Review on Document Clustering using Various Similarity Measures

Komal Gothwal
Assistant Professor,
IT Dept Atharva College of Engineering
Mumbai, Maharashtra, India

Varsha Salunkhe
Assistant Professor,
IT Dept Atharva College of Engineering
Mumbai, Maharashtra, India

***Abstract*—Data mining is a powerful technology whichfocuses on the most important information in the data collected. Various Data mining techniques are used for knowledge discovery from databases. One of the most important data mining or text mining technique is Clustering. Clustering is used to group similar objects together to find necessary structures in data, and organize them into meaningful subgroups. This article discusses about various similarity measure techniques used for clustering. The novel method varies from the existing one as it uses only single view point for clustering whereas Multi-Viewpoint based similarity measure uses many different viewpoints. In MultiViewpoint the two objects measured are assumed not to be in the same cluster. In Hierarchical Clustering with Multiple view points, we use two measures for intercluster and intracluster relation between objects. The former clustering process focuses on partitional clustering of multi viewpoint documents, which are not focused on sparse and high dimensional data.**

Keywords— Document Clustering, Similarity Measure, Text Mining, K-MeanClustering Algorithm, Hierarchical Methods, High Dimensional Data

I. INTRODUCTION

Data mining is the process of extracting the implicit, new and likely useful information from raw data. Data mining techniques are used in a many research areas including Financial Data analysis, cybernetics, Biological data analysis, Telecommunication industry and marketing. Document clustering organizes documents into various groups called as clusters or groups, where the documents in each cluster share some general properties according to specified similarity measure. Document clustering algorithms play a vital role in helping users to navigate, encapsulate and organize the information efficiently.

Due to tremendous growth of accessing information from the web, easy access and knowledge of information are needed critically. The text processing plays a vital role in information retrieval, data mining, and web mining. Text mining attempts to discover new useful information by applying techniques from data mining. Clustering, one of the conventional data mining techniques is an unsupervised learning method where clustering technique try to recognize intrinsic subgroups of the text documents, so that a set of clusters is produced in which clusters show high intra-cluster similarity and low inter-cluster similarity. Generally, text document clustering methods attempt to separate the documents into groups where each group represents some area that is different than those areas represented by the other groups.

II. LITERATURE SURVEY

Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee [1] proposed a novel measure for computing the similarity between two documents. Several features are fixed in this measure. It is a symmetric measure. The difference between occurrence and absence of a feature is considered more important than the difference between the values related with a present feature. The similarity decreases when the number of occurrence- absence features increases. An absent feature has no role to the similarity. The similarity increases as the variation between the two values related with a present feature decreases. This work primarily focuses on textural features. Furthermore, the contribution of the variation is normally scaled. To improve the effectiveness, they have provided an estimate to reduce the complexity involved in the computation. The results have shown that the performance obtained by the proposed measure is better than that achieved by other measures.

GaddamSaidi Reddy and Dr.R.V.Krishnaiah [2] method in finding similarity between text documents or objects while performing clustering is multi-view based similarity. All similarity measures such as cosine, Euclidean, Jaccard, and Pearson correlation are taken under comparison. The conclusion made here is that Ecludean and Jaccard are finest for web document clustering. They both preferred associated afeatures for given subject and calculated distance between two values. Both of them used an algorithm known as Hierarchical Agglomerative Clustering technique in order to perform clustering. Their computational complexity is very high that is the disadvantage of these approaches. Proposed a similarity measure known as MVS (Multi-Viewpoint based Similarity), when it is compared with cosine similarity, MVS is more functional for finding the similarity of text documents. The experimental results and analysis revealed that the proposed system for similarity measure is proficient and it can be used in the real time applications in the text mining area. It makes use of more than one point of reference as contrasting to existing techniques used for clustering text documents. Shady Shehata, FakhriKarray and Mohamed S. Kamel [3] noticed that the most of the common techniques in text mining are based on the statistical analysis of a term, either word or

phrase. Statistical analysis of a term frequency captures the significance of the term within a document only. Text mining model should specify terms that capture the semantics of text. The mining model can capture terms that present the concepts of the sentence, which leads to finding of the topic of the document. The mining model that studies the terms on the sentence, document, and corpus levels are introduced, can effectively differentiate between non important terms with respect to sentence semantics and terms. The term which adds to the sentence semantic is analyzed on the sentence, document, and corpus levels rather than the conventional analysis of the document only.

It is essential to make a note of extracting the relations between verbs and their arguments in the same sentence has the probability for analyzing terms within a sentence. The information about who is doing what to whom clarifies the role of each term in a sentence to the meaning of the main topic of that sentence. It is shown that the standard deviation is enhanced by using the concept-based mining model.

Anna Huang [4] stated that prior to clustering, a similarity distance measure must be determined. The measure shows the degree of closeness or separation of the target objects and should match up to the characteristics that are supposed to discriminate the clusters fixed in the data. It is very complex to perform a methodical study comparing the impact of similarity metrics on cluster quality, because impartially evaluating cluster quality is complex in itself.

The clusters, which are produced in an unverified way, are judge against the pre-defined category structure, which is usually produced by human experts. This sort of evaluation believes that the aim of clustering is to replicate human thinking, so a clustering solution is good if the clusters are reliable with the manually created categories.

