Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DOCUMENT CLUSTERING USING CO-WORD ANALYSIS AND FORMATION OF KEYWORD AGAINST DOCUMENT MATRIX

[1]R. NAGARAJAN, [2] Dr. P. ARUNA
[1]Programmer (Sr. Scale), Department of Computer Science and Engineering, Annamalai University
[2]Professor, Department of Computer Science and Engineering, Annamalai University
E-mail: [1]rathinanagarajan@rediffmail.com, [2]arunapuvi@yahoo.co.in

ABSTRACT

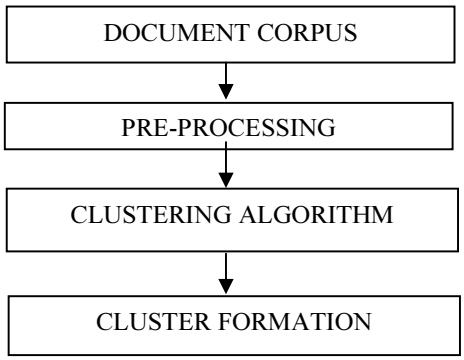
A complexity of the retrieval of relevant document from a large corpus of documents is the most common challenging problem in the areas of web mining and search engines. In addition, the growth of unlabelled and unsupervised documents are also increases this complexity. Document clustering algorithms plays a vital role to reduce this problem. In this paper, an algorithm was proposed to cluster the documents based on their concepts. The proposed algorithm in its first part identifies the vital keywords of the document, those helps to find out the concept of the document, using our new statistical text mining algorithm. In the second part, based on the results derived, the given documents are clustered using Keywords Vs Documents matrix analysis. This proposed Document Clustering algorithm gives more than 90 % accuracy.

Keywords: *Document Clustering, Text Mining, Keyword Extraction, Co-word Analysis, HTML Tags.*

1. INTRODUCTION

Organized documents are very much useful to reduce the time factor as well as accurate retrieval of information against the user's request. A well organization of documents is achieved by grouping or clustering the documents based on their concepts. Statistical text analysis helps us to better clustering of documents. In our new statistical text mining algorithm, a combination of high order frequency terms of the documents and their physical position in the documents are taken to find out the vital keywords of the document. The significance of the terms are identified by the use of HTML Tags. The proposed algorithm relies upon the assumption that, (i) authors of scientific articles choose their technical terms carefully; (ii) when different terms are used in the same articles it is therefore because the author is either recognizing or postulating some non-trivial relationship between their references; and (iii) if enough different authors appear to recognize the same relationship, then that relationship may be assumed to have some significance within the area of science concerned.

Single word never helps to find out the concept of the document. Hence the first part of proposed work find out the significant word pairs or co-words of the document, the co-word construction is done by the calculation of the stronger linkage weight between words, which was achieved by the HTML Tags, For example, Bolder, colored, Italic words are identified by their HTML Tags. Terms in the title, sub title, underlined are also identified and special attention will be given in the algorithm. In the second part of algorithm, with the help of derived keywords, the given documents are clustered by constructing Keyword Vs Document matrix. The paper is organized as follows; brief background discussions on the existing techniques are shown in section 2. Section 3 explains the preprocessing steps before applying clustering algorithm. Section 4 explains the proposed algorithm and document clustering modeling. In Section 5 the experimental results are illustrated. Finally section 6 concludes the paper and discusses some future plans.



ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. RELATED WORK

Usually, document clustering techniques fall into three categories: Graph-based, Hierarchical and Partitioning clustering.

1. Graph based methods model the input documents as vertices of a weighted graph; the edge weight is given by the similarity between two corresponding documents. Accordingly, document clustering problem is turned to graph partitioning based on a certain criteria.

2. Hierarchical clustering, this technique successively builds a tree-like clusters structure which can be constructed in either divisive (top-down) or agglomerative (bottom-up) manner. The main problem with hierarchical clustering is that, it is non-scalable which makes it not suitable for real-time applications and large corpora.

