ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 10, October 2014

Document-Document similarity matrix and Multiple-Kernel Fuzzy C-Means Algorithm-based web document clustering for information retrieval

Poonam Yadav

D.A.V College of Engineering. & Technology, India

Abstract: Due to continuous development of World Wide Web, web database are growing massively where automatic grouping of web documents pose a new challenge for researchers to easily retrieve the information. Literature presents different algorithms for web document clustering useful for information retrieval. In this work, Document-Document similarity matrix and Multiple-Kernel Fuzzy C-Means Algorithm-based web document clustering is developed for information retrieval. At first, web documents are read and initial pre-processing are applied to extract the important words. Then, feature space is constructed using keywords and its frequency. Subsequently, document to document similarity matrix is constructed using the similarity measure, called semantic retrieval measure (SR). The measure considers four different criteria, such as, the probability of occurrence in the document, probability of occurrence in the first document, probability of occurrence in the second document and probability of occurrence in both synonyms set. Based on this measure, D-D matrix is computed to do the final grouping using Multiple-Kernel Fuzzy C-Means Algorithm. The experimentation is done with 100 web documents and the results are evaluated with accuracy and entropy.

Keywords: Information retrieval, Similarity measure, web document clustering, Entropy, Accuracy.

I. INTRODUCTION

Web information retrieval system [1-3] desperately need good document clustering algorithm to categorize documents automatically to retrieve information more easily. The advancement of web usage pose a new challenge for the researchers to develop effective document clustering algorithm to obtain effective results [4, 5, 10,11] with less computational task. Document clustering [6-9] is process of grouping web documents automatically based on occurrence of words as well as semantic information. Document clustering can be done in various ways like, partitional clustering and hierarchical clustering. Among these methods, partitional clustering has received significant attention among the researchers due to its various advantages. K-means, fuzzy c-means and kernel methods are some example of partitional methods.

In this paper, partitional clustering algorithm is developed for web document clustering using Document-Document similarity matrix and Multiple-Kernel Fuzzy C-Means Algorithm. At first, input web documents are pre-processed to find document to document feature space which is then given for MKFCM algorithm which utilizes multiple kernels and FCM algorithm for grouping of documents. The experimentation is done with a set of web documents and the performance is analyzed using clustering accuracy and entropy. The paper is organized as follows: Section 2 presents k-means algorithm and section 3 presents the proposed algorithm. Section 4 provides the experimental results and conclusion is given in section 5.

II. K-MEANS CLUSTERING FOR WEB DOCUMENT CLUSTERING

K-means [13, 14] is one of the partitional clustering algorithms widely applied for grouping of data records. Due to various advantages of k-means clustering algorithm, document clustering is also performed with k-means algorithm. In this algorithm, centroids are randomly chosen and it is updated in every steps using average computation. To find the similarity among data records, Euclidean distance is utilized.

Drawbacks: When performing document clustering using k-means algorithm, two major problems can happen. The first problem is finding the similarity among data. The second problem is that it would require more iteration. The first problem can be easily solved with the similarity measure developed in [12] which has computed similarity measurement through the synonyms and frequency. The second problem can be solved using MKFCM algorithm which converge into a better centroids very fastly.

III.DOCUMENT-DOCUMENT SIMILARITY MATRIX AND MULTIPLE-KERNEL FUZZY C-MEANS ALGORITHM TO WEB DOCUMENT CLUSTERING FOR INFORMATION RETRIEVAL

This section presents the proposed document clustering approach using document-document similarity matrix and Multiple-kernel fuzzy c-means clustering. The block diagram of the proposed approach is given in figure 1. From the figure, four different steps are utilized to perform document clustering. In the first phase, web documents are read and initial pre-processing techniques are applied to

