ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

Feature Values Analysis for Similarity Measure to Text Classification and Clustering

Somnath Mahadev Melasagare[1], Anand B. Bone[2]

P.G. Student, Department of Computer Engineering, SKNSIT's, Lonavala, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, SKNSIT's, Lonavala, Maharashtra, India[2]

ABSTRACT: The similarity between documents are the new innovative concept now days in data mining and information retrieval. These include mainly sponsored search, query reformulation and image retrieval. Standard text similarity measures perform poorly because of data sparseness and the lack of context. Where Document processing plays an important role in data mining, and web search. In text processing, bag-of-words model is used. Measuring the similarity between documents is a main task in the document processing and text classification. In this, a new similarity measure is proposed. To measure the similarity between documents with respect to a feature, the proposed method takes the following cases: a) The feature must be in both documents, b) the feature that appears in only one document, and c) the feature that appears in none of the documents. For first case, similarity increases as the difference between the documents feature values decreases. For second, a fixed value is find out the similarity. For last case, the feature has no contribution to similarity. The effectiveness of measure is evaluated on several real-world data sets for document classification and clustering.

KEYWORDS: document classification; document clustering

I. INTRODUCTION

Text Mining is a method of extracting information from different text documents resources. The goal of text mining is to discover unknown Information. where The challenges that arises due to unstructured text documents are large textual database; all publications are also in electronic form, Very high number of possible word and phrases in the language, Complex and subtle relationships between concepts in text. Information Extraction (IE) is the method of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. This activity relate to processing human language texts by means of natural language processing (NLP).

Document or Text clustering is an automatic document organization, topic extraction or filtering[2]. It is closely related to data clustering. Clustering methods can be used to group the retrieved documents into a list of meaningful categories. clustering involves use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users.

The feature value can be term frequency (the number of occurrences of a term appearing in the document), relative term frequency (the ratio between the term frequency and the total number of occurrences of all the terms in the document set), or tf-idf (a combination of term frequency and inverse document frequency). Usually, the dimensionality of a document is large and the resulting vector is sparse, i.e., most of the feature values in the vector are zero. Such high dimensionality and sparsity can be a severe challenge for similarity measure which is an important operation in text processing algorithms[9].

In this paper, Proposing a new measure for computing the similarity between two documents. Several characteristics are embedded in this similarity measure technique. It is a symmetric 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 increases as the difference between the two values associated with a present feature decreases. Furthermore, the contribution of the difference is normally scaled. The similarity decreases when the number of presence-absence features increases. An absent feature has no contribution to the similarity. The proposed measure is

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

extended to gauge the similarity between two sets of documents. The measure is applied in several text applications, including single label classification, multi-label classification, *k*-means like clustering, and hierarchical agglomerative clustering, and the results obtained demonstrate the effectiveness of the proposed similarity measure.

II. RELATED WORK

Similarity measures extensively used in text classification and clustering. Euclidean distance is usually the default choice of similarity-based methods, e.g. *k*-NN and *k*-means algorithms where the extended Jaccard coefficient can be used for document data and it reduces to the Jaccard coefficient in the case of binary attributes.

1. Dhillon and Modha, "Concept decompositions for large sparse text data using clustering,"
They proposed a spherical *k* means algorithm[10], which we adopted the cosine similarity measure for document clustering. the spherical k means algorithm is used for the document clustering which uses the concept of cosine similarity.

2. D'hondt *et al*., "Pairwise-adaptive dissimilarity measure for document clustering,"
In previous method, for the document clustering the concept of similarity between documents is used but in this pairwise adaptive dissimilarity measure is used for document clustering where this paper adopted a cosine-based pairwise adaptive similarity for document clustering[3].

3 Chim *et al*., "Efficient phrase-based document similarity for clustering"
They performed [11] document clustering based on the proposed phrase-based similarity measure. this technique of document clustering becomes efficient because of phrase based.

III. EXISTING METHOD

In Euclidean distance two documents d1 and d2 represented by term vectors t_a and t_b respectively, the euclidean distance given simply as

$$D_E(\overrightarrow{t_a}, \overrightarrow{t_b}) = (\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2)^{\frac{1}{2}}$$

where the term set is $T = t_i, ..., t_m$.

