International Journal of Pure and Applied Mathematics
Volume 116 No. 9 2017, 409-415
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

AN EFFECTIVE MEASURE FOR RETRIEVAL OF PATTERNS USING CLUSTERING TECHNIQUES

.K.Shanmugapriya[1] Vimala Muthukumar[2]
[1,2] Assistant Professor,,
Department of Computer Science and Engineering,
BIST, BIHER, Bharath University, Chennai, Tamil Nadu, India
[1]sharveshvinay@gmail.com, [2]vimala.cse@bharathuniv.ac.in

Abstract: Among information mining strategy, grouping is a standout amongst the most essential and conventional idea likewise an unsupervised learning worldview. Likeness of a record sets can be measured by coordinating of ideas. Finding or removing the most applicable idea from the records is a challengeable undertaking. To address this issue, in this paper we present an idea of multi see point based closeness measure. Our proposed strategies utilizes various perspective between record sets to remove more applicable match idea instead of extricating just thoughts in view of closeness measure. Utilizing numerous view point, accumulates more data about a specific subject from a wide range of however significant sources or idea. This procedure functions admirably with littler archives yet is particularly powerful with longer records. By social affair more pertinent ideas from the records with numerous perspectives, the report association and recovery can upgrade the capacity to make the most utilization of the archives held away and make recovery of thoughts and also important errand or idea considerably less demanding and speedier. Exploratory outcomes demonstrates that our proposed technique effectively extricate more pertinent idea.

Keywords: Similarity measure, Concept mining, Document grouping.

1. Introduction

Clustering technique is a vital and helpful method that naturally sorts out an accumulation with a significant number of information objects into a considerably more modest number of sound gatherings for promote examination. All the more decisively bunching is the way toward sorting out articles into subgroups whose individuals are comparative somehow [1-3]. We are confronting a regularly expanding volume of content reports. Presence of a lot of writings streaming over the web, archives, some digitized individual data, for example, messages and blog articles are rapidly heaping up each day. Those things brought challenges for successful and productive association of content records. To deal with this issue grouping has turned out to be a compelling methodology. With the advancement of World Wide Web and consequent development of web2.0 this bunching procedure turns out to be all the more fascinating methodology [4]. Consider an illustration, the outcomes returned via web indexes are bunched to help the clients rapidly distinguish and concentrate on the proper arrangement of results. Exact bunching requires an exact meaning of the closeness between a couple of items which thus either as a couple insightful similitude or separation. There exists an assortment of comparability or separation measures have been proposed and connected broadly. Some of them incorporate cosine similitude and Jaccard connection coefficient. A few measures, for example, Euclidean separation and relative entropy have been connected generally in grouping to ascertain the match savvy separations. In content mining idea, the information are separated by processing the recurrence of a term to investigate the significance of the term in a report.

For a moderately little gathering of archives it is conceivable to physically play out the parceling of records into particular districts. What's more, a similar time while apportioning of substantial volumes of content it ought to be bunched in view of closeness measure utilizing some grouping calculation. The primary prerequisites that a bunching calculation ought to fulfill depend on adaptability, capacity to manage commotion and exceptions, Exhibits high dimensionality, ready to manage distinctive sorts of qualities, achieves obtuseness to request of info records, negligible necessities for space information to decide input parameters, finding groups with discretionary shape, encourages interpretability and ease of use[6-8]. The proposed method consists of sentence based concept analysis, document based concept analysis, and concept based similarity measure. To analyze each document on concept based similarity measure the three factors needs to be calculated. The three measures include *ctf (Conceptual term frequency), tf(term frequency), df(Document frequency).* This concept based similarity measure is performed by matching

concepts at the sentence, document rather than individual terms or words only. Our proposed work exploits the information extracted from the concept based analysis algorithm to better judge the similarity between the documents. The similarity measure of document and extracting more relevant concept from the raw documents can be better achieved by employing multiple reference point within the documents. We may have more accurate information regarding the extraction of relevant concept. By standing at various reference point similarity between the documents and the generation of more relevant concept would be achieved. The similarity measure to be compared includes Euclidean distance as a performance metricsp[9-10].