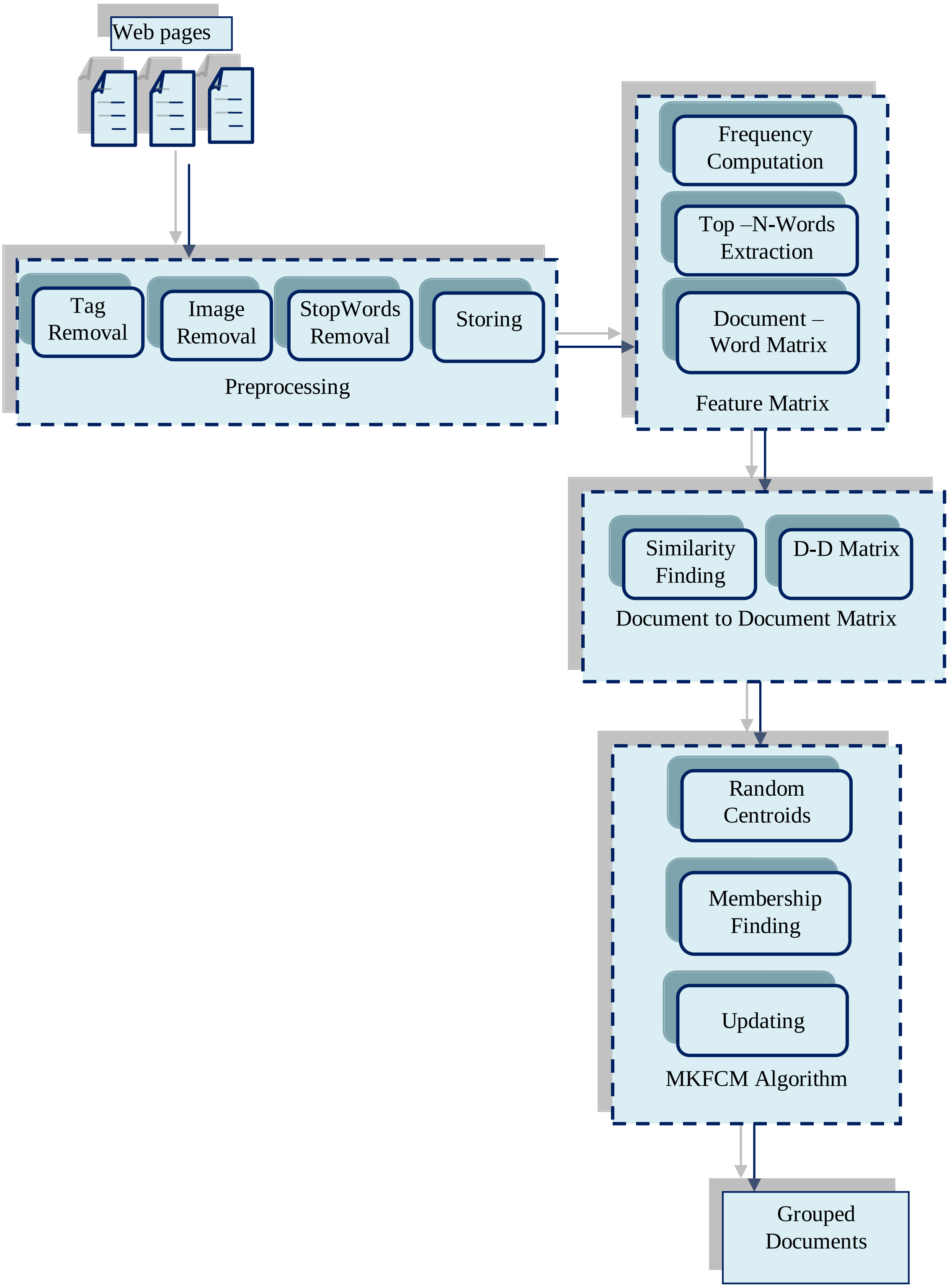

Fig 1. Block diagram of the proposed web document clustering

extract the important words and feature space is constructed using term frequency and keywords. Then, document to document similarity matrix is computed using the similarity measure designed in [12]. Based on this matrix, clustering is performed to group the documents.

A. Pre-processing

Let D_i be web document stored in the web database W which have m number of web documents. Initially, each web documents D_i are taken and pre-processing processes such as tag removal, image removal and stop words removal are applied. Web document which is in the html format is taken to remove the html tags by matching pre-stored html tags. Then, images stored in html document are also removed to extract only the keywords from the html document. After that, stop words like “a, an, the, can, could, may, might” are removed to do the stemming process that converts the words into its original root form.

