Text Classification and Clustering through Similarity Measures

Mirza Ruhi Masuma[1], Prof.V.A.Losarwar[2],

[1]*PG Student, Computer Science & Engg. P.E.S. College of Engineering Aurangabad, India*

[2]*Associate Professor, Computer Science & Engg. P.E.S. College of Engineering Aurangabad, India*

***Abstract*—Similarity measurement is the important process in text processing. It measures the similarities between the two documents. Unlabeled document collections are becoming increasingly large and common and available; mining such data sets is a major contemporary challenge. Words are used as features. Text documents are often represented as high-dimensional and sparse vectors. Measuring the similarity between words, sentences, paragraphs and documents is an important component in various tasks such as retrieval of information, document clustering, word-sense disambiguation, automatic essay scoring, short answer grading, and text summarization. This paper shows the survey of the document clustering.**

***Index Terms*—Text processing, document clustering, similarity measure.**

I. INTRODUCTION

Text processing plays a vital role in information retrieval, web search and data mining, information retrieval, text classification, document clustering, topic tracking, topic detection, questions generation, question answering, essay scoring, short answer scoring, machine conversion, text summarization and others[11]. In text processing, the model bag-of-words is commonly used. The bag of word model is widely used in text mining and information retrieval. Words are counted in the bag, which be different from the mathematical definition of set. Each word is equivalent to a dimension in the resulting data space and every document then becomes a vector which consist of non-negative values on each dimension. Today we are facing an ever increasing volume of text documents. The large numbers of texts flowing over the Internet, huge collections of documents are stored in digital libraries and repositories forms, and digitized personal information such as emails are piling up quickly every day.

These have brought great challenges for the effective and efficient organization of text documents. A document can be defined as any content drawn up or received by the Foundation concerning a matter relating to the policies, decisions falling and activities within its competence and in the framework of its official tasks, in whatever medium (either written on paper or which stored in electronic form , including e-mail, or as a sound, visual or audio-visual recording).A document is represented as a vector in which each component indicates the value of the corresponding feature in the document. The feature value can be term frequency (number of times the term appearing in the document), relative term frequency (it is the ratio between the term frequency and the total number of occurrences of all the terms which is present in the document set), or tf-idf (it is a combination of term frequency and inverse document frequency)[1].

Finding similarity between words is an essential part of text similarity which is then used as a first stage for sentence, document similarities and paragraph. Words can be similar in two ways lexically and semantically. Words are said to be similar lexically if they have a similar character sequence. Words can be said are similar semantically if they have same thing, are opposite of each other, used in the same way, used in the same context and one is a type of another [11].

II .LITERATURE SURVEY

Similarity measures have been largely used in text classification and clustering algorithms. Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee [1] proposed a new measure for determining the similarity between two documents. Several characteristics are embedded in this measure. It is a symmetric measure. The difference between absence and presence of a feature is considered more important than the difference between the values associated with a present feature. The similarity is decreased when the number of absence-presence features increases. The spherical k means algorithm introduced by Dhillon and Modha adopted the cosine similarity measure for document clustering. Zhao and Karypis showed results of clustering experiments with seven clustering algorithms and twelve different text data sets, and showed that the objective function based on cosine similarity it leads to the best solutions irrespective of the number of clusters for most of the data sets. D'hondt et al.[2] adopted a cosine-based pairwise adaptive similarity for clustering of documents. Zhang et al. [3] used cosine similarity to calculate a correlation similarity between two documents in a low-dimensional semantic space and performed clustering of documents in the correlation similarity measure space. Kogan et al. proposed a two step clustering procedure. In this the sPDDP is used to generate initial partitions in the initial step

and a k-means clustering algorithm using the Kullback-Leibler deviation is applied in the second step. Dhillon et al. proposed a discordant information-theoretic feature clustering algorithm for text classification using the Kullback-Leibler divergence. Muhammad Rafi, Mohammad Shahid Shaikh [4] proposed a novel similarity measure based on topic maps representation of documents. Jayaraj Jayabharathy and Selvadurai Kanmani [5] in their papers shows how the emphasis of the work is Dynamic document clustering which is based on Term frequency and Correlated based Concept algorithms, using semantic-based similarity measure. Pallavi J. Chaudhari1 and Dipa D. 1Dharmadhikari [6] show that Multi-viewpoint based similarity measure (MVS) is more suitable for text documents than the popular cosine similarity measure. Lan Huang[7] developed a novel method for learning an inter-document similarity measure from human judgment. The measure predicts similarity more consistently with average human raters than human raters do between themselves, and also outperforms the current state of the art on a standard dataset. Alok Sharma, Sunil PranitLal [8] introduced Tanimoto based similarity measure for host based intrusions using binary aspect set for training and classification. The k-nearest neighbor (kNN) classifier has been utilized to classify a given process as either normal or attack. Gaddam Saidi Reddy and Dr.R.V.Krishnaiah [9] approach in finding similarity between documents or objects while clustering is multi view based similarity. Measures such as Euclidean, cosine, Jaccard, and Pearson correlation are compared. The conclusion made is that Euclidean and Jaccard are best for web document clustering. Their computational complexity is very high which is the drawback of these approaches. Venkata Gopala Rao and S. Bhanu Prasad A[10] shows the document clustering can be applied using concept space and cosine similarity. They found that except the Euclidean distance measure, the additional measures have comparable expected effect for the partitioned text document clustering task.