2. Related work

Looking of related data about the execution change on record grouping is an expanding and furthermore an intriguing exploration zone. Huge numbers of the exploration researchers are attempting until the point that a date for a technique for recovering more significant idea on content grouping. Here, we display a portion of the significant works did before by different researchers. To measure the quality of cluster hierarchy, quality metrics like F-measure and the results compared with various clustering algorithm. This approach was dealt by authors in [9] which are a fast and effective text mining using linear time document clustering. The evaluation considers some feature selection parameters like *tfidf*, feature vector length. But this technique has limitations of scalability. A robust hierarchical clustering algorithm designed in [15-17] that employed link and not distances to measure the similarity between a pair of data points when merging the clusters. This method is useful in situations where a domain expert/similarity table is the only source of knowledge. Efficient document clustering can help automatically organize the document corpus into a meaningful cluster hierarchy for efficient browsing and navigation. Such efficiency can be achieved by the scholars of [18] through non-negative matrix factorization which is combined form of K-means clustering. This method uses an online NMF algorithm to efficiently handle very large scale and streaming datasets. Conventionally this algorithm requires the data matrix to reside in the memory during the solution process leads to problematic approach when the datasets are very large.

Similarity of a xml record also can be detected through the technique evolved in [4]. XML files may be in comparison to their structural similarity and group them into the clusters in order that distinctive garage employer, data retrieval may be exploited greater successfully. The designed method absolutely hit upon the structural similarity between xml documents which notably differs from widespread methods also allow a vast reduction of the computational fees. Based at the paper proposed by author in [7] the aim of cluster analysis is to partition a statistics set of N objects into subgroups such that the ones in each particular group are greater much like each other than to the ones of other businesses. This algorithm enables to find organization of gadgets which have preferentially near values on one-of-a-kind more possibly overlapping, and an characteristic of subsets.

According to the paper [10] to manner large file collections fast and to search collections that overall inside the order of billions to trillions of words then the statistics retrieval ought to needs to be efficient. Since the collection of online records has grown at least as quick as the speed of computers, coping with and retrieval of statistics is very essential. [21] Focus on report similarity based totally on concept tree distance. Each record is taken as a concept tree and employs a tree similarity measure primarily based on tree edit distance to compute similarities among idea trees. This technique constructs the idea trees representing the documents and applies the tree edit distance algorithm to calculate the document similarity.

3. Proposed methodology

In this segment, the proposed method for finding the greater applicable concept with multi view-factor primarily based similarity degree is specific. The normal method flow diagram is also depicted on this phase. The algorithm that shows the distinct computation of locating relevant concept the usage of our proposed technique is represented in this segment.

3.1 Similarity Degree

The similarity among the files is based totally at the aggregate of sentence-based, corpus-based totally, and record-based totally analysis. Document similarity may be measured by way of matching of ideas among document pairs. Significant results at the clustering high-quality due to similarity's insensitivity to noisy terms cause wrong similarity. Extracting concepts are much less touchy to noise in terms of calculating file similarity. The standards are originally extracted by way of the semantic role labeler and these concepts are analyzed with admire report, sentence and corpus degrees. The proposed concept based similarity measure extracts more relevant concept in report pairs.

3.2 Semantic position labeler

Semantic shape of a sentence may be characterised with the aid of a form of verb argument structure. Consider an e.G. Difficulty-verb-item. Now, earlier than u item and

saying the words have some meaning however now not structure. Assume sentences,

"Dog bites man"

"Man bites dog"

In the above sentences, the phrases have meaning, also the identical time two sentences containing the equal phrases inside the identical structure but may have one of a kind meaning depending on in which within the shape, the words seem. Move to 1st case, canine that bites man, then second case the person who bites the canine. By converting the position of "words" otherwise said to be "terms" inside the shape adjustments the which means of shape. Each time period has a semantic function within the sentence is known as "concept". Labels are assigned to every sentence the use of semantic function labeler.

Concepts may be either words or phrase and are absolutely depending on the semantic structure of the sentence. Our proposed approach output the matched concepts by means of this semantic position evaluation in a sentence.

3.3 Concept mining

The objective behind our proposed model is to achieve an accurate analysis of concept on the sentence, corpus level and also the document and finding more relevant and matched concept using multi view point based similarity measure. The following factors needs to be computed to find similarity based on concept analysis.

