights of terms (Sahami & Heilman 2006). One of the issues with this method is that, it is applicable for short texts only, because passing a long text as a query, is an expensive operation. Becker and Kuropka developed Topic-based Vector Space Model (TVSM) to measure the document similarity. In this approach, each term in the document is represented as vector space and the weight of the term is defined as calculated as the length of vector. In the paper an example is shown to illustrate how the direction and length of the vectors play a role in finding the important terms in the documents and weight assignments. Terms which have the longer magnitude (which meet the predefined length threshold) and tighter angle (close to 0), are considered as the important terms. Then after the weights of terms in each document is used to calculate the Cosine similarity between each pair of documents. The main concept of this work is similar to our study as they took the magnitude and direction of vectors into account as well (Becker & Kuropka 2003). Nelson $et\ al.$ studied and compared two recommendation server methodologies implemented for the NASA Technical Report Server (NTRS). One method is the log analysis and the other one is VSM. They measured the similarities using Cosine similarity to recommend top 10 similar documents. After running the experiments they found out, in general, Cosine outperforms the log analysis (Nelson et al. 2004). The result of their study and other studies (Chim & Deng 2008, Cross et al. 2002) shows VSM gives better results on measuring similarity among documents and that is why we emphasize on VSM model.

III. Background

In this section, first we describe the data preparation including text preprocessing and how the Vector Space Model forms vectors from body of texts. Then we briefly explain two traditional geometric similarity measures namely Cosine similarity and Euclidean distance.

A. Data Preparation

In order to measure the similarity among documents meaningfully, the features have to be chosen carefully to

be on comparable scale and the result has to reflect the underlying semantics (Strehl et al. 2000). The magnitude of vectors (which reflects the importance of terms to a document in term of frequency) can be used as an influential and powerful tool in computing the similarity (Becker & Kuropka 2003). In this study, we have avoided normalizations due to its significant affect on term frequencies and consequently on magnitudes of vectors which leads to inaccurate similarity measurement (Das et al. 2009, Singhal et al. 1996, Strehl et al. 2000). First step is text preprocessing to transform texts into a set of useful data which can be a set of keywords or phrases. To this end, feature extraction technique is applied in order to extract specific bits of information (Heidarian 2011). That is, first we tokenize all texts to split up a string of characters into a set of tokens such as words, punctuation marks, symbols, or other meaningful elements (Laboreiro et al. 2010). After tokenization, Feature Selection stage reduces the dimensionality of tokenized texts by removing irrelevant, redundant and noisy data such as hashtags, links, punctuation, emotion text icons, linking words and stop words (Heidarian 2011). Finally we use stemming technique to reduce inflected words to their stems (Porter 1980). Consequently each document is represented as a bag of words cleared from noisy data and ready to be represented as a vector based on the search query (all text preprocessing steps applied on searched query as well). Finally we measure the similarity between documents by computing the similarity between the associated vectors.

B. Vector Space Model

Vector Space Model (VSM) is a set of vectors with addition and scalar multiplication which introduced by Salton (Salton & Buckley 1988) to describe objects (documents and corpus of texts) using n-dimensional vectors which each dimension representing the frequency of a certain term in a document using TF-IDF model. One of the main drawbacks of TF-IDF is its lack of accuracy due to length of the documents. Long documents have higher term frequencies as they repeat the same term more often (Singhal et al. 1996). The main solution to this problem is different TF normalization components such as Euclidean normalization that dampens the quantity of TF significantly to a unified length (Das et al. 2009, Singhal et al. 1996) and may strongly affect on measurements in a negative way in some cases (Strehl et al. 2000). In order to use the term frequency in similarity measurements, we do not decrease or dampen the term frequencies and their distributions because they play a crucial role in our new similarity computation. To handle the mentioned drawback, we use TF-IDF Ranking algorithm (Wu et al. 2010) to avoid TF-IDF to bias toward long sentences: $\text{TF-IDF}(d,t) = \frac{\text{TF}(d,t)}{\sum_D \text{W}_{D,d}} \cdot \log(\frac{\text{U}}{\text{df}(t)})$ where TF(d,t) is the frequency of the term t within the texts of document d, U is the total number of documents, df(t) is the number of documents which their texts contain the term t and $\text{W}_{D,d}$ is the total number of words in the text from document d.

C. Similarity Measures

There are several algorithms used for measuring similarity between documents such as Pearson (Kornbrot 2005), Spearman (Zar 1998), Kullback-Leibler divergence (Kullback & Leibler 1951), Jaccard coefficient (Jaccard 1912), Shannon (Wartena & Brussee 2008), Euclidean distance and Cosine similarity (Salton & Buckley 1988). Most commonly used algorithms for measuring pairwise similarities between documents after they represented as vector spaces are Cosine similarity and Euclidean distance (Chim & Deng 2008). Although the results of many studies (Salton & Buckley 1988, Nelson et al. 2004) show the geometric and VSM based models are more robust in measuring similarities among documents compared to non-geometric models and this study emphasizes on geometric measures, in this section we briefly explain some of the most popular non-geometric similarity measures.