Similarity Measure is defined as the root of square differences between the respective coordinates of d1 and d2,

$$d_{Euc}(\mathbf{d}_1, \mathbf{d}_2) = [(\mathbf{d}_1 - \mathbf{d}_2)\cdot(\mathbf{d}_1 - \mathbf{d}_2)]^{1/2}$$

Cosine similarity is a measure of similarity between two vectors used for analyzing the text or string or document similarity

- The similarity scales between 0 and 1(maximum)
- The bigger the return value is, the more similar two texts are
- Cos 0^0 =1 , 0^0 means that two texts are equal
- Cos 90^0=0, 90^0 means that two texts are totally different
- Example

('H','e','l','l','o') and ('H','e','l','l','o') Degree bet these two 0

('H','e','l','l','o') and ('x','y') Degree bet these two increasing to 90 as it differs

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

IV.PROPOSED SYSTEM

In this work, a document **d** with *m* features *w*1, *w*2, . . . , *wm* be represented as an *m*-dimensional vector, i.e., **d** = < *d*1, *d*2, . . . , *dm* >. If *wi*, $1 \leq i \leq m$, is absent in the document, then *di* = 0. Otherwise, *di* > 0. The following properties, among other ones, are preferable for a similarity measure between two documents:

Similarity between Two Documents

Based on the preferable properties mentioned above, we propose a similarity measure, called SMTP (Similarity Measure for Text Processing), for two documents **d**1 = < *d*11, *d*12, . . . , *d*1*m* > and **d**2 = < *d*21, *d*22, . . . , *d*2*m* >. Define a function *F* as follows:

$$F(d1, d2) = \frac{\sum_{j=1}^{m} \mathrm{N}_*(\mathrm{d}_{1\mathrm{j}}, \mathrm{d}_{2\mathrm{j}})}{\sum_{j=1}^{m} \mathrm{N}_{\mathrm{U}}(\mathrm{d}_{1\mathrm{j}}, \mathrm{d}_{2\mathrm{j}})}$$

where

$$N_*(d_{1j}, d_{2j}) = \begin{cases} 0.5\left(1 + \exp\left\{-\left(\frac{d_{1j}-d_{2j}}{\sigma_j}\right)^2\right\}\right), \\ \quad \text{if } d_{1j}d_{2j} > 0 \\ 0, \text{ if } d_{1j} = 0 \text{ and } d_{2j} = 0 \\ -\lambda, \text{ otherwise,} \end{cases}$$

$$N_{\cup}(d_{1j}, d_{2j}) = \begin{cases} 0, \text{ if } d_{1j} = 0 \text{ and } d_{2j} = 0 \\ 1, \text{ otherwise.} \end{cases}$$

Then proposed similarity measure, *SSMTP*, for **d**1 and **d**2 is

$$SSMTP(\mathrm{d1}, \mathrm{d2}) = \frac{F(d1, d2) + \lambda}{1 + \lambda}$$

where proposed measure takes into account the following three cases:

a) The feature considered appears in both documents,

b) the feature considered appears in only one document, and

c) the feature considered appears in none of the documents.

For first case, setting lower bound 0.5 and decrease the similarity as the difference between the feature values of the two documents increases, scaled by a Gaussian function where σj is the standard deviation of all non-zero values for feature *wj* in the training data set. For second case, we set a negative constant $-\lambda$ disregarding the magnitude of the non-zero feature value. For last case, the feature has no contribution to the similarity.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

V. METHODOLOGY

System Architecture

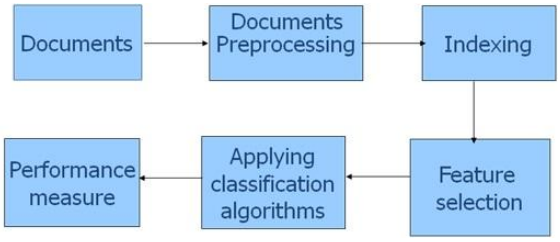

Fig. 1 System Architecture

In above Fig. 1 System Architecture is given which includes several modules for analysis of similarity measures between documents and clustering process for documents.

1. Documents Representation

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. Recently, the bag-of-words model has also been used for computer vision.