3.4 Sentence based concept analysis

This approach analyzes each concept at the sentence level using Idea based frequency degree which is denoted through conceptual time period frequency (ctf). The ctf calculation of idea c in sentence s and report d is expressed as,

a) Calculating ctf issue of idea c in sentence s:

The ctf is the range of occurrence of idea c in verb argument of sentence s. If the idea c appears often in one-of-a-kind verb argument of equal shape of identical sentence s then it plays a predominant function of contributing to meaning of s.

b) Calculating ctf thing of concept c in document d:

The ctf fee of idea c in file is computed if the concept c takes many ctf values in one of a kind sentences within the identical record d. It is denoted in equation (1) as,

$$\text{Ctf} = \frac{\sum_{n=1}^{sn} ctf_n \sum_{n=1}^{sn} ctf_n}{sn \qquad sn} \quad (1)$$

sn is the total number of sentences that contain concept c in document d. The average value of ctf is computed to make the overall importance of concept c.The overall flow of our proposed model is depicted .

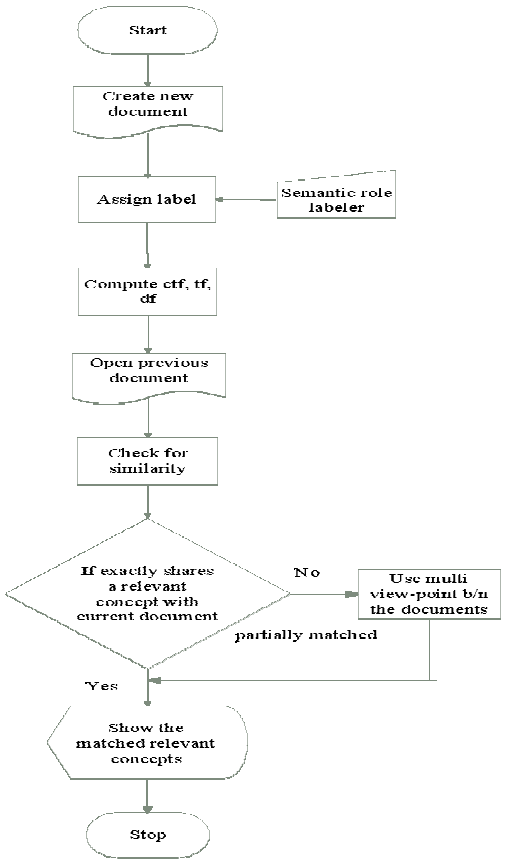

Figure 1. Overall Flow of our proposed model

3.3 Document based concept analysis

The concept based term frequency *tf* is computed to analyze each concept at the document level. It denotes the number of occurrence of a concept c in original document.

3.4 Corpus based concept analysis

The number of documents containing concept c is finding out by calculating document frequency*df*which extract concept that can discriminate between documents.

Similarity between the documents is checked by matching of relevant concept. The output may be exact match or partial match but the efficiency can be achieved by extracting exact concepts within the documents. The proposed method attains overall accuracy and efficiency by using multiple reference point between the document pairs. The similarity between the documents (di, d_j) is obtained with respect to the angle between the documents.

$$Sim\ (di,\ d_j) = \frac{\frac{1}{n-n_r}\frac{1}{n-n_r}\sum_{d_h} sim(d_i - d_h, d_j - d_h}{\sum_{d_h} sim(d_i - d_h, d_j - d_h)} \quad (2)$$

By standing at various reference point d_h to view document similarity based on concept between two documents d_i and d_j, more exact matching of concepts can be extracted within the document d. By this way our proposed method achieves efficiency.

4. Algorithm

An algorithm that depicts our computation of match concepts is given above:

1. Algorithm: Cluster the documents based on matching relevant concept
2. Input: Raw text documents
3. Output: Clustered documents with matched relevant concept
4. Begin
5. Create new document d_{doci}
6. $S_{doci}$is a sentence in d_{doci}
7. Label each sentence by semantic role labeler
8. Construct concept list $C_{doci}$from s_{doci}
9. For each concept c_i € $C_i$do
10. Compute ctf_i of c_i in d_{doci}
11. Compute tf_i of c_i in d_{doci}
12. Compute df_i of c_i in d_{doci}
13. Open previous document d_k
14. S_k is a sentence in d_k
15. Build concept list C_k from s_k
16. For each concept c_j € $c_k$do
17. If (c_i == c_j) then
18. Update df_i of c_i
19. Compute ctf_{weight}= avg (ctf_i, ctf_j)
20. Insert new concept matches to L
21. End if
22. End for
23. End for
24. Show the matched concept list L
25. If L is a partial match then
26. Employ multiple reference point b/n doc(d_{doci} , d_k)
27. Establish exact match
28. End if
29. Output the more relevant matched concepts