1) Non-geometric measures: Pearson correlation coefficient computes similarity between two variables bounded to -1 and +1. Coefficient 1 shows correlation is positive and two data objects are correlated perfectly, -1 indicates total negative correlation (Kornbrot 2005, Zar 1998). Jaccard Coeficient (Jaccard 1912) divides intersection of the objects by their unions. The produced coefficient ranges between 0 and 1. If two documents are same, the coefficient is 1 and it is 0 if there is no similarity between them. Kullback-Leibler Divergence (KLD) (Kullback & Leibler 1951) measures differences between two probability of distributions. It makes the automatic use of term sets for each category. Hence only those terms which belong to predefined category-term lists are taken into consideration and they are compared with each category term probability distribution. It means when a document contains only limited number of terms in comparison to the number of words in categories, the term frequency of many terms in that document is zero (Bigi 2003). Moreover it is asymmetric measure. Jensen-Shannon divergence (JSD) is based on the KLD with the difference that it is symmetric (Wartena & Brussee 2008).

2) Cosine Similarity: Cosine similarity computes the pairwise similarity between two documents using dot product and magnitude of vector document A and vector document B in high-dimensional space (Salton & Buckley 1988). The following formula calculates the Cosine similarity between vector (documents) A and vector B in n dimensional space:

$$\text{V} = \text{cosine}(A,B) = \frac{\sum_{n=1}^{k} A(n) \cdot B(n)}{|A| \cdot |B|} \qquad (1)$$

The resulting similarity ranges from minimum 0 to maximum 1. If the degree between A and B is 0, it means two

vectors are overlapped, in this condition two documents have the maximum similarity and its result is 1 (Cosine 0=1).

3) Euclidean distance: Euclidean distance (ED) is another geometrical measure used to measure similarity of two documents. Each document is represented as a point in space based on term frequency of n terms (representing n dimension). ED computes the difference between two points in n dimensional space based on their coordinate using following equation:

$$\mathrm{ED}(A,B) = \sqrt{\sum_{n=1}^{k} \big(A(n) - B(n)\big)^2} \qquad (2)$$

Using ED the highest similarity between two vectors happens when they are plotted in the same point in space and ED between them is 0. Overall, among geometric measures on tasks which are involved in text similarity, cosine was the best measure (Turney et al. 2010).

IV. Motivation

The main purpose of this study is to measure the similarities among documents with high accuracy in such a way that one hopes to better understand which documents are more similar (or less similar). We call this concept *similarity level* and *accuracy* refers to the power of a measure in differentiating the similarity level among documents in such a way that one can understand which documents are less, more and most similar. This power of differentiation between more or less similarity can be significantly useful for recommendation systems and clusterings. In future works, the *similarity level* can be used to identify top n similar documents to a particular topic (search query) or a document. Moreover, as K-Means algorithm (explained in Section VI-B3) uses the distance among data points to detect the closest points and form the clusters, high *accuracy* in measuring similarities helps not only to cluster all similar data points together in one cluster, but cluster most similar data points to one cluster and less similar ones to another. The results in Table VIII which will be explained later, supports this idea. In this section we explain why existing VSM measures are not robust enough to measure the similarity level accurately.

A. Boolean results

One of the limitations associated to VSM is Boolean values (Wong et al. 1985). The Boolean values which are more witnessed in Cosine Similarity, gives a generic view over similarity among documents. For instance, Cosine gives value 1 when two vectors (documents) are overlapped (regardless of the vectors' magnitudes) and shows two documents are similar, but does not show how similar two documents are in terms of a desired topic/searched query. We will clarify this drawback with an example in Section IV-B.

B. Cosine Drawbacks

Cosine similarity is a powerful method for measuring the difference between two documents based on their orientations but not their magnitudes. Hence magnitude of vectors which is dealing with term frequency does not play any role in this similarity measurement. Figure 1a shows two critical situations wherein Cosine similarity's weaknesses may produce inaccurate similarity results toward vectors' magnitudes. Based on Cosine similarity metric, the similarity between document A and document B, document A and document C and finally between document A and document D are equal. Despite the identical angel between them, there is a huge difference between the vectors' magnitudes. As A and B have the higher proportion of the terms in their texts, it seems A and B are more dedicated to the terms from searched query than C and D. Hence A has higher similarity to B, less similarity to C and least similarity to D. On the other hand, based on Cosine similarity, document B has the same similarity of 1 (Boolean value) to document C and document D, while it is obvious that B and D are less similar because there is a longer distance between B and D so the difference is higher. Hence Cosine similarity cannot be that proper and accurate method for measuring similarity level between vectors.