B. Document-Document similarity matrix computation

Once we obtain keywords from the web document, document to document similarity matrix (D-D) is computed. This matrix is generated by finding similarity among all the documents. The document to document similarity matrix is indicated as D-D matrix which is in the size of $m*m$. Every element within matrix is similarity between two web document having extracted keywords. The similarity is computed based on measure, called semantic retrieval measure (SR) [12]. The proposed measure considers four different criteria, such as, the probability of occurrence in the document, probability of occurrence in the first document, probability of occurrence in the second document and probability of occurrence in both synonyms set. Based on these four criteria, the following formula is formulated.

$$SR_{measure}=\left[\frac{\alpha*P_r\left(D_1,D_2\right)+\beta*P_r\left(D_1,\neg D_2\right)+\gamma*{}_r\left(\neg D_1,D_2\right)+\eta*P_r\left(\not\neg D_1,\not\neg D_2\right)}{\alpha+\beta+\gamma+\eta}\right]$$

Where, α,β,γ,η are constants. The values of $P_r\left(D_1,D_2\right)$, $P_r\left(D_1,\neg D_2\right)$, $P_r\left(\neg D_1, D_2\right)$ and $P_r\left(\not\neg D_1,\not\neg D_2\right)$ are defined as follows,

$$P_r\left(D_1,D_2\right)=\frac{1}{m}\sum_{i=1}^{m}2\left(\frac{f_{D_1}+f_{D2}}{\max\left(f_{D_1},f_{D_2}\right)}\right)$$

$$P_r\left(D_1,\neg D_2\right)=\frac{1}{m}\sum_{i=1}^{m}\frac{f_{D_1}}{\left(f_{D_1}+f_{D_2}\right)}$$

$$P_r\left(\neg D_1, D_2\right)=\frac{1}{m}\sum_{i=1}^{m}\frac{f_{D2}}{\left(f_{D_1}+f_{D_2}\right)}$$

$$P_r\left(\not\neg D_1,\not\neg D_2\right)=\frac{1}{m}\sum_{i=1}^{m}2\left(\frac{\not\neg f_{D_1}+\not\neg f_{D_2}}{\max\left(\not\neg f_{D_1},\not\neg f_{D_2}\right)}\right)$$

From the above equation, m is the unique keywords presented in both the documents, f_{D_1} is frequency of the keywords in D_1, f_{D2} is the frequency of keywords in D_2, $\not\neg f_{D_1}$ represents the frequency of keywords in the synonyms set, $\not\neg f_{D_2}$ is the frequency of keywords in the synonyms set. Here, synonyms set are computed by giving the keywords of document to the wordnet ontology.

- .

C. Web document clustering using Document-Document similarity matrix and Multiple-Kernel Fuzzy C-Means Algorithm

D-D matrix constructed from the previous step is given to MKFCM algorithm [15] for clustering. The objective function of MKFCM algorithm is then reformulated for web document clustering as,

$$OF=\sum_{i=1}^{c}\sum_{j=1}^{m}u_{ij}(1-k(D_j,R_i))$$

Here, $1-k(D_j,o_R)$ can be considered as a similarity measurement derived in the kernel space. In MKFCM algorithm, random representatives are chosen from D-D similarity matrix. In the next step, membership matrix is computed as per the following equation.

$$u_{ij}=\frac{(1-k(D_j,R_i))^{-1/n-1}}{\sum_{l=1}^{c}(1-k(D_j,R_l))^{-1/n-1}}$$

From the membership matrix, new representative is computed as per the following equation.

$$R_i=\frac{\sum_{l=1}^{m}u_{il}*k(D_l,R_i)*D_l}{\sum_{l=1}^{m}u_{il}*k(D_l,R_i)}$$

The kernel function $k(D_i,x_{Rj})$ is calculated based on product of the two Gaussian kernel functions. The definition is given as,

$$k(D_i,x_{Rj})=\exp\left(-\frac{\left|D_i-{}_{Rj}\right|^2}{r^2}\right)\exp\left(-\frac{\left|\bar{D}i-\bar{D}j\right|^2}{r^2}\right)$$

Based on this new representative, membership matrix is updated and this process is iterated until there is no change in the representatives. The final iteration provide the grouped web documents.