3.5 Experimental evaluation:

The goal of our proposed method that uses multiple reference point between the document pairs is to maximize the retrieval of more relevant and matched concepts. The experiment was conducted to test the effectiveness of relevant concept matching in determining an accurate measure of the similarity between documents. The effectiveness of our proposed approach are evaluated based on precision, recall, the prediction accuracy and the time taken for processing and retrieving the relevant match concepts from the documents. The following figures show comparison between our proposed and existing method. The existing method used concept based mining model to retrieve the match concepts whereas our proposed method uses multiple reference point within the concept based mining model to retrieve the more relevant exact match concepts and cluster the documents based on exactly matched concepts.

Precision is a measure that expresses the fraction of returned or retrieved documents that are more relevant and matched concepts. This measure shows the consistent results which are purely based on the measure and understanding of relevance. Precision can be calculated using the equation (3),

$$\text{Precision} = \frac{\text{No.ofrelevant document retrieves}}{\text{Total no.of document retrieved}} \quad (3)$$

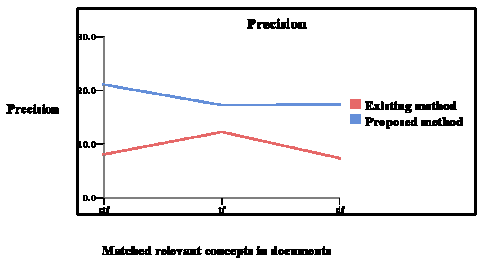

Figure 2. Precision study for existing & proposed

Fig2.illustrates that our proposed method has higher precision value than the existing one for matched concept retrieval.

Similarly the existing and proposed approaches are compared using Recall which is another measuring factor. The estimation of recall value is calculated through equation (4).

$$\text{Recall} = \frac{\text{No.of relevant document retrieved}}{\text{Total no.of existing relevant document}} \quad (4)$$

Figure 4 reveals the recall values for proposed and existing techniques.

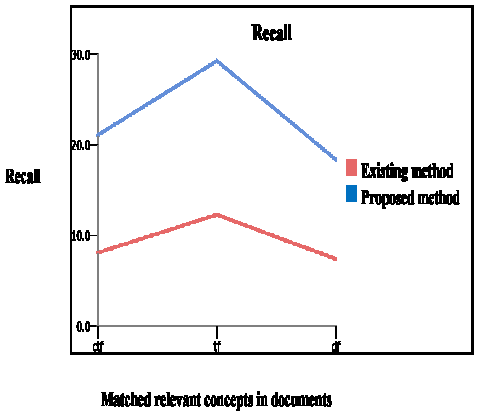

Figure 3. Recall study for existing & proposed

It also reveals that the proposed method consumes less time than the existing method. This explicitly denotes that proposed method retrieves more relevant matched concept in documents faster than the existing method. The computation time is calculated in seconds. The experimental results shows that our proposed method retrieves more relevant matched concepts in document pairs with an accuracy rate of 79% which is 13.7% greater than the existing method. It is also shown that our proposed method extracts relevant matched concepts in much fast seconds when compare to the existing technique.

5. Conclusion

A new technique to empower the document clustering by extracting the more relevant concept can be achieved by employing multiple reference point on the documents. The proposed methodology gains more efficiency by gathering more relevant match concepts from the documents with multiple points of reference at the document pair, both the document organization and retrieval can enhance the ability to make the most use of the documents held in storage and make retrieval of ideas as well as relevant task or concept much easier and faster.

The experimental results proves that our proposed method extract document with an accuracy of 79% which are more relevant and matched concepts. It also consumes less time than the existing method. Our proposed method outperforms the existing method. To enhance the proposed approach, carry out the same work to web document clustering.