C. Euclidean Distance drawbacks

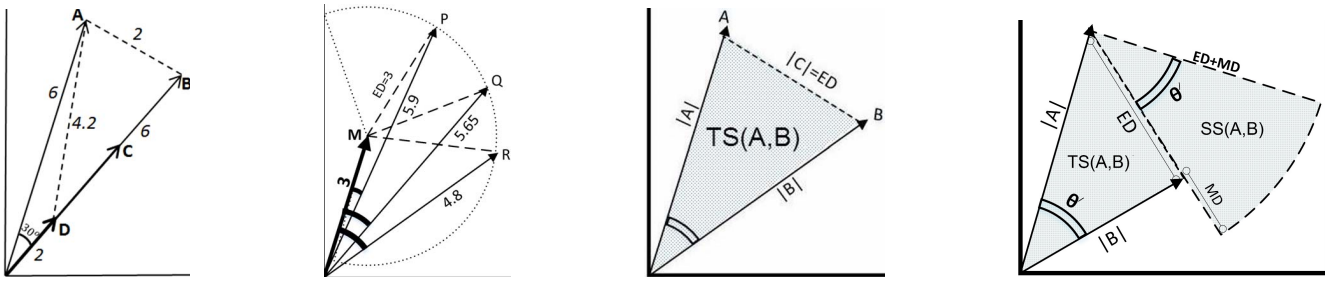
Figure 1b shows the main ED's drawback clearly. As it can be seen many vectors such as P, Q and R can be drawn from M with the same ED and despite the huge difference between them, ED shows P, Q and R have got the same similarity of 3 to M.

V. Novel Geometric Similarity Measurement Method

As it was mentioned earlier, the similarity measurement has to take underlying semantic features into consideration in order to reflect the meaningful results (Strehl et al. 2000). In the other word, by using the magnitude of vectors as an influential and powerful tool, we can compute similarity among documents more accurate. In order to interpolate the magnitudes into similarity measurement, it is required to include more parameters than the angel and ED between vectors. In this research, a new algorithm called TS-SS computes the similarity between vectors from two divers prospective and generates the similarity value between two vectors not only from the angel and ED between them, but also the difference between their magnitudes (Becker & Kuropka 2003). The model is examined on different datasets to prove its accuracy and robustness in clustering and measuring similarity level.

A. Triangle's Area Similarity (TS)

By looking at vectors in Figure 1c, it is obvious that a triangle can be formed as the ED is drawn between them. As



(a) (b) (c) (d)

Figure 1: (a)Example of Cosine drawback. (b)Example of ED drawback. (c)Triangle Similarity (TS). (d)Triangle Similarity-Section Similarity (TS-SS)

the ED between two vectors decreases and they get closer to each other (so the angel between them gets tighter also), the area of the triangle decreases. This decrement has the inverse relation to similarity between two documents.

1) Calculation and strength of TS: By calculating the area of the triangle between two vectors and using it as similarity metric, in fact three characteristics have been included in computing the similarity, namely angle between vectors, magnitude of vectors and ED. We call this method TS (Triangle's Area Similarity) and present it as TS(A,B) for two vectors of A and B. We use SAS (Side Angle Side) formula to calculate the area of the triangle. SAS formula uses the length of two sides and the Sin of included angle. To that end, it is required to measure the magnitude of the vectors and the angle between them: $|A| = \sqrt{\sum_{n=1}^{k} A_n^2}$, $|B| = \sqrt{\sum_{n=1}^{k} B_n^2}$ where n is dimension of vector (x,y,z,...). In order to compute the angle between vectors, first Equation 1 is used to get V (the cosine value of θ), then using the following cosine rule, the new angle (θ') is calculated:

$$\theta' = \cos^{-1}(V) + 10 \tag{3}$$

The idea of forming the triangle fails when vectors are overlapped like B and C in Figure 1a. To overcome this problem, we increased theta by 10 degree. We used 10 degrees as minimum to keep it rounded and the reason to do not use any value less than 10 is to avoid more decimal points and keep our calculations simple.

$$\mathrm{TS}(A,B) = \frac{|A| \cdot |B| \cdot \sin(\theta')}{2} \tag{4}$$

Using this approach, as two vectors get closer to each other and the area of triangle between them decreases, proportionally the dissimilarity decreases and the similarity increases and the maximum similarity is met when two vectors with the same length are overlapped. Unlike ED, using TS method, the measured similarity of M and P, is different with similarity of M and Q in Figure 1b. That is, TS(M,P) $<$ TS(M,Q) which is an expected result as P is more similar to Q. As the area of triangles among pairs of vectors vary due to differences in their magnitudes and angles between them, TS generates different similarity values.

2) Weakness of TS: TS accuracy fails in Figure 1a. Obviously A and B look more similar as they have the same magnitude, while D is smaller than A and ED(A,D) is approximately three times more than ED(A,B). Despite these obvious distance differences, it can be seen TS(A,D) $<$ TS(A,B) and it conveys that A is more similar to D rather than to B which is not a true conclusion.

