International Engineering Journal For Research & Development

Vol.3
Issue 2

E-ISSN NO:-2349-0721

Impact factor : 2.25

A NOVEL APPROACH FOR TEXT SIMILARITY MEASURE AND CLASSIFICATION

[1]Ms. Bhawna Gayakwad
MTech. Computer Science & Engineering
SBITM College of Engineering, Betul, India
Email: bhawnagayakwad@gmail.com

[2]Dr. S. D. Choudhari
SBITM College of Engineering
Betul, India
Email: choudhari.sachin1986@gmail.com

ABSTRACT:

In the text processing field finding the similarity between multiple documents is an important operation. In this paper, we proposed a new similarity measure for document clustering. To figure out the similarity between multiple documents with respect to a feature, our proposed similarity finding measure takes the following cases into account:

1) The selected feature may appear in both documents, 2) the selected feature appears in only one document, and 3) the selected feature appears in none of the documents. In the first case, the documents similarity actually increases as the difference between the selected involved features values are less. Moreover, the involvement of the difference is normally scaled by feature values. However in the second case, a constant value is involved to find the similarity and in the last case, the selected feature are absent between the documents and thus has no contribution to the document similarity. Our proposed measure is extended to estimate the appropriate similarity between two document sets to get effective results with better performance.

Keywords—Document similarity, document clustering, entropy, accuracy, clustering algorithms.

1 INTRODUCTION

Text document processing plays an important role in information retrieval, data mining, as well as web search [27], [31]. In document processing, the bag-of-words model is frequently used [26], [29]. A document is typically represented as a vector of selected feature in which each component indicates the value of the corresponding feature in the document. The feature value can be term frequency, relative term frequency (i.e. 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) [25].Typically, the dimensionality of a document is very 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 strict challenge for similarity measure which is an important operation in text processing algorithms . A bunch of similarity measures have been proposed for computing the similarity between two vectors. Cosine document similarity [25] is a measure taking the cosine of the angle between two feature vectors. The Kullback-Leibler divergence is a non-symmetric measure of the

difference between the probability distributions associated with the two vectors[35]. Euclidean distance is a well-known document text similarity metric taken from the Euclidean geometry field [45]. Manhattan distance is also similar to Euclidean distance and also known as the taxicab metric, is another similarity metric[45]. The Canberra distance metric [45] is used in situations where nonnegative elements are present in vector. The Jaccard coefficient [21] is used for comparing the similarity of two sample document sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. The Hamming distance [21], [22] between two feature vectors is the number of positions at which the corresponding symbols are different. The extended Jaccard coefficient and the Dice coefficient retain the sparsity property of the cosine similarity measure while allowing discrimination of collinear vectors.

IT-Sim , An information-theoretic measure for document set similarity, was proposed in [8]. Chim et al. [11] also proposed a phrase-based similarity measure to compute the Suffix Tree Document (STD) model. Similarity measures have been widely used in text classification and clustering algorithms.

We propose a document similarity measure for finding the similarity between two documents. Multiple characteristics are involved in this measure. It is a symmetric measure. Here the difference between presence and absence of a selected key feature is considered more essential than the difference between the values associated with a present feature. The similarity value increases as the difference between the two values associated with a present feature decreases. Also, the contribution of the difference is generally scaled. The similarity decreases when the number of presence and absence feature increases. An absent feature has no contribution to the document similarity.

Our proposed measure is progressively extended to determine the similarity between two sets of documents appropriately. The similarity measure is going to applied in several text document set, and apply k-means clustering to demonstrate the accuracy and effectiveness of the proposed similarity measure.

IEJRD
E-ISSN NO:2349-0721

RELATED WORKS

There are some similarities measures which have been commonly adopted for finding the similarity between two documents are described here.

Consider td1 and td2 be two documents represented as feature vectors. The Euclidean distance [45] measure is defined as the root of square differences between the respective coordinates of td1 and td2, i.e.

dEuc(td1, td2) = [(td1 − d2)•(td1 − td2)]1/2 ----------(1)

where A•B denotes the inner product of the two vectors A and B. Cosine similarity [25] measures the cosine of the angle between td1 and td2 as follows:

SCos(td1, td2) = td1•td2 (td1•td1)1/2(td2•td2)1/2----------(2)

Pairwise-adaptive similarity [17] dynamically selects a number of features out of td1 and td2 and is defined to be

dPair(td1, td2) = td1,K • td2,K (td1,K•td1,K)1/2(td2,K•td2,K)1/2---------- (3)

where tdi,K is a subset of tdi, i = 1, 2, containing the values of the features which are the union of the K largest features appearing in td1 antd td2, respectively.