3. Partitioning clustering is to directly partition a dataset into K groups such that documents in a group are more similar to each other than any document from another group. The disadvantage of this technique is that it can converge to a sub-optimal solution. K-mean clustering is a typical example to this kind.

In this section we review previous work on document clustering algorithms and discuss how these algorithms measure up to the requirements of the Web domain. In [1] the steps involved in the construction of text clustering framework and the challenges faced by text clustering algorithms has been discussed. In [2], less similarity based clustering method for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity has been discussed. In [3], the Cuckoo Search Optimization algorithm for Web Document Clustering based on the obligate behavior of some cuckoo species in combining with the levy flight has been discussed. In [4], Bees Algorithm, optimizing the document clustering problem, which simulates the foraging behavior of honey bees swarms in collecting nectar from flower patches by performing global and local search simultaneously; this improves avoidance of local minima convergence has been discussed. In [5], Correlation Preserving Indexing (CPI) method is to find an optimal semantic subspace by maximizing the correlation between the documents in the local patches and simultaneously correlation in the patches outside are minimized has been discussed. In [6], to cluster the documents, the term weight is calculated not only based on their frequency, but also in their physical position of document. The method to construct co-words and the construction of co-words Vs documents matrix has been discussed. In [7] Existing literature and explored the main techniques and methods for automatic documents classification , classifier construction and knowledge extraction and the issues along with the approaches and opportunities has been discussed. In [8], Web Searching algorithms and hierarchical classification techniques for increasing a search service's understanding of web queries has been analyzed. In [9] classification-based algorithm for text detection using a sparse representation with discriminative dictionaries has been proposed and also the adaptive run-length smoothing algorithm and projection profile analysis has been used to further refine the candidate text areas. In [10] the approach of Web Service classification where text mining and classification of WSDL (Web Service Description Language) documents has been analyzed based on association rules i.e. association rules are applied to analyze the degree of dependency between contents of WSDL and category of the Web Services. In [11] Classification of text document to the predefined classes by using various text representation schemes has been discussed. In [12] associative classification, an integrated framework, was used for text categorization, and an algorithm was proposed for text classification uses words as features, to derive feature set from pre-classified text documents has been discussed.

3. METHODOLOGY

3.1 Keyword Construction

In the first part of the proposed algorithm, the documents (Di) taken for analysis are converted in to HTML documents (HDi). Then by analyzing the HTML Tags, the terms of the each document are extracted under three groups. The terms in the titles, subtitle, figures and images are extracted with their frequencies by analyzing the HTML Tags <t>, <t1>, <img> and grouped under the heading First Priority Terms (FPT). Terms with special effects such as Bold, Bold italic, Colored, Underlined are identified with their frequencies and grouped as Second Priority Terms (SPT). To find out the significant terms in the body of document, our proposed algorithm extract the sentences (HDiSj), then the raw terms in each sentence (HTMDiSjTk) with its frequencies are extracted, The Third

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Priority Terms (TPT) are derived by constructing Terms Vs Sentences Matrix. The terms extracted in all the three groups were ranked based on their frequencies.

The terms with high frequency are considered to be very closer to the concept of the documents. By analyzing and comparing the higher order frequency terms in all the three groups the keywords of the documents are derived. The collated keywords were standardized keywords with the help of standard thesaurus. The steps involved in this first part of the algorithm as follows,

Step 1 : Uniform conversion of Documents into HTML documents ($HTMD_i$)

Step 2 : By analyzing the HTML Tags like <T>, <H>,<img> terms in the title, subtitle, tables, figures, image etc., are extracted and grouped under First Priority Terms(FPT)

Step 3 : By analyzing the HTML Tags like <B>,<BI>,<Colored>,Terms with special effects such as Bold, Italic, Bold Italic, colored etc., are extracted and grouped under Second Priority Terms (SPT).

Step 4 : Sentences in the documents areextracted separately and labeled ($HTMD_iS_j$)

Step 5 : Terms in all the sentences are extracted and labeled with its Sentence ID,

Step 6 : Terms Vs Sentence Matrix is Constructed and Third Priority Terms (TPT) are derived with its frequencies.