B. Sector's Area Similarity (SS)

Merely TS alone is not robust enough to interpolate vectors' differentiations precisely to produce accurate similarity results due to missing components. By looking at Figure 1a, it can be understood that one of the missing components is the difference between the vectors' magnitudes. In this study the magnitude difference between two vectors is called MD and represented as:

$$\mathrm{MD}(A,B) = \left| \sqrt{\sum_{n=1}^{k} A_n^2} - \sqrt{\sum_{n=1}^{k} B_n^2} \right| \tag{5}$$

MD has a one way direct relation to ED. That is, as MD increases, proportionally ED increases. Combination of ED and MD might be helpful but in order to leverage their effects on similarity values we need the angle between vectors which is called angular difference and represented as AD(A,B) = θ' for two vector of A and B. AD can be calculated using Equation 3. Like the relation between MD and ED, AD and ED also has the one way direct relation. In the other word, in all cases, increasing and decreasing the angle between vectors effects on ED directly. That is why AD is another important component in measuring similarity. ED and MD can be combined by summation and AD can accentuate the power of this combination by forming a circular section with a radius of ED+MD length (Figure 1d). By measuring the area of the formed section, we can calculate the similarity between two vectors from another

prospective. This similarity is called SS (Section's Area Similarity) and computed using the following formula:

$$\mathrm{SS}(A,B) = \pi \cdot (\mathrm{ED}(A,B) + \mathrm{MD}(A,B))^2 \cdot \left(\frac{\theta'}{360}\right) \quad (6)$$

C. TS-SS method

TS and SS complete each other and that is the reason we combine them by multiplying them together. The range of TS-SS measure is from 0 to ∞. The reason for choosing multiplication but not summation to combine TS and SS is that in some cases the value of TS and SS are disproportionate where one is extremely larger then the other one. For example, in Figure 1a, TS similarity is too big (TS(A,B)=5.91) while SS similarity is too small (SS(A,B)=0.047). If we use summation, they can not effect on each other significantly to give the realistic similarity value and we get TS(A,B)+SS(A,B)=5.95 $>$ TS(A,D)+SS(A,D)=2.96 which is a false result because A is more similar to B as discussed earlier. But if we use multiplication we get TS(A,B)$\cdot$ SS(A,B)=0.27 $<$ TS(A,D)$\cdot$SS(A,D)=1.71 which is a true result. Using TS-SS method, similarity of 0 happens only when ED=MD=0 and it shows two vectors are absolutely identical in term of direction and magnitude which indicates the maximum similarity between two documents. This novel similarity method is called TS-SS and is presented as following:

$$\frac{|A| \cdot |B| \cdot \sin(\theta') \cdot \theta' \cdot \pi \cdot \big(\mathrm{ED}(A,B) + \mathrm{MD}(A,B)\big)^2}{720} \quad (7)$$

VI. Experiments

In this section first we briefly describe the four datasets used in this study. Then we will explain the evaluation models and finally we compare the three metrics based on the results of evaluations. We assume that end user enters three keywords in search engine and is interested in finding the clusters of similar documents where each document contains at least one of searched query's keyword in a selected dataset (the measures accept more then three keywords obviously, but three is chosen for this study to keep the calculations and explanations simple). Based on this assumption, at first we apply text preprocessing mentioned in Section III-A on all texts in each dataset and create a list of top 600 keywords with high TF-IDF score from each benchmark. Then after we create twenty search queries, each consisted of three keywords picked from the list of top keywords randomly. Finally we create Cosine similarity matrix, ED similarity matrix and TS-SS similarity matrix for each search query in each dataset (For future works, we will measure the similarity among documents based on top n keywords with highest TF-IDF values from the entire corpus). In order to evaluate and show the correctness of TS-SS measure, we compare similarity matrices from the three geometrical metrics by running four different evaluations. Finally we represent a test case including four documents derived from Classic4 and show the significance of some drawbacks mentioned in Section IV in real world.

A. Datasets

The four chosen datasets are labeled by their topics manually and are widely used for data classifications, text categorization and clustering.

1) 20 News Group: This dataset contains 20 different categories and each category has 1000 newsgroup documents (*The 20 Newsgroups data set* 2008). For this dataset we have computed the similarity values for 20 search queries.

2) 7 sectors: This dataset consists of classified documents from seven industrial sections, and each section has around 7 subsections (*CMU World Wide Knowledge Base Project* 2011). Totally there are 44500 labeled documents in this dataset. For this dataset, we have computed the similarity values for 10 search query.

3) WebKB: This dataset is the collection of web pages collected from computer science department of diverse universities in 1997 and manually classified into seven classes (*WebKB* 2010). This database contains 8334 documents. For this dataset we have computed the similarity values for 10 search queries.

4) Classic4: The Classic4 dataset contains 7095 labeled documents from abstract of scientific papers in four divers categories (*Classic3 and Classic4 DataSets* 2010). For this dataset we have computed the similarity values for 10 search queries.

B. Evaluations

After computing similarity values for all search queries in each dataset using Cosine, ED and TS-SS metrics, we compare these three metrics by using four different evaluations namely Uniqueness, Number of booleans, Minimum gapscore and Purity.

1) Uniqueness: The purpose of this evaluation is to compute the percentage of unique values in each similarity matrix. The outcome of this evaluation indicates the existence possibility of drawbacks mentioned in Section IV. When we have more unique similarity values, it means the measure is robust to recognize the similarity level among documents even when there is a small difference among documents, rather than generating same similarity values among different documents.