IV. Results and discussion

This section presents the experimental results and discussion of the proposed D-D matrix-based MKFCM clustering algorithm This section presents the experimental results and discussion of the proposed D-D matrix-based MKFCM clustering algorithm.

A. Evaluation with clustering accuracy

The proposed D-D matrix-based MKFCM clustering algorithm is implemented with 100 web documents having two groups, one is related with sports articles and other one is related with politics' related articles. Every group contains 50 documents and it is given as input to the algorithm. The clustering output is evaluated with clustering accuracy.

$$CA=\frac{1}{n}\sum_{i=1}^{k}Most_i$$

Where, $Most_i$ is the majority number of documents having identical class labels in the ith cluster, n_i is the number of documents in the ith cluster. The performance plot of the proposed D-D matrix with MKFCM algorithm and D-D matrix with k-means algorithm is given in figure 2. From the figure, we can easily understand that the proposed algorithm providing good accuracy for all the different clusters compared with existing algorithm. For the cluster of six, the proposed algorithm reached of about 85% accuracy as compared with existing algorithm reaches the value of 80%.

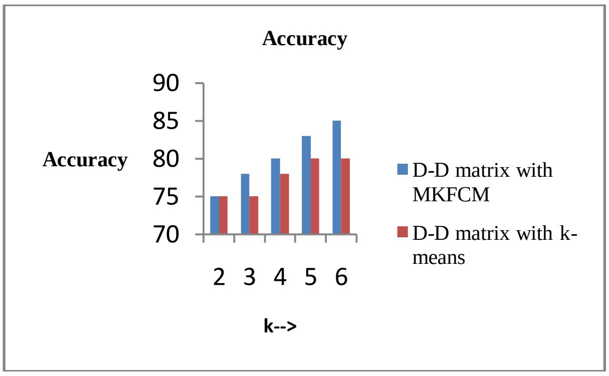

Fig. 2. CA plot in between the proposed and existing

B. Evaluation with entropy

The proposed D-D matrix with MKFCM algorithm is implemented with 100 web documents. The clustering output is evaluated with entropy.

$$E = \frac{\sum_{i=1}^{k} n_i \left(\sum_{j=1}^{p} -\frac{n_i^j}{n_i} \log \frac{n_i^j}{n_i} \right)}{(\log p) n}$$

Where, n_i is the number of documents in the ith cluster, and n_i^j is the number of documents with label j in the ith cluster. $\sum_{j=1}^{p} -\frac{n_i^j}{n_i} \log \frac{n_i^j}{n_i}$ of the numerator calculates the randomness of each cluster. The numerator gives the total sum of randomness contributed by all the clusters. The denominator purports to normalize the E value with maximum being 1.

The performance plot of the proposed D-D matrix with MKFCM and D-D matrix with k-means algorithm based on entropy is given in figure 3. From the figure 3, we can easily understand that the proposed algorithm providing good and less entropy for all the different clusters compared with existing algorithm. For the cluster of two, the proposed algorithm reached the entropy value of 0.8 as compared with existing algorithm reaches the value of 0.85.

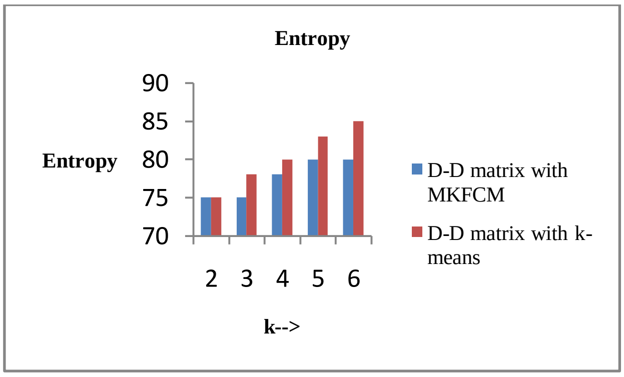

Fig. 3. Entropy plot in between the proposed and existing

V. Conclusion

In this work, Document-Document similarity matrix and Multiple-Kernel Fuzzy C-Means Algorithm-based web document clustering was proposed for information retrieval. At first, web documents are given to pre-processing which extract only the important words. Then, feature space is constructed using keywords and its frequency to find document to document similarity matrix using semantic retrieval measure which considered four different considerations. Based on this measure, D-D matrix was computed to do the final grouping using Multiple-Kernel Fuzzy C-Means Algorithm. The experimentation is performed with 100 web documents and results are evaluated with accuracy and entropy. The proposed algorithm reached of about 85% accuracy as compared with existing algorithm which has reached the value of 80%. Also, proposed algorithm reached the entropy value of 0.8 as compared with existing algorithm reaches the value of 0.85.