Step 7 : Third Priority Terms with higher order frequency are compared with other two group terms and the common terms are extracted and labeled as a keywords of the document

3.2 Construction of co-words

A single word is not enough for clustering; hence our proposed algorithm was designed to find out significant word pairs or co-words of the documents. The Standardized keywords derived in the previous stage are taken as input for co-word construction. The prefix and suffix words of our keywords with its frequency are identified, then the binding strength between the prefix and keyword and suffix and keyword are calculated by the formulae

$$S(C_i.C_{ij})=C_{ij}*C_{ij}/C_iC_j 0<S<1 \quad (1)$$

where 'Cij' is the number of co-occurrence of terms i and j. The higher (1) or the lower (0) strength of the co-occurrence is determined by the number of times a co-word pair occurs together in a given document. As the strength of the co-occurrence ranges between 0 and 1, it is necessary to fix certain thresholds in order to filter the 'uninteresting co-occurrences'. This implies that the co-word pairs with a co-occurrence frequency under a certain threshold value are filtered out. The higher the occurrence of the co-word pair, the greater will be the strength of their association and vice-versa. From the strength of association one could identify the semantic relationship existing between the terms. In this manner, co-words for the entire set of sample documents were collected with their parent document identification. Steps involved in co-word construction is summarized as follows

Step 1: For a single keyword prefix and suffix word are extracted from the document

Step 2: Frequencies of the prefix + keyword & suffix + keywords are calculated

Step 3: Matrix of the co-word pairs is constructed based on the higher order frequencies of the co-occurrences of keywords

Step 4: Strength of the linkage 'S' of the co-occurrences are measured by using the formula: $S(C_i.C_{ij})=C_{ij}*C_{ij}/C_iC_j$ $0<S<1$ where 'C_{ij}' is the number of co-occurrence of terms i and j Step 5 : Semantic relationship between the terms are identified

Step5 : Repeat step 1 over step5 for all the other documents

3.3 Keywords Vs Document Matrix Construction

The second part of the algorithm concentrated on the construction of Co-word Vs Document matrix. In this stage, stronger linkage co-words derived from the first part of the algorithm are compared with the documents. The frequencies of the co-words in the documents with their document ID are

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

calculated, and the higher order frequency co-words in all the documents are labeled as higher order frequency co-words. The common top ranking frequency co-word presented documents are clustered. Steps involved in document clustering are summarized as follows

Step 1 : co-words of the higher order frequencies in all the documents are extracted and labeled as high frequency co-words.

Step 2 : cluster are formed by analyzing high frequency co-words Vs Documents

4. EXPERIMENTAL RESULTS

4.1 Keyword Extraction

In this section, the performance of our proposed algorithm had been tested with the corpus of 170 documents in four different areas (Computer Graphics, Computer Networks, Data Mining and Document Clustering).

Step 1: All the taken 170 documents are first converted uniformly into HTML documents and labeled with an unique ID like $HTMD_1$, $HTMD_2$, $HTMD_3$….. $HTMD_{170}$

Step 2 : Based on the HTML tags like <T>, <H>, <img> terms in the title ,subtitle, tables, figures, images documents are extracted under First Priority Terms (FPT) with its document ID, End of this stage 2210 terms were identified as FPT.

Step 3: By analyzing the HTML special effected tags like <B>, <BI>, <Colored> the Second Priority Terms (SPT) are extractedwith its document ID. In this stage 2907terms were identified as SPT.

Step 4 : Sentences of the each and every 170 documents are identified separately with the help of the separator '.', and labeled with their document Id ($HTMD_iS_j$) , End of this stage,7140 sentences were extracted in all documents .

Step 5 : In this stage, Terms in the sentences derived in the previous stage are extracted with their frequencies, then the standard stop words list were compared with derived terms and the matched terms were removed, the remaining terms are summarized with the help of standard thesaurus. End of this stage 42142 terms were extracted.

Step 6: In this stage, for each and every 170 documents, Terms Vs Sentence Matrix was constructed, The idea of the proposed algorithm of this stage is, to find out the identification of terms in more sentences and also labeled as Third Priority Terms.

