International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 4, Issue 10, October 2015 **ISSN 2319 - 4847**

Review on Similarity Measures for Document Clustering

Neha Chopade[1] , Dr. Jitendra Sheetlani[2]

[1]Associate Professor,SIES College of Management Studies, Navi Mumbai

[2]Associate Professor, Shri Satya Sai University of Technology and Medical Sciences, Sehore

ABSTRACT

This paper reviews various similarity measures used in document clustering algorithms. The similarity measures reviewed in this paper are not conventional similarity measures. The clustering of web documents is based on the similarities and dissimilarities between the web pages.. The basic idea of document clustering is to add new coming web page into the cluster to which it is more similar.

Keywords: Document Clustering, Inter Cluster Similarity Measure, Intra Cluster Similarity Measure, Degree of Coherency.

1. INTRODUCTION

Nowadays, the people from every age group are using internet as their main source of information, since it is the largest data depository where almost all type of data is available. In every second millions of pages are getting uploaded on the internet and due to this abundance of information, the process of information retrieval becomes tedious. The internet users have one common expectation to retrieve fast and best matching information of their interest. This objective can be easily achieved by applying a technique called document clustering. Document clustering is a text mining technique to efficiently explore useful information from textual data. It is a process of grouping documents into clusters on the basis of their similarities. According to this technique the clusters must exhibit high intra cluster similarity and low inter cluster similarity. In any clustering problem it is a challenge to identify the features of document which needs to be considered discriminatory [1]. In other words we must choose the attributes like words, phrases or links of the documents on which the clustering will be based which indicates a document model. When documents are represented by a bag of words, the resulting document-word matrix typically represents data in 1000+dimensions. Several attempts have emerged to efficiently cluster documents that are represented in such high dimensional space. If clusters are to be meaningful, the similarity measure should be invariant to transformations natural to the problem domain [2]. Any clustering technique relies on four concepts:

1. A data representation model,
2. a similarity measure,
3. a cluster model, and
4. a clustering algorithm that builds the clusters using the data model and the similarity measure.

In this paper we are focusing on various similarity measure used in document clustering algorithms.

2. SIMILARITY MEASURES

2.1 A Phrase-Based Similarity Measure

The Phrase similarity [4] was introduced by Hammouda, K.M. & Kamel M.S. in 2004 . They devised a similarity measure based on matching phrases rather than individual terms.between two documents is calculated based on the list of matching phrases between the two documents. Frequency of phrases is an important factor in this similarity measure. The frequent the phrase appears in both documents, the more similar they tend to be. The cosine and Jaccard measures are indeed of such nature, but they are essentially used as single term based similarity. This phrase based similarity measure is a function of four factors:

- The number of matching phrases P,
- The lengths of the matching phrases

$$(l_i : i= 1,2,\ldots\ldots,P).$$

- The frequencies of the matching phrases in both documents (f_{1i} and f_{2i} : i= 1,2,…….,P), and
- The levels of significance (weight) of the matching phrases in both document (w_{1i} and w_{2i} : i= 1,2,…….,P).

The phrase similarity between two documents, d_1 and d_2 , is calculated using the following empirical equation:

$$\text{sim}_p(d_1, d_2) = \frac{\sqrt{\sum_{i=1}^{P} [g(l_i) \cdot (f_{1i}\ w_{1i} + f_{2i}\ w_{2i})]^2}}{\sum_j |s_{1j}| \cdot w_{1j} + \sum_k |s_{2k}| \cdot w_{2k}}$$

where $g(l_i)$ is a function that scores the matching phrase length, giving higher score as the matching phrase length approaches the length of the original sentence, $|s_{1j}|$ and $|s_{2k}|$ are the original sentence lengths from document d_1 , and d_2 , respectively.

2.2 Histogram Based Similarity measure

It is an incremental dynamic method of building the clusters which was proposed by Hammouda, K.M. & Kamel M.S[5]. The key focus of this concept is to keep each cluster at a high degree of coherency at any time. It is a concise statistical representation of the set of pair wise document similarities distribution in the cluster. To find out the quality of cluster cohesiveness they calculated the ration of the count of similarities above a certain similarity threshold S_T to the total count of similarities. The higher this ratio, the more cohesive is the cluster.
Let n_c be the number of documents in a cluster. The number of pair wise similarities in the cluster is $m_c = n_c(n_c+1)/2$. Let S = { s_i : I = 1, ….. , m_c } be the set of similarities in the cluster. The histogram of the similarities in the cluster is represented as :
H = { h_i : i=1,….,B}
h_i = count(s_k) $\quad s_{li} \le s_k < s_{ui}$
Where B : the number of histogram bins,
h_i : the count of similarities in bin i
s_{li} : the lower similarity bound of bin i and
s_{ui} : the upper similarity bound of bin i.
The histogram ratio (HR) of a cluster is the measure of cohesiveness of the cluster is calculated as :