2) Number of booleans: One of the main issues with VSM is generating many boolean similarities (Wong et al. 1985) which highly impact on measuring similarity level. This evaluation counts the number of boolean values in the similarity matrices generated by each measure and shows how many percent of values are booleans. The main purpose of this evaluation is to show which measure produces more boolean values and gives less variation. In fact this is another attempt to support the mentioned drawbacks in Section IV.

3) Purity: Purity is widely used for measuring the quality of clusters based on the label of documents in each cluster. Hence first we need to cluster the documents.We cluster documents based on K-Means algorithm using WEKA. K-Means (*KMeans Clustering* 2015) is a supervise and nondeterministic algorithm which clusters data points into K clusters based on K given centroids and for different number of seeds generates different clusters. K is given based on the number of classes in each dataset and in order to get the robust result, we run the K-Means algorithm with diverse number of seeds for 100 times on all search queries' similarity matrices and compute the average as the final result. Finally to measure the purity of each cluster each cluster is assigned to the category which is the most frequent in the cluster. Then the assignment accuracy of cluster C_i which contains n_i documents is measured by the following formula (*Evaluation of clustering* 2009):

$$P(C_i) = \frac{d_i}{n_i} \tag{8}$$

where d_i is the number of documents from the dominant category in cluster C_i.

4) Minimum Gapscore: In this test we compare the similarity matrices with their associated oracle. As the documents are labeled, constructing the oracle matrix is achievable.

Minimum Boolean Gap Score

Begin
Input: `real` similarity[n][n]; `bool` oracle[n][n]
Sim = **vector** [];
curGapScore = 0
For $j = 1$ **to** $n - 1$ **do**
  For $k = j + 1$ **to** n **do**
    Sim.append(**pair**(similarity[j][k],oracle[j][k]))
    If oracle[j][k]==1 **then**
      curGapScore = curGapScore+1
Sort Sim **by decreasing order** key1
then increasing order key2
minGapScore = curGapScore
For each (s,b) **in** Sim **do**
  If b==1 **then**
    curGapScore = curGapScore-1
  else
    curGapScore = curGapScore+1
  If curGapScore $<$ minGapScore **then**
    minGapScore = curGapScore
Return minGapScore
End

For a fixed set of documents, let S be a Boolean Oracle that returns $S[i, j] = 1$ (= true) if and only if document i is similar to document j. For a similarity measure α we define its *gap score* (with respect to S) as

$$\min_{B_c} \sum_{1 \le i < j \le n} \big|\, B_c(\alpha(i,j)) - S[i,j] \,\big|,$$

$$\text{where } B_c(x) = \begin{cases} 1 & \text{if } x \ge c \\ 0 & \text{if } x < c \end{cases}$$

Furthermore, for two α_1 and α_2 we define a quasi order $\alpha_1 \le_S \alpha_2$ if the gap score of α_1 is at most the gap score of α_2.

5) Test case: We selected four documents from Classic4 which every one has at least one of the keywords in the search query (randomly generated) of *"alveolar aneurysm car"*. For simplicity we have changed the name of the documents to A, B, C and D and the text preprocessing (e.g. stemming, removing stopwords and noisy data) has applied on text belonging to each document in Table III. Based on the similarity values, keyword frequencies and number of words in each document, we show the robustness of each measure in computing similarity levels.

VII. Results

As explained earlier, different number of search queries have been used to compute the similarities in each dataset. That is, for each search query, we have three similarity matrices namely Cosine, ED and TS-SS similarity matrix. We have applied the evaluation techniques on all matrices and the results shown in all tables represent the average of results. Also the results presented in percentage in Table I and Table II indicate in average, how many percent of similarity values are unique and boolean respectively.

Table I: Uniqueness Results

Dataset	Cosine	ED	TS-SS
20NewsGroup	2.19%	88.13%	**88.92%**
7sector	2.48%	**97.92%**	**97.92%**
WebKB	1.37%	99.50%	**99.51%**
Classic4	4.79%	**98.14%**	**98.14%**

Table II: Number of booleans

Dataset	Cosine	ED	TS-SS
20NewsGroup	98.71%	0.089%	**0.087%**
7sector	99.64%	**0.2%**	**0.2%**
WebKB	99.61%	**0.12%**	**0.12%**
Classic4	99.87%	**0.44%**	**0.44%**

As Table I shows Cosine similarity has the very low percentage of uniqueness and it conveys that the drawbacks mentioned about this model is a significant issue. This low accuracy in recognizing the differences among similarity values causes the lack of distinguishment between documents with higher similarities and documents with lower similarities. Although intangible difference between ED uniqueness and TS-SS uniqueness indicates that the ED