References

[1] Kushchu, I., "Web-based evolutionary and adaptive information retrieval", IEEE Transactions on Evolutionary Computation, Vol. 9, NO. 2, PP-117-125, 2005.

[2] Mukherjea, S. ; Bamba, B. ; Kankar, P., "Information retrieval and knowledge discovery utilizing a biomedical patent semantic Web", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, NO. 8, PP. 1099 - 1110, 2005.

[3] Ming Chen ; Hofestadt, R., "Web-based information retrieval system for the prediction of metabolic pathways", IEEE Transactions on NanoBioscience, Vol. 3, NO. 3, PP. 192 - 199, 2004.

[4] Sumiya, K., Kitayama, D. ; Chandrasiri, N.P., "Inferred Information Retrieval with User Operations on Digital Maps", IEEE Internet Computing, vol. 18, no. 4, pp. 70-73, 2014.

[5] Xiaogang Han, Wei Wei ; Chunyan Miao ; Jian-Ping Mei ; Hengjie Song, "Context-Aware Personal Information Retrieval From Multiple Social Networks", Computational Intelligence Magazine, IEEE, vol. 9, no. 2, 2014.

[6] Sanghamitra Bandyopadhyay,"Multiobjective Simulated Annealing for Fuzzy Clustering With Stability and Validity", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 41, No. 5, pp. 682-691, Sept. 2011.

[7] Siti Noraini Sulaiman and Nor Ashidi Mat Isa, "Adaptive Fuzzy-K-means Clustering Algorithm for Image Segmentation", IEEE Transactions on Consumer Electronics, Vol. 56, No. 4, pp. 2661-2668, Nov. 2010.

[8] Pradipta Maji,"Fuzzy–Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 41, No. 1, pp. 222-233, Feb. 2011.

[9] Yun Yang and Ke Chen, "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations", IEEE transactions on knowledge and data engineering, Vol. 23, NO. 2, February 2011.

[10] Junnila, V., Laihonen, T., "Codes for Information Retrieval With Small Uncertainty", IEEE Transactions on Information Theory, vol. 60, no. 2, pp. 976-985, 2014.

[11] Böhm, T, Klas, C.-P. ; Hemmje, M., "ezDL: Collaborative Information Seeking and Retrieval in a Heterogeneous Environment", computer, IEEE, vol. 47, no. 3, pp. 32-37, 2014.

[12] Poonam Yadhav, "SR-K-MEANS clustering algorithm for semantic information retrieval",

[13] Xiaojun Chen ; Xiaofei Xu ; Huang, J.Z. ; Yunming Ye, "TW-k-means: Automated two-level variable weighting clustering algorithm for multiview data", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, NO. 4, PP. 932 - 944, 2011.

[14] Jiye Liang ; Liang Bai ; Chuangyin Dang ; Fuyuan Cao, "The K - Means-Type Algorithms Versus Imbalanced Data Distributions", IEEE Transactions on Fuzzy Systems, Vol. 20, NO. 4, PP. 728-745, 2012.

[15] Long Chen ; Chen, C.L.P. ; Mingzhu Lu, "A Multiple-Kernel Fuzzy C-Means Algorithm for Image Segmentation", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 41, no. 5, pp. 1263-1274, 2011.

BIOGRAPHY

Dr. Poonam Yadav obtained B.Tech in Computer Science &Engg. fromKurukshetra University Kurukshetra and M.Tech in Information Technology from Guru Govind Singh Indraprastha University in 2002 and 2007 respectively. She had Awarded Ph.D inComputer Science& Engg. fromNIMS University, Jaipur. She is currently working as Principal in D.A.V College of Engg. & Technology, Kanina (Mohindergarh). Her research interests include Information Retrieval, Web based retrieval and Semantic Web etc. Dr. PoonamYadav is a life time member of Indian Society for Technical Education.