$$HR_c = \frac{\sum_{i=T}^{B} h_i}{\sum_{j=1}^{B} h_j}$$

$T = \lfloor S_T \cdot B \rfloor$
Where HR_c : the histogram ratio of cluster c,
S_T : the similarity threshold, and
T : the bin number corresponding to the similarity threshold.

2.3 Multi View point Similarity Measure

Cosine similarity is a single view point measure since it is calculated by finding the cosine angle between the two document vectors at the origin i.e vector 0. In 2012 D. T. Nguyen, L. Chen, and C. K. Chan proposed a new similarity measure to obtain a more accurate assessment of how close or distant a pair of document points (d_i and d_j) is, if we could measure them by standing at more than just one viewpoint as references. For example, from a third point d_h , the direction and distances to d_i and d_j are indicated by two new vectors $(d_i - d_h)$and $(d_j - d_h)$respectively. Therefore, working on different vectors with a number of different viewpoints, the similarity between a pair of document is defined as:

$$\underset{d_i, d_j \in S_r}{Sim(d_i, d_j)} = \frac{1}{n - n_r} \sum_{d_h \in S \setminus S_r} Sim(d_i - d_h, d_j - d_h)$$

2.4 Binary Method of finding Intra and Inter Cluster Similarities

This method [3] was proposed by S. Mimaroglu and A. M. Yagci in 2010 for finding the inter and intra cluster similarities.For a clustering, $(D) = \{C_1, C_2, \ldots\ldots, C_{|\pi(D)|}\}$, intra cluster similarity is measured as follows:

$$ICS(\pi(D)) = \sum_{i=1}^{|\pi(D)|} \frac{1}{|C_i|^2} \sum_{d, d' \in C_i} similarity(d, d')$$

and inter cluster similarity is measured as follows:

$$ECS(\pi(D)) = \sum_{i=1}^{|\pi(D)|} \sum_{j=i+1}^{|\pi(D)|} \frac{1}{|C_i||C_j|} \sum_{d \in C_i, d' \in C_j} similarity(d, d')$$

3. CONCLUSION AND FUTURE WORK

Various similarity measures have been used in document clustering. Each of them is differently defined by its own properties. In this review paper, we collected four similarity measures used over the last few years. The similarity measures discussed in this paper are already compared against the traditional method of finding similarity between documents while they proposed and they produce better results. Future work will aims to the comparisons of these similarity measures to select the accurate measure for document clustering.

REFERENCES

[1]. http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf [Accessed : Aug. 10, 2015]
[2]. Shrehl A., Ghosh J., and Mooney R "Impact of Similarity Measures on Web-page Clustering" Workshop of Artificial Intellegence for Web Search , July 2000, pp . 58-64.
[3]. Mimaroglu S. and Yagci A.M. "A Binary Method for Fast Computation of Inter and Intra Cluster Similarities for Combining Multiple Clusterings" in Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human , 2009, pp. 452-456.
[4]. Hammouda K. M., and Kamel M.S. "Efficient Phrase-Based Document Indexing for Web Document Clustering" IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 10 , October 2004.
[5]. Hammouda K.M. , and Kamel M.S. "Incremental document Clustering Using Cluster Similarity Histograms" in Web Intellegence IEEE/WIC Int. Conference Proceedings, pp. 597-601 , 2003.
[6]. Nguyen D.T. , Chen L., and Chan C.K. " Clustering with Multiviewpoint –Based Similarity Measure" IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 6, October 2012.

AUTHORS

Neha Chopade received the M.C.A degree from Govt. Geetanjali Girls College , Barkatullah University,Bhopal in 2000. She is having teaching experience of 16 years. Her area of interest for research are web mining. Currently she is working with SIES college of Management studies.

Dr. Jitendra Sheetlani received the M.C.A degree from Pt. Ravishankar University, Raipur in 2003. He has done his Ph.D.(Computer Science) on "An Efficient Implementation of Concurrency Control in Distributed Database Transaction Process", from NIMS University Jaipur, Year 2013. Now he is working with Sri Satya Sai University of Technology & Medical Sciences as Associate Professor