- Each word corresponds to a dimension in the resulting data space
- Each document then becomes a vector consisting of non-negative values on each dimension these values are corresponds to feature
- Let D = {d1,d2,d3,d4 ....} be the set of documents & T = {t1,t2,t3,t4....tm} be the set of unique terms in D
- A document is then represented as a m dimensional vector, td = (tf(d, t1), . . . , tf(d, tm)))

2. Documents Preprocessing

Many words are not informative and thus irrelevant are remove from the document representation. (e.g.) the, a, an, and, there, their, is, was, were, where, etc. These word typically about 400 to 500. It is used to improve the efficiency and potential problems of removing stop words[8].

Reducing words from their root form. A document may contain several occurrences of words like fish, fishes and fishers. An advantage of stemming is to improve the effectiveness to match similar words. Reduce indexing size to combing words with same roots may reduce indexing size as much as 40-50%. Porter algorithm is used to stem the words[8].

Porter Algorithm:

Step 1: Gets rid of plurals and -ed or -ing suffixes
Step 2: Turns terminal y to i when there is another vowel in the stem
Step 3: Maps double suffixes to single ones:
-ization, -ational, etc.
Step 4: Deals with suffixes, -full, -ness etc.
Step 5: Takes off -ant, -ence, etc.
Step 6: Removes a final –e

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

3. Indexing

The Extracted text documents are converted into Boolean weighting by using the indexing technique of Term Frequency – Inverse Document Frequency. TF–IDF is the product of two statistics, term frequency and inverse document frequency. The **term frequency** tf (*t*, *d*), the simplest choice is to use the raw frequency of a term in a document, i.e. the number of times that term *t* occurs in document *d*. The raw frequency of *t* by f (*t*, *d*), then the simple tf scheme is tf (*t*, *d*) = f (*t*, *d*). Other possibilities include

- Boolean "frequencies": tf (*t*, *d*) = 1 if *t* occurs in *d* and 0 otherwise;
- Logarithmically scaled frequency: tf (*t*, *d*) = log (f (*t*, *d*) + 1);

The **inverse document frequency** is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

TF–IDF is calculated as tf – idf (t, d, D) = tf (t, d) × idf (t, D)

4. Feature Selection

- The similarity of two patterns and feature selection are the critical factors affecting the performance of classifier, especially for classification based on pattern similarity theory
- Feature selection is a technique of selecting subset of relevant features for building robust learning model

5. Applying classification algorithm

- In this randomly selected training documents are used for training/validation and the testing documents are used for testing
- The predesignated training data are used for training/validation and the predesignated testing data are used for testing
- Note that the data for training/validation are separate from the data for testing in each case

6. Performance Measure

- The effectiveness of measure will be evaluated on several real-world data sets for text classification and clustering problems
- The results will show that the performance obtained by the proposed measure is better than that achieved by other measures

VI. CONCLUSION

In this work, the aim has been to investigate and discover efficient technique for similarity measure by comparing different techniques. with this in mind the Future research in the data mining similarity measure will strive towards improving the computational speed, accuracy and precision.

REFERENCES

[1] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering*" IEEE Trans. Knowl. Data Eng.*, VOL. 26, no. 7, July 2014.
[2] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2005.
[3] 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.
[4] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA, USA: Elsevier, 2006.
[5] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1217–1229, Sept. 2008.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

[6] A. Strehl and J. Ghosh, "Value-based customer grouping from large retail data-sets," in *Proc. SPIE*, vol. 4057. Orlando, FL, USA, Apr. 2000, pp. 33–42.
[7] T. Zhang, Y. Y. Tang, B. Fang, and Y. Xiang, "Document clustering in correlation similarity measure space," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1002–1013, Jun. 2012.
[8] D. Renukadevi , S. Sumathi "TERM BASED SIMILARITY MEASURE FOR TEXT CLASSIFICATION AND CLUSTERING USING FUZZY C-MEANS ALGORITHM" International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 4, April 2014
[9] P. K. Agarwal and C. M. Procopiuc, "Exact and approximation algorithms for clustering," in *Proc. 9th Annu. SODA*, Philadelphia, PA, USA, 1998, pp. 658–667.
[10] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Mach. Learn.*, vol. 42, no. 1, pp. 143–175, 2001.
[11] H. CHIM AND X. DENG, "EFFICIENT PHRASE-BASED DOCUMENT SIMILARITY FOR CLUSTERING," *IEEE TRANS. DATA ENG.*, VOL. 20, PP. 1217–1229, SEPT. 2008.