Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Survey on Similarity Measure for Clustering

P. H. Govardhan, Prof. K . P. Wagh, Dr. P.N. Chatur

Dept. of Computer Science and Engineering,

SGB Amravati University, India

Abstract -- The amazing progress of computer technology in the few decades has led to large supplies of powerful and affordable computers. Increase in the number of electronic documents it is hard to visualize these documents efficiently by putting manual effort. These have brought challenges for the effective and efficient organization of web page documents automatically. Extracting features from web pages is initial task found in mining. On the basis of extracted features similarity between web pages are going to be calculated. There is various similarity measures are pointed out for work. To implement the efficient similarity measure one has to do survey on outcomes. This paper helps to study similarity measure for web page clustering.

Keywords: Dynamic page rank , equivalence measure, data analysis , term frequency, degree of closeness, data mining, information extraction

I. INTRODUCTION

Data mining is the process of extracting the implicit, previously unknown and potentially useful information from data. Document clustering, subset of data clustering, organizes documents into different groups called as clusters, where the documents in each cluster share some common properties according to defined similarity measure. Document clustering algorithms play an important role in helping users to effectively navigate, summarize and organize the information.

Due to explosive growth of accessing information from the web, efficient access and exploration of information are needed critically. The Text processing plays an important role in information retrieval, data mining, and web search. Text mining attempts to discover new, previously unknown information by applying techniques from data mining. Clustering, one of the traditional data mining techniques is an unsupervised learning paradigm where clustering methods try to identify inherent groupings of the text documents, so that a set of clusters is produced in which clusters exhibit high intra-cluster similarity and low inter-cluster similarity. Generally, text document clustering methods attempt to segregate the documents into groups where each group represents some topic that is different than those topics represented by the other groups.

II. DIFFERENT METHODS

A. Related work

Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee [1] proposed a new measure for computing the similarity between two documents. Several characteristics are embedded in this measure. 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 decreases when the number of presence-absence features increases. An absent feature has no contribution to the similarity. The similarity increases as the difference between the two values associated with a present feature decreases. This work mainly focuses on textural features. Furthermore, the contribution of the difference is normally scaled. To improve the efficiency, they have provided an approximation 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. Gaddam Saidi Reddy and Dr.R.V.Krishnaiah [2] approach in finding similarity between documents or objects while performing clustering is multi-view based similarity. All measures such as cosine, Euclidean, Jaccard, and Pearson correlation are compared. The conclusion made here is that Ecludean and Jaccard are best for web document clustering. They both selected related attributes for given subject and calculated distance between two values. Both of them used an algorithm known as Hierarchical Agglomerative Clustering in order to perform clustering. Their computational complexity is very high that is the drawback of these approaches. Proposed a similarity measure known as MVS (Multi-Viewpoint based Similarity), when it is compared with cosine similarity, MVS is more useful for finding the similarity of text documents. The empirical results and analysis revealed that the proposed scheme for similarity measure is efficient and it can be used in the real time applications in the text mining domain. It makes use of more than one point of reference as opposed to existing algorithms used for clustering text documents. Shady Shehata, Fakhri Karray and Mohamed S. Kamel [3] mentioned 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 importance of the term within a document only. Text mining model should indicate terms that capture the semantics of text. The mining model can capture terms that present the concepts of the sentence, which leads to discovery of the topic of the document. The mining model that analyzes terms on the sentence, document, and corpus levels are introduced, can effectively

discriminate between non important terms with respect to sentence semantics and terms. The term which contributes to the sentence semantics is analyzed on the sentence, document, and corpus levels rather than the traditional analysis of the document only.

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

Anna Huang [4] declared that before clustering, a similarity/distance measure must be determined. The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. It is very difficult to conduct a systematic study comparing the impact of similarity metrics on cluster quality, because objectively evaluating cluster quality is difficult in itself.

The clusters, which are generated in an unsupervised way, are compared to the pre-defined category structure, which is normally created by human experts. This kind of evaluation assumes that the objective of clustering is to replicate human thinking, so a clustering solution is good if the clusters are consistent with the manually created categories.

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

Hung Chim and Xiaotie Deng [5] found that the phrase has been considered as a more informative feature term for improving the effectiveness of document clustering. They proposed 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 unique feature term in the Vector Space Document (VSD) model, the phrase-based document similarity naturally 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 and developed a new document clustering approach. Their evaluation experiments indicate that the new clustering approach is very effective 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 important roles in text-based information retrieval. The concept of the suffix tree and the document similarity are quite simple, but the implementation is complicated. Investigation 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 expanded feature vector of the traditional single-word terms in the VSD model.

Yanhong Zhai and Bing Liu [6] studied the problem of extracting data from a Web page that contains several structured data records. The objective is to segment these data records, extract data items/fields from them and put the data in a database table.

They proposed approach to extract structured data from Web pages. Although the problem has been studied by several researchers, existing techniques are either inaccurate or make many strong assumptions.

Jacob Kogan, Marc Teboulle and Charles Nicholas [7] argue that the choice of a particular similarity measure may improve clustering of a specific dataset. They called this choice the "data driven similarity measure". They found that the overall complexity of large data sets motivates application a sequence of algorithms for clustering a single data set. Their results of numerical experiments indicate, 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 techniques significantly reduce the dimension.

Syed Masum Emran and Nong Ye [9] said distance metric value is used to find the similarity or dissimilarity of the current observation from the already established normal profile. To find the distance between normal profile and current observation value, one can use many distance metrics. Alexander Strehl, Joydeep Ghosh, and Raymond Mooney [10] studied if clusters are to be meaningful, the similarity measure should be invariant to transformations natural to the problem domain. The features have to be chosen carefully. They conducted a number of experiments to assure statistical significance of results. Metric distances such as Euclidean are not appropriate 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] found that in terms of similarity measure for information retrieval, difficult it is to discriminate between the populations. R. A. Fisher introduced the criteria for sufficiency required that the statistic chosen should summarize the whole of the relevant information supplied by the sample. Mei-Ling Shyu, Shu-Ching Chen, Min Chen & Stuart H. Rubin [12] mentioned that compared to the regular documents, the major distinguishing characteristics of the Web documents is the dynamic hyper-structure. In their experimental results they found that the Euclidean distance gives the worst performance, followed by the cosine coefficient. N. Sandhya, Y. Sri Lalitha, Dr. A. Govardhan & Dr. K. Anuradha [13] analyzed text document clustering plays an important role in providing intuitive navigation, there is no systematic comparative study of the impact of similarity measures on cluster quality. They conducted a number of experiments and used entropy measure to assure statistical significance of results. Cosine, Pearson correlation and extended Jaccard similarities emerge as the best measures to capture human categorization behavior, while Euclidean measures perform poor.

They found that the measures have significant effect on clustering of text documents. Considering the type of cluster analysis involved in their study, they got that there are three components that affect the final results representation of the documents, distance or similarity measures considered, and the clustering algorithm itself.

III. CONCLUSION

In this survey, the aim has been to investigate and compare different techniques for similarity measures. Future research in the data mining similarity measure will strive towards improving the 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] Gaddam Saidi 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, Fakhri Karray, 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] Yanhong Zhai 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 Masum Emran 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, Joydeep Ghosh, 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.Sri Lalitha, Dr.A.Govardhan and Dr.K.Anuradha "Analysis of Similarity Measures for Text Clustering", *GRIOET*, Hyderabad, India.