References

[1] Khaled M.Hammouda "*Efficient Phrase-Based Document Indexing for Web Document Clustering*", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 10, OCTOBER 2004.

[2] Dekang Lin. "*Automatic retrieval and clustering of similar words.* In Proceedings of theCOLING-ACL, Montreal, Canada, 1998.

[3] Fei Wang and Chenhao Tan, *"Efficient Document Clustering via Online Nonnegative Matrix Factorizations",* Cornell university, Newyork

[4] S. Flesca, G. Manco, E. Masciari, L. Pontieri, and A. Pugliese, "*Fast Detection of xml Structural Similarity*," IEEE Trans. Knowledge andData Eng., vol. 17, no. 2, pp. 160-175, Feb. 2005

[5] Javed Aslam, Katya Pelekhov, and Daniela Rus, "*A Practical Clustering Algorithm for Static and Dynamic Information Organization"*, Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management, Bethesda, Maryland, USA, Pages 208-217, November3-7, 1998

[6] D. Boley, M. Gini, R. Gross, S. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, "*Partitioning-Based Clusteringfor Web Document Categorization*", Decision Support Systems, vol. 27, pp. 329-341, 1999

[7] J. Friedman and J. Meulman, "*Clustering Objects on Subsets of Attributes,*" J. Royal Statistical Soc. Series B Statistical Methodology, vol. 66, no. 4, pp. 815-839, 2004

[8] I.S. Dhillon, S. Mallela, and D.S. Modha, "*Information-Theoretic Co-Clustering*," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003

[9] Bjorner Larsen and Chinatsu Aone, "*Fast and Effective Text Mining Using Linear-time Document Clustering"*, KDD-99, San Diego, California, 1999

[10] C.D. Manning, P. Raghavan, and H. Schu¨ tze, "*An Introduction to Information Retrieval"*. Cambridge Univ. Press, 2009

[11] D. Gildea and D. Jurafsky, "*Automatic Labeling of SemanticRoles,"* Computational Linguistics, vol. 28, no. 3, pp. 245-288, 2002.

[12] S. Shehata, F. Karray, and M. Kamel, "*Enhancing Text Clustering Using Concept-Based Mining Model*," Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM), 2006

[13] W. Wong and A. Fu, "*Incremental Document Clustering for Web Page Classification*," Proc. 2000 Int'l Conf. Information Soc. in the 21st Century: Emerging Technologies and New Challenges (IS2000), 2000.

[14] S. Zhong, "*Efficient Online Spherical K-means Clustering*," Proc. IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005

[15] S.Y. Lu and K.S. Fu, "*A Sentence-to-Sentence Clustering Procedure for Pattern Analysis,*" IEEE Trans.

Systems, Man, and Cybernetics, vol. 8, no. 5, pp. 381-389, May 1978
[16] R. Nock and F. Nielsen, "*On Weighting Clustering*," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1223-1235, Aug. 2006
[17] SudiptoGuha, Rajeev Rastogi, and Kyuseok Shim, " *A Robust Clustering Algorithm for Categorical Attributes"*, In Proceedings of the 15th International Conference on Data Engineering, 1999
[18] S. Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky, "*Shallow Semantic Parsing Using Support Vector Machines*," Proc. Human Language Technology/North Am. Assoc. for Computational Linguistics (HLT/NAACL), 2004
[19] DapheKoller and Mehran Sahami, "*Hierarchically classifying documents using very few words"*, Proceedings of the 14th International Conference on Machine Learning (ML), Nashville, Tennessee, July 1997, Pages 170-178
[20] S. Pradhan, K. Hacioglu, W. Ward, J.H. Martin, and D. Jurafsky, "*Semantic Role Parsing: Adding Semantic Structure to Unstructured Text*," Proc. Third IEEE Int'l Conf. Data Mining (ICDM), pp. 629-632, 2003
[21] P. Lakkaraju, S. Gauch, and M. Speretta, "*Document Similarity Based on Concept Tree Distance,*" Proc. 19th ACM Conf. Hypertext and Hypermedia, pp. 127-132, 2008
[22] M. Steinbach, G. Karypis, and V. Kumar, "*A Comparison of Document Clustering Techniques*," Proc. Knowledge Discovery and Data Mining (KDD) Workshop Text Mining, Aug. 2000.