Web Document Clustering using Proposed Similarity Measure

P. H. Govardhan
Department of Computer Science and Engineering, Government College of Engineering, Amravati.

K. P. Wagh
Department of Information Technology, Government College of Engineering, Amravati.

P. N. Chatur, Ph.D.
Department of Computer Science and Engineering, Government College of Engineering, Amravati.

ABSTRACT
Recent advance research in data warehousing and data mining emerges various types of information sources. Web documents are the most useful information resources in this era. Efficient uses of these resources are most important for knowledge discovery. Bunch of documents providing related information is to be grouped in one cluster. Finding the similarity between documents is tedious task. There are various similarity measures introduced earlier to solve the problems related to clustering. Proposing new similarity measure to get better results of clustering is reason behind this paper work. As before concern to previous research, there is no consideration of present and absent features in documents. Proposed similarity measure concentrates on both present and absent features in the documents. Concentrating on similarity measure will help to mining technique.

Keywords
Cluster, Document Vector, Inverse Document frequency, Similarity Measure, Term Frequency, Web Document.

1. INTRODUCTION
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. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances. However, even in this case the Euclidean distance can sometimes be misleading.

A document is usually represented as a vector in which each component indicates the value of the corresponding feature in the document. Selecting the similarity measure can be a severe challenge which is an important operation in text processing. A lot of measures have been proposed for computing the similarity between two vectors. It proposes a measure for computing the similarity between documents which embeds several characteristics. 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. The similarity decreases when the number of presence-absence features increases.

2. LITERATURE SURVEY
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.

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. Their computational complexity is very high that is the drawback of these approaches.

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. The mining model that analyzes terms on the sentence, document, and corpus levels are introduced, can effectively discriminate between non important terms.

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. 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.

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. 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.

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.

They 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.

Researcher 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.

Observer 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.

3. PROPOSED WORK

In this section, discussion will lead with proposed work. Similarity measures have been extensively used in text classification and clustering algorithms.

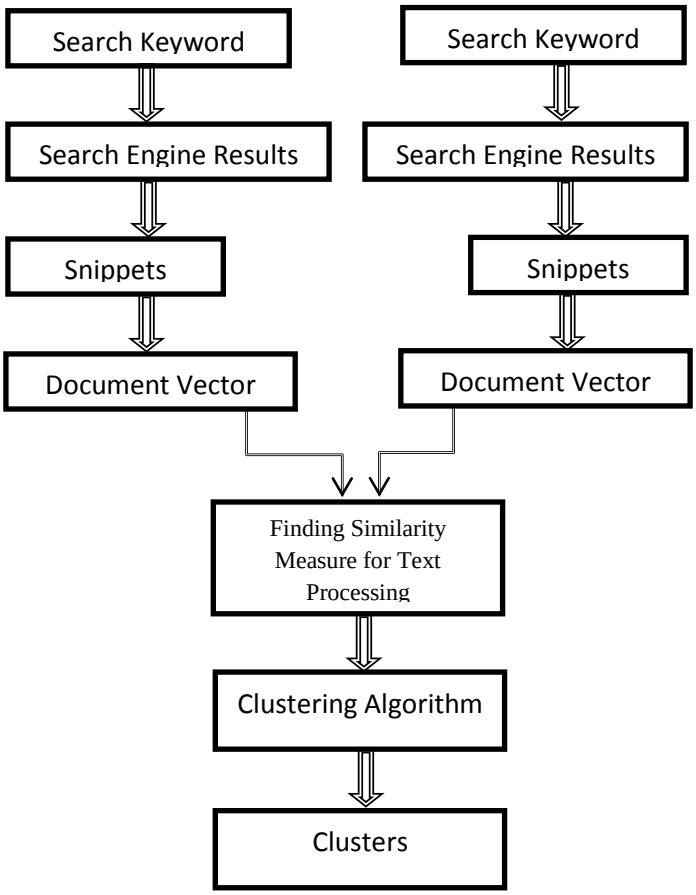

Fig 1: System Architecture for Proposed Work

A measure for computing the similarity between documents which embeds several characteristics i.e. 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. The similarity decreases when the number of presence-absence features increases.

3.1 Search Keyword

As we are discussing about searching keyword. Proposed work will lead by searching keyword. Searching keyword will be direct input to the system. This input will be provided for search engine where search engine will try to find out searching keyword results. According to the code design for number of searching results for system, we will get the specific number of searching results for further process. This search engine results known as snippets which holds basic information related to search keyword. Overall search keyword plays efficient role for proposed system.

3.2 Search Engine Results

As compare to searching keyword, our proposed system will provide this keyword to search engine. Search engine will find out search results for our proposed system. To provide search keyword for search engine through proposed system we require internet connection for system. The system will be design to hold search engine results as per code of line written for number of search engine results require for system. Search engine results hold the information as similar to search keyword. Every result will have healthy information regarding search keyword. This information is more likely beneficial for data mining.