Table III: The test case from Classic 4

Doc.ID	Text
A	acut experiment pneumococc type pneumonia mous migrat leucocyt pulmonari capillari alveolar space reveal electron microscop preliminari studi experiment pneumococc pulmonari pneumonia mous leucocyt observ pass capillari interstiti tissu eventu alveolar space intercellular junction endotheli epitheli cell membran
B	light electron microscop studi develop respiratori tissu rat light microscop observ develop rat lung shown presenc glandular canalicular alveolar stage stage identifi electron microscopi present differ part lung e g 40 45 mm c r length glandular stage lung tissu immatur appear light microscopi electron microscopi individu cell immatur respect organell glycogen present immatur cell canalicular stage lung tissu vascular stage develop duct air space line continu complet epithelium blood vessel complet endothelium lamel inclus bodi present epitheli endoderm cell earli stage develop micropinocytot vesicl present larg number epitheli endotheli cytoplasm suggest foetus indic absorpt amniot fluid alveolar space mechan alveolar distens discuss natur remain uncertain respiratori tissu rat fulli differenti birth import fact human infant discuss 10 adult blood air barrier consist epithelium zona diffusa endothelium vari thick project perform whilst receipt grant medic research council canada gratitud express gratitud express miss sylvia smith type manuscript
C	pathogenesi viral influenz pneumonia mice pathogenesi influenz pneumonia mice studi electron microscopi mice inocul ld pr8 influenza virus kill vari interv inocul observ light microscopi correl electron microscopi order evalu lesion produc peripheri earliest lesion focal area edema alveolar line cell capillari endothelium interpos basement membran caus appreci thicken blood air pathway hypertrophi degener desquam alveolar line prolifer alveolar macrophag result complet consolid progress week infect central area lung affect somewhat differ day infect noncili bronchiolar cell show consider hyperplasia endoplasm reticulum apic cytoplasm edema viral particl matur lumen surfac cell releas bronchiolar lumen bronchiolar cell ciliat noncili underw degener slough bronchiolar lumen regener epithelium stratifi surfac cell elong flatten peribronchiolar interstiti tissu gradual total infiltr cell mononuclear type
D	role alveolar inclus bodi develop lung develop alveolar epithelium man rat contain characterist inclus bodi heterogen structur basic consist membran profil limit membran unit type inclus bodi appear result focal cytoplasm degrad occur rapid chang cuboid alveolar epithelium inclus bodi develop rat lung similar call lamellar transform mitochondria evid present suggest alter cytoplasm membran involv process inclus bodi format certain imag associ golgi complex interpret earli form inclus bodi evid inclus bodi enlarg accret membran final extrud alveolar space inclus bodi form secret greater number late fetal life earli infanc e cuboid alveolar epithelium differenti matur flatten type contain inclus bodi basi morpholog characterist inclus bodi distribut acid phosphatas reaction conclud inclus bodi lysosom structur activ remodel develop alveolar epithelium possibl interrelationship inclus bodi pulmonari surfact discuss

Table IV: Term frequency (TF), total number of words (W) and their proportion (TF/W) in each document of the test case

Doc.	W	TF of *"alveolar "*	TF/W
A	37	2	0.05
B	147	3	0.02
C	119	3	0.02
D	127	6	0.04

Table V: TS-SS similarity values for the test case

Doc.	A	B	C	D
A	0	4.00E-5	3.68E-05	3.76E-06
B	4.00E-5	0	3.96E-07	2.26E-05
C	3.68E-05	3.96E-07	0	1.91E-05
D	3.76E-06	2.26E-05	1.91E-05	0

drawback mentioned in Section IV-C is not a critical issue in this research, it could be a critical issue in larger datasets.

Table II shows that in overall, cosine generated around 99% boolean results while the other two metrics did not do so for the same documents. In general the higher percentage

Table VI: ED similarity values for the test case

Doc.	A	B	C	D
A	0	0.198112	0.168760	0.038742
B	0.198112	0	0.029352	0.159370
C	0.168760	0.029352	0	0.130017
D	0.038742	0.159370	0.130017	0

Table VII: Cosine similarity values for the test case

Doc.	A	B	C	D
A	1	1	1	1
B	1	1	1	1
C	1	1	1	1
D	1	1	1	1

of uniqueness and number of boolean values in ED and TS-SS supports the claim that many document pairs which have the same similarity values based on cosine metric, are not exactly same, therefore we believe cosine is not robust enough to distinguish similarities in high level.

Table VIII and Figure 2 represent the most significant results in comparing the three metrics. As the Figure shows, ED is the weakest model for clustering. In our biggest dataset, 20News, TS-SS outperforms Cosine with a significant difference, while in other datasets TS-SS outperforms Cosine slightly. In fact in the small datasets, there are few types of documents and the chance that documents of the same type get clustered together is higher than the condition where there are several types of documents like twenty types of documents in 20News dataset. Therefore, the significant better result of TS-SS in 20News dataset justifies the robustness and reliability of the model for big data and real world data where the variety of documents/texts are high. As mentioned earlier, due to stochastic outcome of K-Means, we run the algorithm 100 times and selected the best purity result of each measure shown in Table VIII. In

order to show the consistency of the results over 100 runs, the standard deviation of all 100 purities of each dataset for each measure is shown in the same table.

Table VIII: The results of purity test and the standard deviation of purity values.