Step 7: In this final step, Higher order frequency Third Priority Term identified in the previous step were compared with First and Second priority Terms , then the common higher frequency terms in all the groups were identified as the Keyword of a Document. By repeating the Step 2 to Step 7, the keywords of the each and every 170 documents were extracted. For all the 170 documents, 1687 (kwk) keywords with its frequency value were extracted. Figure 1 shows the flow diagram of our Keyword Extraction.

4.2 Co-Words

As per the Second part of the algorithm, the keywords extracted from the first part of the algorithm were taken to construct the Co-words. For each and every keyword (1687), the neighboring terms – both prefix and suffix terms were taken with their frequencies. The strength of the linkage value 'S' of the co- occurrences values of the words were calculated by the formula-1, to view the linkage strength value 'S', the top 10 keyword kw1,kw2,kw3,…kw10 were taken and the table value shows that their linkage .

It is absorbed from the table values that, for the threshold value 0.800 >, the highest linkage strength value of 0.927 says that the words kw4 &kw5 are the stronger valued co-word for the sampledocuments, likewise the value of kw4 & kw9 is 0.900,,the linkage strength value of kw1 & kw4 is0.898 ; the value of kw7 & kw10 is0892, the value of kw2 & kw7 is 0.877, the value of kw1 & kw6 is0.870, the value of kw3 & kw5 is 0.852, the value of kw2 & kw6 is 0.841, the value of kw8 & kw9 is 0.830. It is also absorbed from the table values, the weaker linkage strength of word kw1 are kw1&kw2, kw1&kw3, kw1&kw3, kw1&kw5, kw1&kw7, kw1&kw8, kw1&kw9, kw1&kw10, because their values are less than the threshold value of 0.800 >. Likewise the lower linkage strength co-words are identified by their value and are not taken next stage.

It is interesting to note (vide Table-2) the words Kw1, Kw2, Kw4, Kw5, Kw6, Kw7, Kw9 are contributed in more than one co-word construction, i.e. the contribution of kw1 in co-words Kw1 & Kw4 and Kw1 & Kw6. Likewise Kw2 in co-words Kw2 & Kw6 and kw2 & Kw7. The contribution of Kw4 is high because it forms three co-words kw4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

& kw1, kw4 & kw5 and kw4 & kw9, kw5 contributes to construct Kw5 & Kw3 and Kw5 & Kw4.

4.3 Document Clustering By Constructing Of Co-Words Vs Document Matrix

It is noticed from the table-I that, the highest linkage strength value (0.927) keywords are kw4 & kw5; forms the co-word cw1, the next highest linkage strength value (0.900) keywords are kw7 & kw9 forms the co-word cw2, the keywords kw1 & kw4 (0.898) forms the co-word cw3, kw7 & kw10 (0.892) forms the co-word cw4, kw2 & kw7 (0.877) forms the co-word cw5, kw1 & kw6 (0.870) forms the co-word cw6, kw1 & kw5 (0.852) forms the co- word cw7, kw2 & kw6 (0.841) forms the co-word cw8, the keywords kw8 & kw9 forms the co word cw9.

The occurrences of co-words cw1 to cw9 in our document corpus Di were calculated to find the clusters, the value of occurrences of the co-words in documents are viewed in Table-2.

From the Table-2, for the threshold value 10, the occurrences of cw1, cw3 & cw6 in Documents D1, D4, D5 & D6 are higher than 10, which clearly shows that the documents D1, D4,D5 & D6 discuss the same concepts, and they forms the Cluster -1, likewise the occurrences of the co-words cw4, cw7,cw8 in documents D2, D3,D8 & D9 are also higher than the threshold value 10, hence the documents D2, D3,D8 & D9 deals the same type of concepts and forms the Cluster-2. It is absorbed from the table values that the occurrences values of the co-words cw1 to cw9 are very low in D7,which shows that that the document D7 have no common concept with other Documents.