The extended Jaccard coefficient [48], [49] is an extended version of the Jaccard coefficient [21] for data processing:

SEJ(td1, td2) = td1•td2 td1•td1 + td2•td2 − td1•td2----------(4)

PROPOSED SIMILARITY MEASURE

The presence or absence of a key feature is more essential than the difference between the two values associated with a present feature. Consider two key features wi and wj and two documents td1 and td2. Suppose wi does not appear in td1 but it appears in td2. Then wi is considered to have no relationship with td1 while it has some relationship with d2. In this case, td1 and td2 are dissimilar in terms of wi. If wj appears in both td1 and td2. Then wj has some relationship with td1 antd td2 simultaneously. In this case, td1 and td2 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 td1 and td2. For example, assume that wi is absent in td1, i.e., d1i = 0, but appears in td2, e.g., d2i = 2, and wj appears both in td1 and td2, e.g., td1j = 3 and td2j = 5. Then wi is considered to be more essential than wj in determining the similarity between d1 and td2, although the differences of the feature values in both cases are the same.

2) The similarity degree should increase when the difference between two non-zero values of a specific feature decreases. For example, the similarity involved with td13 = 2 and td23 = 20 should be smaller than that involved with td13 = 2 and td23 = 3.

3) The similarity degree should decrease when the number of presence-absence features increases. For a presence absence feature of td1 and td2, td1 and td2 are dissimilar in terms of this feature as commented earlier. Therefore, as the number of presence-absence features increases, the dissimilarity between td1 and td2 increases and thus the similarity decreases. For example, the similarity 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 >.

Then our proposed similarity measure, SSMTP , for td1 and d2 is

SSMTP (td1, td2) = F(td1, td2) + λ1 + λ

Algorithm:

Step1: Find document set feature key words based on term frequency value

Step2: Create feature document vector based on feature words find through step1

Step3: Input the different document set to create document vector for each document based on feature document vector.

Step4: Map and save document vector value to database for further similarity measure

Step5: Calculate document similarity measure based on document vector value to depict similarity between two documents

Step6: cluster the documents based on its similarity value

CONCLUSION

We have proposed a similarity measure between two documents or two document set. Here multiple key properties of document are considering the presence or absence of a feature is more essential than the difference between the similarities values associated with a current document feature. The document similarity degree increases when the number of presence and absence features pair's are less. Two documents are actually least similar to each other if none of the key features have non-zero values in both documents. Besides, it is desirable to consider the value distribution of a feature for its contribution to the similarity between two documents.

The proposed methodology has also been extended to measure the similarity between two documents set. To improve the competence, we have provided an approximation to reduce the complexity involved in the computation. The λ values used are selected between range 0.6∼1.0, can be set for applications where many features appear commonly in the documents being compared, and a small value of λ, e.g., 0.01∼0.0001, can be set for applications where many features appear in one document but not in other documents. In this work, we are focusing on the efficiency and performance for document similarity measures in different classification/clustering algorithms.

REFERENCES:

[1] http://web.ist.utl.pt/ acardoso/datasets/.

[2] http://www.cs.technion.ac.il/ ronb/thesis.html.

[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[4] http://www.dmoz.org/.

[5] P. K. Agarwal and C. M. Procopiuc. Exact and approximation algorithms for clustering. Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 658–667, 1998.

[6] D. W. Aha. Lazy learning: Special issue editorial. Artificial Intelligence Review, 11(1-5):7–10, 1997.

[7] G. Amati and C. J. V. Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems, 20(4):357–389, 2002. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

[8] J. A. Aslam and M. Frost. An information-theoretic measure for document similarity. Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pages 449–450,2003.

[9] G. H. Ball and D. J. Hall. A clustering technique for summarizing multivariate data. Behavioral Science, 12:153–155, 1967.

[10] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. IEEE Transactions on Knowledge and Data Engineering, 17(12):1624–1637, 2005.

[11] H. Chim and X. Deng. Efficient phrase-based document similarity for clustering. IEEE Transactions on Knowledge and Data Engineering, 20(9):1217 – 1229, 2008.

[12] S. Clinchant and E. Gaussier. Information-based models for ad hoc IR. Proceedings of 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 234–241, 2010.