	Purity			Standard deviation		
Dataset	Cosine	ED	TS-SS	Cosine	ED	TS-SS
20News	0.46	0.47	**0.86**	0.043	0.036	**0.033**
7Sector	0.63	0.69	**0.75**	0.054	**0.027**	0.056
WebKB	0.83	0.74	**0.85**	0.063	0.091	**0.038**
Classic4	0.92	0.80	**0.95**	0.097	0.120	**0.086**

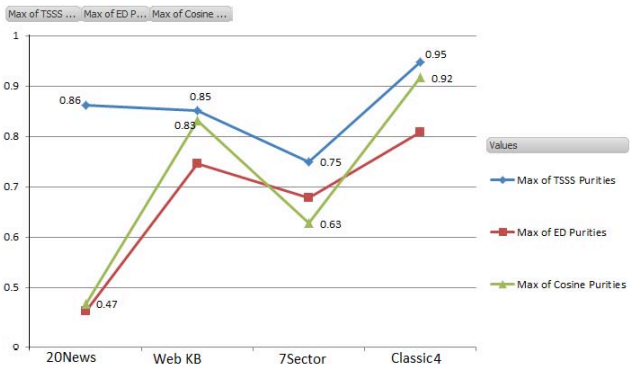


Figure 2: Purity values

Table IX shows the overall minimum gap score of each measure. Based on our definition in Section VI-B4, the less gap score means more similarity to oracle. The result shows the only significant outperformance belongs to TS-SS in 20NewsGroup and in other datasets results are approximately same for all measures.

Table IX: Minimum Gapscore

Dataset	Cosine	ED	TS-SS
20NewsGroup	61876.25	61962.2	**61682.1**
7sector	**6083.3**	6085.7	6085.7
WebKB	**32065.44**	32066.22	32066.22
Classic4	21335	21336	**21334.7**

As Table III shows, each document has only one of the keywords in the search query and the common keyword is *"alveolar"*. By looking at proportion of term frequency (TF) to total number of words (W) column in Table IV, we notice that document A has the highest similarity to document D and it's similarity to B and C is equal, but the numbers in column W, indicate A is slightly more similar to C rather than to B in term of number of words. Based on the similarity values measured by Cosine in Table VII, it can be concluded that all four documents are equally similar to each other, and no similarity level is measured(boolean values). Unlike Cosine, Ed and TS-SS measure the similarity levels as expected. As shown in Table V and Table VI, Sim(A,D) $>$ Sim(A,C) $>$ Sim(A,B) where Sim(A,B), Sim(A,C) and Sim(A,D) represents the similarity between A and B, A and C and A and D respectively. This result shows when the mentioned drawbacks about ED does not exist, ED is as powerful as TS-SS in measuring the similarity level, but its poor results in purity shows it is not as reliable as TS-SS for document clustering.

VIII. Time Complexity

For n documents, there are n^2 relationships. For each relationship the similarity can be measured using α (Cosine, ED and TS-SS). Let $Q = \{t_1, \ldots, t_k\}$ be the set of k terms (keywords) where the searched query is a subset. Measuring similarity for each measure can be done in $O(k)$ and need to process $O(n^2)$ combinations of n documents, so for all relationships the similarity can be computed in $O(n^2\ k)$ for all three measures.

For measuring the uniqueness and number of booleans, we take the entries above the main diagonal only. For n users there are $\frac{n^2-n}{2}$ elements above the diagonal. Hence counting the unique values and booleans each can be done in $O(n^2)$.

K-Means algorithm clusters n documents in $O(i\ c\ n^2\ k)$ where i is the number of iteration, c is the number of clusters (number of clusters is equal to number of categories in each similarity matrix) and k is the vector dimension (number of terms). In computing purity values, group the documents of the same cluster is done in $O(n)$ in worse case and finding the dominant category in all clusters is done in $O(c\ log\ n)$ if we have more than one cluster, and $O(n)$ if we have only one cluster. Finally calculating purity is done in constant time of $O(c)$. Overall, computing the purity is done in $O(n)$.

For finding the minimum gap score, the sorting algorithm is computed in $O(n^2\ log\ n)$ and it dominates the other steps of the algorithm.

IX. Conclusion

The main purpose of this study is to measure the similarity level among documents accurately, in such a way that one hopes to better understand which documents are more (or less) similar. *accuracy* refers to the power of a measure in differentiating the similarity level among documents in such a way that one can understand which documents are least, less, more and most similar. This power of differentiation can be significantly useful for recommendation systems and clusterings. In this study, first we represent documents as vectors using VSM technique. Then we propose a new method called TS-SS to measure the similarity level among documents for desired keyword(s) based on the geometrical similarities. By looking at popular geometrical similarity methods such as Euclidean distance and Cosine similarity we notice, in some special conditions, they are not robust enough to identify the similarity level. TS-SS helps one to understand which documents are more

(or less) similar by taking more specific geometrical criteria. We apply Euclidean distance, Cosine similarity and our new method on four labeled datasets. Finally we report how these three geometrical similarity measures perform in terms of similarity level and clustering purity using four evaluation techniques. The evaluations' results show that our new model outperforms the other measures. TS-SS clusters documents with better purity and is more reliable for measuring the similarity level.

ACKNOWLEDGEMENT

We thank Dr. Yun Sing Koh for assistance and support on an early version of the manuscript. We would also thank Prof. Jim Warren for comments that greatly improved the manuscript.