3.3 Snippets Retrieval

Next task relevant to further processing is snippets retrieval. Snippets are nothing but search engine results provided to proposed system from search engine. This snippet holds the basic information as per search keyword. This information is more relevant to search keyword. So this information will have to retrieve from search engine results. Processing of this information for similarity measure is more beneficial than processing whole information of web page. Snippets hold the basic information relevant to web page in the form of bunch of few words. This bunch of words plays important role for whole web page. Processing this bunch of words is easy as compare to processing done for whole page. As per number of researchers done work on web page clustering, they evaluated that processing snippet for web page clustering is more easy and efficient processing than concentrating on whole web document.

3.4 Document Vector

Further processing leads with the creating document vector. Processing whole document with respect to text is very tedious. Creating document vector for processing whole text is easier than considering whole document for processing. Document vector contains numerical values which are carried out from calculation like term frequency-inverse document frequency. This numerical values further will used for calculation and represents document.

Here, we can derive document vector with example how the actual method works; such as follow,

Suppose we have two documents:

Document 1:
Data mining technique finds solution for mining problems. Provides traditional way solution for warehouse.

Document 2:
Processing big data of warehouse requires solution for knowledge discovery. Mining techniques always provides solution.

Assume that we use word count as feature values and we consider7 features which are data, mining, warehouse, knowledge, technique, solution, and discovery respectively. Then we have two corresponding vectors d1and d2 as

d1 = <1, 2, 1, 0, 1, 2, 0 >,

d2 = <1, 1, 1, 1, 1, 2, 1>

3.5 Finding Similarity Measure for Text Processing

Proposed similarity measure concentrates on both present and absent features in the documents. Concentrating on similarity measure will help to mining technique. Based on the preferable properties mentioned above, we propose a similarity measure, called SMEC (Similarity Measure for Efficient Clustering),for two set of documents such as follow. Let G1 and G2 be two document sets containing k1 and k2 documents, respectively, i.e.,

G1 = {d11, d12, . . . , d1k1} and

G2 = {d21, d22, . . . , d2k2}.

The function F between G1 and G2is defined to be

Define a function F as follows:

$$F(G1,G2) = \frac{\sum_{l=1}^{k1} \sum_{m=1}^{k2} \sum_{n=1}^{z} N*(d_{ln}^1, d_{mn}^2)}{\sum_{l=1}^{k1} \sum_{m=1}^{k2} \sum_{n=1}^{z} N_U(d_{ln}^1, d_{mn}^2)}$$

Where,

$N_*(d_{ln}^1, d_{mn}^2)$ =

$$\begin{cases} 0.5\left(1+\exp\{-\left(\frac{d_{ln}^1 - d_{mn}^2}{\sigma_k}\right)^2\}\right), & \text{if } d_{ln}^1 . d_{mn}^2 > 0 \\ 0, & \text{if } d_{ln}^1 = 0 \text{ and } d_{mn}^2 = 0 \\ -\lambda, & \text{otherwise} \end{cases}$$

$N_U(d_{ln}^1, d_{mn}^2)$ =

$$\begin{cases} 0, & \text{if } d_{ln}^1 = 0 \text{ and } d_{mn}^2 = 0 \\ 1, & \text{otherwise} \end{cases}$$

SMEC (Similarity Measure for Efficient Clustering) for G1 and G2is considering as follow:

$$S_{\text{SMEC}}(G1,G2) = \frac{F(G1,G2)+\lambda}{1+\lambda}$$

The proposed SMEC (Similarity Measure for Efficient Clustering) for G1 and G2 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.

3.6 Clustering

In data mining, processing is a key task to ensure reliability and quality of the knowledge extracted by the whole mining process. Web page clustering deal with a set of web pages. As the amount of data to process is potentially infinite if dynamic web pages are considered, the need of processing this information seems necessary to deal with this computational problem. Web page clustering appears as a reasonable solution. These techniques group pages together based on some kind of relationship measure. Pages in the same cluster will be considered as a single item for further data analysis steps. The goal of clustering is to reduce the large amount of raw data by categorizing in smaller sets of similar items. Similarity measures have been extensively used in text clustering algorithms. Clustering means storing related objects close together on secondary storage so that when one object is accessed from disk; all its related objects are also brought into memory. Then access to these related objects is a main memory access that is much faster than a disk access. There are several clustering algorithms available such as k-means, k-medoid, HAC, BuckShot, Expectation Maximization (EM) etc. These algorithms can be used for proposed work.

4. ACKNOWLEDGMENTS

The authors are grateful to the anonymous reviewers for their comments, which were very helpful in improving the quality and presentation of the paper.

5. CONCLUSION

The process of Data mining is used to uncover hidden or unknown information that is not apparent, but potentially useful. Meaningful information lives in the form of text which is extracted from web pages. On the basis of preprocessing results of web pages; measuring the similarity between documents is an important operation in the text processing field. Hence, the study of similarity measure for clustering is initially motivated by a research on automated text categorization. Proposed work covers all the possibilities such as present and absent feature for considering each case related to document. The application of document clustering to information retrieval has been motivated by the potential effectiveness gains postulated by the cluster hypothesis.

6. 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] 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 ComputerIntrusion Detection", IEEE Workshop onInformation 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.