Many measures have been proposed for computing the similarity between two vectors. The Kullback-Leibler divergence is said to be a non-symmetric measure of the difference between the probability distributions which is related with the two vectors. Euclidean distance is a well-known similarity metric which is taken from the Euclidean geometry field. Manhattan distance like to Euclidean distance and is also known as the taxicab metric is another similarity metric.

A. Character-Based Similarity Measures

Longest Common Sub String (LCS) is an algorithm that considers the similarity between two strings is based on the length of adjacent chain of characters that exist in both the strings.

Damerau-Levenshtein defines that the distance between two strings by counting the minimum number of operations needed to transform from one string into the other, where an 654 operation is defined as a deletion,insertion or substitution of a single character, or a transposition of two adjacent characters.

Jaro is based on the number and order of the common characters between two strings. It takes into consideration typical spelling deviations and it is mainly used in the area of record linkage.

Jaro–Winkler is an extension of Jaro distance. Jaro-Winkler uses a prefix scale which gives more positive ratings to strings that go with from the beginning for a set prefix length.
Needleman-Wunsch algorithm is an example of dynamic programming, and was the very first application of dynamic programming to the biological sequence comparison. It performs a global arrangement in a straight line to find the best alignment over the two sequences. It is appropriate when the two sequences are of similar length, with a important degree of similarity throughout.

Smith-Waterman is another example of dynamic programming. It performs a spatial alignment to find the best alignment over the preserved domain of two sequences. It is useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context.

N-gram is a sub-sequence of n items from a given sequence of text. N-gram similarity algorithms compare the n-grams from each character or word in two strings. The Distance is computed by dividing the number of like n-grams by maximal number of n-grams.

B. Knowledge-Based Similarity

Knowledge-Based Similarity is one of a semantic similarity measures that is based on identifying the degree of similarity between words using information which is derived from semantic networks. WordNet is one of the most popular semantic network in the area of Knowledge-Based similarity between words WordNet is a large lexical database of English where verbs adjectives nouns and adverbs are all grouped into sets of perceiving synonyms(synsets) each expressing a distinct concept. Synsets are interlinked by means of theoretical-semantic and lexical relations.

C. Phrase-Based Document Clustering

Techniques of Document clustering generally rely on single term analysis of the document data set, such as the Vector Space Model. To get more errorless document clustering, more informative features including phrases and their weights are primarily important in such scenarios. Document clustering is particularly useful in lots of applications such as automatic categorization of documents, building a taxonomy of documents, grouping search engine results, and others.

D. Critical Analysis

Character-Based measures operate on character sequences and character composition.

Knowledge-Based similarity is one of semantic similarity measures that is based on identifying the degree of similarity among words using information derived from semantic networks.

Phrase-based similarity measure is capable of accurate calculation of pair-wise document similarity. To achieve more correct document clustering, more informative features including phrases and their weights are mainly important in such scenarios.

Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee presented a novel similarity measure between two documents.Several desirable properties are embedded in this measure and concluded that the similarity measure they proposed is better than they achieve by other measure. The similarity measure is shown below.

III. SIMILARITY MEASURE

Consider a document d with m features w1, w2, . . . , wm be represented as an m-dimensional vector, i.e., d = <d1, d2, . . . , dm >. If wi, $1 \leq i \leq m$, is not present in the document then di= 0. Otherwise, di>0. The following properties among other ones are desirable for a similarity measure between two documents[1]

The absence or presence of a feature is necessary than the difference between two values associated with a present feature. Here we consider two features wi and wj and two documents d1 and d2.

Let wi does not appear in d1 but it does appears in d2, then wi have no relationship with d1 while it has some relationship with d2.

If case d1 and d2 are dissimilar in terms of wi. And if wj appears in both document d1 and d2 then wj has some relationship with d1 and d2 simultaneously. Here in this case d1 and d2 are similar to some degree in terms of wj. For the above two cases it is reasonable to say that wi carries more weight than wj in determining the similarity degree between documents d1 and d2.

Lets assume that wi is absent in d1 i.e., d1 i = 0 but appears in d2 e.g., d2i = 2 and wj appears both in d1 and d2 e.g., d1j = 3 and d2 j = 5. Then wi is considered to be more essential than wj in determining the similarity between document d1 and d2 although the differences of the feature values in both cases are the same.