It is found that there is no measure that is best across the world for all kinds of clustering problems. The outcome of the cosine similarity, Jaccard correlation and Pearson's coefficient are very close in terms of performance, and are notably better than the Euclidean distance measure experimented with the web page documents.

Hung Chim and Xiaotie Deng [5] noticed that the phrase has been considered as a more helpful feature term for improving the effectiveness of document clustering. They planned a phrase-based document similarity to compute the pairwise similarities of documents based on the Suffix Tree Document (STD) model. By mapping each node in the suffix tree of STD model into a distinctive feature term in the Vector Space Document (VSD) model, the phrase-based document similarity logically inherits the term tf-idf weighting scheme in computing the document similarity with phrases. They applied the phrase-based document similarity to the group- average Hierarchical Agglomerative Clustering (HAC) algorithm technique and developed a new document clustering method. Their evaluation researches specifies the new clustering method is valuable on clustering the documents of two standard document benchmark corpora OHSUMED and RCV1.Finally they found that both the traditional VSD model and STD model play vital roles in text-based information retrieval. The theory of the suffix tree and the document similarity are quite easy, but the execution is complicated. Analysis is required to improve the performance of the document similarity. They conclude that the feature vector of phrase terms in the STD model can be considered as an advanced feature vector of the traditional single-word terms in the VSD model.

YanhongZhai and Bing Liu [6] studied the problem of removing data from a Web page that contains several ordered data records. The purpose is to segment these data records, extract data items/fields from them and put the data in a database table. They projected approach to extract structured data from Web pages. Although the problem has been studied by many researchers, existing methods are either incorrect or make many strong assumptions.

Jacob Kogan, Marc Teboulle and Charles Nicholas [7] argue that the choice of a particular similarity measure may get better clustering of a specific dataset. They called this choice the ―data driven similarity measure‖. They analysed that the overall complexity of huge data sets encourages application a sequence of algorithms for clustering a single data set. Their results of numerical research show, however, that the best clustering results can be obtained for intermediate parameter values.

Inderjit Dhillon, Jacob Kogan& Charles Nicholas [8] found that in particular, when the processing task is to partition a given document collection into clusters of similar documents a choice of good features along with good clustering algorithms is of paramount importance. Feature or term selection along with a number of clustering strategies. The selection methodologies considerably decreases the dimension.

Syed Masum Emran and Nong Ye [9] stated that distance metric value can be used to find the similarity or dissimilarity of the present observation from the already established normal profile. One can use many distance metrics to get the distance between normal profile and current observation value.

Alexander Strehl, Joydeep Ghosh, and Raymond Mooney [10] stated that if clusters are to be meaningful, the similarity measure should be invariant to transformations natural to the problem domain. The features have to be selected cautiously. They conducted a number of experiments to assure statistical significance of results. Metric distances such as Euclidean are not correct for high dimensional, sparse domains. Cosine, correlation and extended Jaccard measures are successful in capturing the similarities implicitly indicated by manual categorizations as they seen for example in Yahoo.

S. Kullback and R. A. Leibler [11] mentioned that in terms of similarity measure for text information retrieval, difficult it is to differentiate between the populations. R. A. Fisher introduced the criteria for

adequacy required that the statistic selected should recapitulate the whole of the applicable information supplied by the model.

Mei-Ling Shyu, Shu-Ching Chen, Min Chen & Stuart H. Rubin [12] mentioned that as compared to the regular documents, the major unique characteristics or features of the Web documents is the dynamic hyperstructure. In their experimental results and analysis they found that the Euclidean distance gives the bad performance, followed by the cosine coefficient.

N. Sandhya, Y. Sri Lalitha, Dr. A. Govardhan& Dr. K. Anuradha [13] analyzed text document clustering technique plays an vital role in providing instinctive navigation, there is no methodical comparative study of the effect of similarity measures on cluster quality. They conducted a number of experiments and used entropy measure criteria to assure statistical importance of results. Cosine, Pearson correlation and extended Jaccard similarities appear as the best similarity measures to capture human categorization activities, while Euclidean measures perform worst. Analysis involved in their research, they got that there are three mechanism that influence the final outcome representation of the text documents, distance or similarity measures considered, and the clustering algorithm technique itself.

III. CONCLUSION

In this survey, the aim of exploring and comparing different clustering techniques for similarity measures has been studied. Future research in the text mining similarity measure will strive towards improving the entropy, accuracy, precision, and computational speed.

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] 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.

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

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

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

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

[7] J. Kogan, M. Teboulle and C. K. Nicholas, ―Data driven similarity measures for k-means like clustering algorithms‖, Information Retrieval, Vol. 8, No. 2, pp. 331–349, 2005.

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

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

[10] Alexander Strehl, JoydeepGhosh, and Raymond Mooney,‖ Impact of Similarity Measures on Web-page Clustering‖, Workshop of Artificial Intelligence for Web Search, July 2000.

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

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

[13] N. Sandhya, Y.SriLalitha, Dr.A.Govardhan and Dr.K.Anuradha

[14] Analysis of Similarity Measures for Text Clustering‖, GRIOET, Hyderabad, India.