5 CONCLUSION AND FUTURE WORK

This paper has proposed a new algorithm, not as usual text mining algorithm, it finds the term weight based on both the frequencies and their physical position in the document, and the proposed work also constructs the co words by finding the linkage weight between the words, which helps to construct Co-word Vs Documents matrix to find out the clusters. This study explains the feasibility of co-word construction as a Viable approach for document clustering, On the whole, the problem of complexity of the retrieval of relevant document from a large corpus of documents is reduced through this proposed algorithm. Our future plan will focus on the integration of our proposed algorithm with other evolutionary algorithms and perform the comparison with the other available models.

REFERENCES :

[1] Francis M. Kwale, "An Efficient Text Clustering Framework", *International Journal of Computer Applications*, vol. 79 No. 8., 2013

[2] Manjot Kaur and Navjot Kaur, "Web Document Clustering Approaches Using K-Means Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, No 5, 2013, pp .861-864.

[3] Moe Moe Zaw and Ei Ei Mon, "Web Document Clustering Using Cuckoo Search Clustering Algorithm based on Levy Flight", *International Journal of Innovation and Applied Studies*, Vol. 4 No. 1, 2013, pp.182-188.

[4] Nihal M. AbdelHamid, M. B. Abdel Halim and M. Waleed Fakhr, "Bees Algorithm-Based Document Clustering", *ICIT 2013 The 6th International Conference on Information Technology*,2013.

[5] L. Prabhakar, Srikanth. B, Hierarchical "Document Clustering Using Correlation Preserving Indexing", *International Journal of Computer Science and Mobile Computing*, Vol. 2, No. 8, 2013, pp .237 – 242.

[6] R. Nagarajan and Dr. P. Aruna, "Document Clustering using Keyword standardization and Co-word analysis", *International Journal of Information Technology and Engineering*, Vol. 3 , No 1-2, 2012, pp. 85-89.

[7] Aurangzeb Khan □, Baharum B. Bahurdin, and Khairullah Khan, "An Overview of E-Documents Classification", *International Conference on Machine Learning and Computing IPCSIT,* vol.3 ,2011.

[8] Yogendra Kumar Jain and Sandeep Wadekar, "Classification-based Retrieval Methods to Enhance Information Discovery on the Web", *International Journal of Managing Information Technology (IJMIT)* Vol.3, No.1, 2011.

[9] Ming Zhao , Shutao Li , and James Kwok, "Text detection in images using sparse representation with discriminative dictionaries", *Elsevier: Image and Vision Computing,* Vol. 28, 2010.

[10] Shailja Sharma and Shalini Batra, "Applying Association Rules for Web Services Categorization", *International Journal of*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Computer and Electrical Engineering, Vol. 2, No. 3, 2010.

[11] B S Harish, D S Guru and S Manjunath, "Representation and Classification of Text Documents: A Brief Review", *IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition RTIPPR,* 2010.

[12] Padmavati Shrivastava and Uzma Ansari, "Text Categorization based on Associative Classification", *Special Issue of IJCCT Vol.1 Issue 2, 3, 4; 2010 for International Conference [ACCTA-2010]*, (2010).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