The similarity degree should increase when the difference between two values (that are non zero) of a specific feature decreases. For example the similarity that is involved with d13 = 2 and d23 = 15 should be smaller than that involved with d13 = 2 and d23 = 4.

The similarity degree should decline when the number of absence-presence features increases. For a presence-absence feature of d1 and d2, d1 and d2 are unalike in terms of this feature as commented earlier. We consider for example, the likeness between the documents < 1, 0, 1 > and < 1, 1, 0 > should be smaller than that between the documents < 1, 0, 1 > and < 1, 0, 0 >.

Two documents are considered to be least similar to each other if none of the features have non-zero values in both documents. Let d1 = < d11, d12, . . . , d1m > and d2 = < d21, d22, . . . , d2m >. If

d1 id2i = 0,
d1i + d2i > 0

for $1 \leq i \leq m$. Then d1 and d2 are least similar to each other.

Similarity measure should be symmetric. The similarity degree between d1 and d2 should be same as that between d2 and d1.

The standard deviation of the feature is taken into account for its input to the similarity between two documents feature with a superior spread offers more involvement to the similarity between d1 and d2.

A SIMILARITY BETWEEN TWO DOCUMENTS

The similarity measure, called SMTP (Similarity Measure for Text Processing), for two documents d1 = < d11, d12, . . . , d1m > and d2 = < d21, d22, . . . , d2m >. It defines a function F as follows:

$$F(d_1, d_2) = \frac{\sum_{j=1}^{m} N_*(d_{1j}, d_{2j})}{\sum_{j=1}^{m} N_\cup(d_{1j}, d_{2j})} \text{ ……..eq(1)}$$

Then our proposed similarity measure, for d1 and d2 is

$$s_{SMPT}(d_1, d_2) = \frac{F(d_1, d_2) + \lambda}{1 + \lambda} \text{ ……eq(2)}$$

This similarity measure takes into account following three cases: a) The feature considered should appears in both documents. b) the feature considered should appears in only one document and c) the feature considered should appears in none of the documents. For the first case, set a lower bound 0.5 and reduce the similarity as the difference between the feature values of the two documents increases. For the second case then set a negative constant $-\lambda$ which pays no attention to the magnitude of the non-zero feature value. For the last case, the feature has no involvement to the similarity.

IV. CONCLUSIONS

In this survey four text similarity approaches were discussed; Character based, Knowledge-Based, Phrase-Based Document Clustering and SMTP (Similarity Measure for Text Processing). Knowledge-Based similarity is one of semantic similarity measures. Character-Based measures operate on character sequences and character composition. Phrase-based similarity measure is capable of errorless calculation of pair-

wise document similarity. Several desirable properties are embedded in Similarity measure for text processing. For example, the similarity measure is symmetric. The presence or absence of a feature is considered more important than the difference between the values associated with a present feature. The similarity degree increases when the number of absence-presence features pairs decreases. Two documents are least similar to one another if none of the features have non-zero values in both documents. The similarity measure is much more better than that other measures.

REFERENCES

[1]. 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, 2013.

[2]. J. D'hondt, J. Vertommen, P.-A. Verhaegen, D. Cattrysse, and J. R. Duflou, "Pairwise-adaptive dissimilarity measure for document clustering," *Inf. Sci.*, vol. 180, no. 12, pp. 2341–2358, 2010.

[3]. T. Zhang, Y. Y. Tang, B. Fang, and Y. Xiang, " Document clustering in correlation similarity measure space," *IEEE Trans. Knowl. DataEng.*, vol. 24, no. 6, pp. 1002–1013, Jun. 2012.

[4]. Muhammad Rafi, Mohammad Shahid Shaikh, " An improved semantic similarity measure for document clustering based on topic maps,"Computer Science Department, NU-FAST, Karachi Campus, Pakistan, 2013.

[5]. Jayaraj Jayabharathy and Selvadurai Kanmani, " Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature", Decision Analytics, Springer open journal, Puducherry, India, 2014.

[6]. Pallavi J. Chaudhari1, Dipa D. Dharmadhikari, " Clustering With Multi-Viewpoint Based Similarity Measure: An Overview," International Journal of Engineering Inventions ISSN: 2278-7461, www.ijeijournal.com Volume 1, Issue 3, pp 01-05, September 2012.

[7]. Lan Huang, "Learning a Concept-based Document Similarity Measure",Department of Computer Science, University of Waikato, New Zealand,2012.

[8]. Alok Sharma, Sunil PranitLal, "Tanimoto Based Similarity Measure for Intrusion Detection System", *Journal of Information Security*, 2, 195-201,2012.

[9]. GaddamSaidi 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.

[10]. VenkataGopalaRao S. Bhanu Prasad A," Space and Cosine Similarity measures for Text Document Clustering", International Journal of Engineering Research & Technology (IJERT)Vol. 2 Issue 2, February- 2013

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