[13] M. Craven, D. DiPasquo, D. Freitag, A. K. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge form the world wide web. Proceedings of 15th National Conference on Artificial Intelligence, 1998.

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

[15] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. Machine Learning, 42(1):143–175, 2001.

[16] I. S. Dhillon, S. Mallela, and R. Kumar. A Divisive infromation-theoretic feature clustering algorithm for text classification. Journal of Machine Learning Research, 3:1265–1287, 2003.

[17] J. D'hondt, J. Vertommen, P.-A. Verhaegen, D. Cattrysse, and J. R. Duflou. Pairwise-adaptive dissimilarity measure for document clustering. Information Sciences, 180:2341–2358, 2010.

[18] R. O. Duda, P. E. Hart, and D. J. Stork. Pattern Recognition. Wiley, 2001.

[19] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Science, 95(25):14863–14868, 1998.

[20] H.'Fang, T. Tao, C. Zhai. A formal study of heuristic retrieval constraints. Proceedings of 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 49–56, 2004.

[21] C. G. Gonz´alez, W. B. Jr., and A. L. V. Rodrigues. Density of closed balls in real-valued and autometrized boolean spaces for clustering applications. 19th Brazilian Symposium on Artificial Intelligence, pages 8–22, 2008.

[22] R. W. Hamming. Error detecting and error correcting codes. Bell System Technical Journal, 29(2):147–160, 1950.

[23] K. M. Hammouda and M. S. Kamel. Efficient phrase-based document indexing for web document clustering. IEEE Transactions on Knowledge and Data Engineering, 16(10):1279 – 1296, 2004.

[24] K. M. Hammouda and M. S. Kamel. Hierarchically distributed peer-topper document clustering and cluster summarization. IEEE Transactions on Knowledge and Data Engineering, 21(5):681 – 698, 2009.

[25] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann, Elsevier, 2006.

[26] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. International Conference on Machine Learning, pages 143–151, 1997.

[27] T. Joachims and F. Sebastiani. Guest editors' introduction to the special issue on automated text categorization. Journal of Intelligent Information Systems, 18(2/3):103–105, 2002.

[28] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7):881–892, 2002.

[29] H. Kim, P. Howland, and H. Park. Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research, 6:37–53, 2005.

[30] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng. Some effective techniques for naive bayes text classification. IEEE Transactions on Knowledge and Data Engineering, 18(11):1457 – 1466, 2006.

[31] K. Knight. Mining online text. Communications of the ACM, 42(11):58–61, 1999.

[32] J. Kogan, C. Nicholas, and V. Volkovich. Text mining with information theoretic n clustering. Computing in Science and Engineering, 5(6):52–59, 2003.

[33] J. Kogan, M. Teboulle, and C. K. Nicholas. Data driven similarity measures for k-means like clustering algorithms. Information Retrieval, 8(2):331–349, 2005.

[34] S. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. Seventh Annual European Symposium on Algorithms, pages 362–371, 1999.

[35] S. Kullback and R. A. Leibler. On information and sufficiency. Annuals of Mathematical Statistics, 22(1):79–86, 1951.

[36] S.-J. Lee and C.-S. Ouyang. A neuro-fuzzy system modeling with selfconstructing rule generation and hybrid SVD-based learning. IEEE Transactions on Fuzzy Systems, 11(3): 341–353, 2003.

[37] V. Lertnattee and T. Theeramunkong. Multidimensional text classification for drug information. IEEE Transactions on Information Technology in Biomedicine, 8(3):306 – 312, 2004.

[38] D. D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, 5:361–397, 2004.

[39] D. Lin. An information-theoretic definition of similarity. Proceedings of 15th International Conference on Machine Learning, 1998.

[40] M. G. Michie. Use of the bray-curtis similarity measure in cluster analysis of foraminiferal data. Mathematical Geology, 14(6):661–667, 1982.

[41] T. Mitchell. Machine Learning. McGraw-Hill, 1997.

[42] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. Machine Learning, 39(2/3):103– 134, 2000.

[43] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Learning to classify text from labeled and unlabeled documents. Proceedings of 15th National Conference on Artificial Intelligence, 1998.

[44] G. Salton and M. J. McGill. Introduction to Modern Retrieval. McGraw-Hill Book Company, 1983.

[45] T. W. Schoenharl and G. Madey. Evaluation of measurement techniques for the validation of agent-based simulations against streaming data. International Conference on Computational Science, 2008

IEJRD
E-ISSN NO:2349-0721