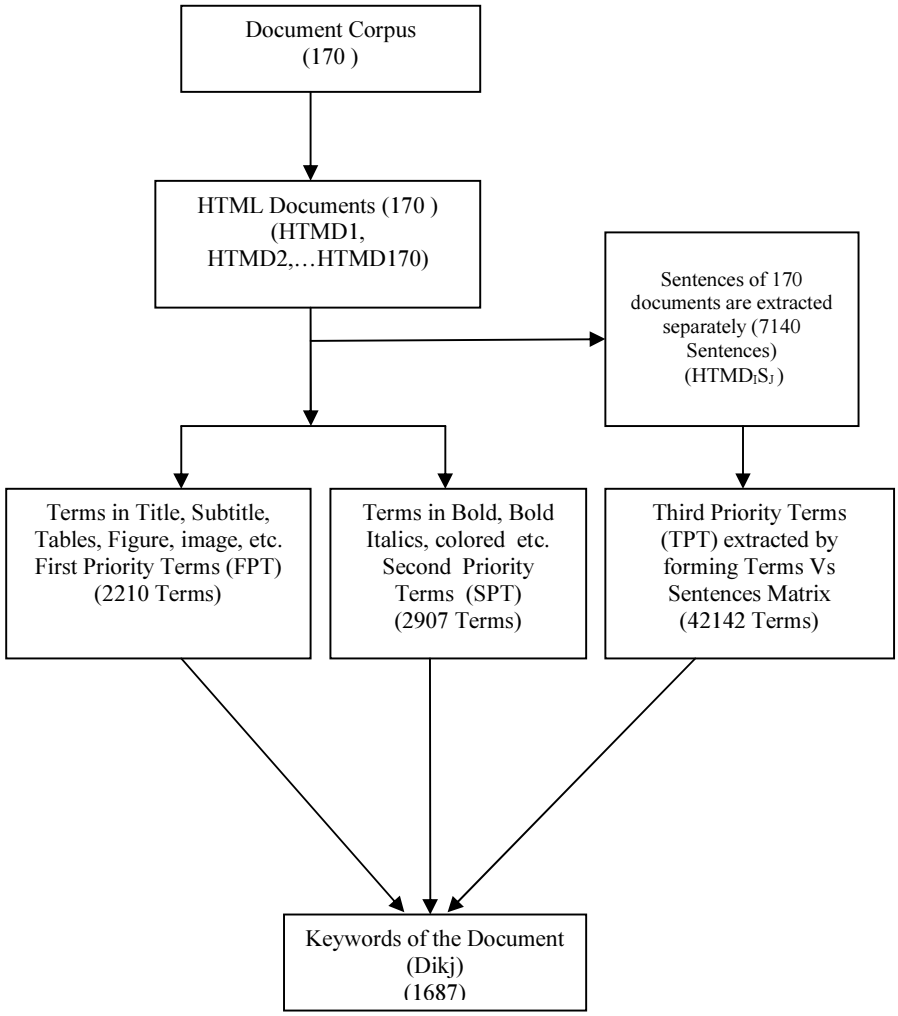


Figure 1: Keyword Extraction

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1 : Construction of co-words(based on thier Linkage Strength 'S")

Kwi	Kw1	Kw2	kw3	Kw4	kw5	kw6	kw7	kw8	Kw9	kw10
Kw1	1	0.158	0.246	0.898	0.356	0.870	0.187	0.207	0.324	0.424
Kw2	0.158	1	0.327	0.302	0.284	0.841	0.877	0.457	0.359	0.419
Kw3	0.246	0.327	1	0.100	0.852	0.182	0.262	0.284	0.245	0.700
Kw4	0.898	0.302	0.100	1	0.927	0.288	0.192	0.192	0.900	0.098
Kw5	0.356	0.284	0.852	0.927	1	0.412	0.321	0.192	0.287	0.306
Kw6	0.870	0.841	0.182	0.288	0.412	1	0.192	0.187	0.176	0.084
Kw7	0.187	0.877	0.262	0.192	0.321	0.192	1	0.626	0.628	0.892
Kw8	0.207	0.457	0.284	0.192	0.192	0.187	0.626	1	0.830	0.626
kw9	0.324	0.359	0.245	0.900	0.287	0.176	0.628	0.830	1	0.324
kw10	0.424	0.419	0.700	0.098	0.306	0.084	0.892	0.626	0.324	1

Table 2 : Co-word Vs Documents

Co-words Vs. Documents	D1	D2	D3	D4	D5	D6	D7	D8	D9
CW1	12	4	-	15	12	16	-	3	1
CW2	7	6	3	8	10	11	-	2	3
CW3	15	1	-	11	10	12	-	-	2
CW4	2	12	11	4	5	3	1	10	13
CW5	2	1	2	1	3	1	2	1	3
CW6	10	3	4	13	11	13	-	3	2
CW7	3	10	12	3	2	1	2	14	14
CW8	2	12	14	4	1	1	1	12	11
CW9	3	9	8	2	3	1	10	7	8