REFERENCES

Becker, J. & Kuropka, D. (2003), Topic-based vector space model, *in* 'Proceedings of the 6th International Conference on Business Information Systems', pp. 7–12.

Bigi, B. (2003), *Using Kullback-Leibler distance for text categorization*, Springer.

Chim, H. & Deng, X. (2008), 'Efficient phrase-based document similarity for clustering', *Knowledge and Data Engineering, IEEE Transactions* **20**(9), 1217–1229.

Classic3 and Classic4 DataSets (2010). http://www.dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets/

CMU World Wide Knowledge Base Project (2011). http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb

Cross, R., Borgatti, S. P. & Parker, A. (2002), 'Making invisible work visible: Using social network analysis to support strategic collaboration', *California management review* **44**(2), 25–46.

Das, S., Egecioglu, Ö. & El Abbadi, A. (2009), 'Anonymizing edge-weighted social network graphs', *Computer Science, UC Santa Barbara, Tech. Rep. CS-2009-03* .

Elsayed, T., Lin, J. & Oard, D. W. (2008), Pairwise document similarity in large collections with mapreduce, *in* 'Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers', Association for Computational Linguistics, pp. 265–268.

Evaluation of clustering (2009). http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

Hammouda, K. M. & Kamel, M. S. (2002), Phrase-based document similarity based on an index graph model, *in* 'Proceedings of the 2002 IEEE International Conference on Data mining', IEEE, pp. 203–210.

Heidarian, A. (2011), Multi-clustering users in twitter dataset, *in* 'International Conference on Software Technology and Engineering, 3rd (ICSTE 2011)', ASME Press.

Jaccard, P. (1912), 'The distribution of the flora in the alpine zone. 1', *New phytologist* **11**(2), 37–50.

KMeans Clustering (2015). http://home.deib.polimi.it/matteucc/Clustering/

Kornbrot, D. (2005), 'Pearson product moment correlation', *Encyclopedia of Statistics in Behavioral Science* .

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* pp. 79–86.

Laboreiro, G., Sarmento, L., Teixeira, J. & Oliveira, E. (2010), Tokenizing micro-blogging messages using a text classification approach, *in* 'Proceedings of the fourth workshop on Analytics for noisy unstructured text data', ACM, pp. 81–88.

Lakkaraju, P., Gauch, S. & Speretta, M. (2008), Document similarity based on concept tree distance, *in* 'Proceedings of the nineteenth ACM conference on Hypertext and hypermedia', ACM, pp. 127–132.

Lebanon, G. (2006), 'Metric learning for text documents', *Pattern Analysis and Machine Intelligence, IEEE Transactions* **28**(4), 497–508.

Nelson, M. L., Bollen, J., Calhoun, J. R. & Mackey, C. E. (2004), User evaluation of the nasa technical report server recommendation service, *in* 'Proceedings of the 6th annual ACM international workshop on Web information and data management', ACM, pp. 144–151.

Porter, M. F. (1980), 'An algorithm for suffix stripping', *Program: electronic library and information systems* **14**(3), 130–137.

Sahami, M. & Heilman, T. D. (2006), A web-based kernel function for measuring the similarity of short text snippets, *in* 'Proceedings of the 15th international conference on World Wide Web', AcM, pp. 377–386.

Salton, G. & Buckley, C. (1988), 'Term-weighting approaches in automatic text retrieval', *Information processing & management* **24**(5), 513–523.

Singhal, A., Buckley, C. & Mitra, M. (1996), Pivoted document length normalization, *in* 'Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 21–29.

Strehl, A., Ghosh, J. & Mooney, R. (2000), Impact of similarity measures on web-page clustering, *in* 'Workshop on Artificial Intelligence for Web Search (AAAI 2000)', pp. 58–64.

The 20 Newsgroups data set (2008). http://qwone.com/~jason/20Newsgroups/

Turney, P. D., Pantel, P. et al. (2010), 'From frequency to meaning: Vector space models of semantics', *Journal of artificial intelligence research* **37**(1), 141–188.

Wan, X. & Peng, Y. (2005), The earth mover's distance as a semantic measure for document similarity, *in* 'Proceedings of the 14th ACM international conference on Information and knowledge management', ACM, pp. 301–302.

Wartena, C. & Brussee, R. (2008), Topic detection by clustering keywords, *in* '19th International Workshop on Database and Expert Systems Application, 2008. DEXA'08.', IEEE, pp. 54–58.

WebKB (2010). http://www.csmining.org/index.php/webkb.html

Wong, S. M., Ziarko, W. & Wong, P. C. (1985), Generalized vector spaces model in information retrieval, *in* 'Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 18–25.

Wu, W., Zhang, B. & Ostendorf, M. (2010), Automatic generation of personalized annotation tags for twitter users, *in* 'Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, pp. 689–692.

Zar, J. H. (1998), 'Spearman rank correlation', *Encyclopedia of Biostatistics* .

Zhang, T., Tang, Y. Y., Fang, B. & Xiang, Y. (2012), 'Document clustering in correlation similarity measure space', *Knowledge and Data Engineering, IEEE Transactions* **24**(6), 1002–1013.