IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727,
PP 09-11
www.iosrjournals.org

Clustering of Documents Based on Semi-supervised Method

Ms. Dipika.L.Tidke[1], Ms. Madhuri V. Malode[2]
[1](Department of MCA(Engg.), NDMVPS's KBT COE,Nashik./SPPU,Maharashtra, India.
[2](Department of MCA(Engg.), NDMVPS's KBT COE,Nashik./SPPU,Maharashtra, India.

Abstract: *To improve the performance of document clusters ,it is very important to find effectiveand efficient mechanism to minimize the processing time without affecting the total numberof documents given as input required application. The Semi-supervised mechanism usedhere addressed the above need.The proposed approach is designed 1) to group documents into a set of clusters and the number of document clusters formed is determined automatically. 2) To distinguish the discriminative words and non-discriminative words and separate them from unrelated noise words.*

Keywords- Semi-*supervised Clustering, feature partition, Pattern Recognition.*

I. Introduction

Document clustering, means combination of unlabeled text documentsinto significant cluster, is of considerable interest in numerous applications. One assumption, taken by customary document clustering approaches, as in [1], [2],[3], is that the number of clusters N which is to be generated in the process of document clustering is user-defined. N is viewed as a predefined value. However, in realism, to produce the correct value of N is a difficult problem. This is not only time consuming but also impracticable especially when document data sets are bulky. Besides, an incorrect assessment of N might deceived the clustering process. Clustering accuracy reduces considerably if a greater or a lesser number of clusters are used.

Semi-supervised clustering lies in between automatic tagging and auto-organization. It is assumed that it is not essential for the manager is to specify a set of modules, but only to make available a set of texts grouped by the criteria to be used to form the group.Thus if properly prepared, the algorithm is able to remove the noisy terms and to increase the parting among the documents in the different clusters using the consistencies available in the large unlabeled collection. In the experiments the algorithm showed very good performance even when only few starting topics are designated.

The main purpose semi-supervised clustering algorithm is to maximize the throughput power. These algorithms are not just related to maximize the total throughput of the clustering but also time saving. Semi-supervised algorithm is based on the two metrics: i) minimize total processing time. ii) Maximizing efficiency. The first metric focuses on the total time required to generate the clusters based on given threshold value. Second metric focuses on the generation of distinct discriminative words getting high frequency count.

II. Related Work

In [4] ,authors challenge to group documents into an optimum number of clusters while the number of clusters M is revealed mechanically. They develop a Dirichlet Process Mixture (DPM) model to partition documents. It shows promising results for the clustering problem when the number of clusters is unknown. The basic idea of DPM model is to jointly consider both the data likelihood and the clustering property of the Dirichlet Process (DP) prior that data points are more likely to be related to popular and large clusters.

A variational inference algorithm is inspected to assume the document collection configuration as well as the partition of document words at the same time. For the algorithm of variational inference, it could be applied to understand the document collection structure in a much faster way. The Gibbs sampling algorithm is also considered for assessment. However, this is very time consuming process.

Nigam et al. [3] recommended a multinomial mixture model. It relates to the EM algorithm for document clustering supposing that document emphases multinomial distribution. Deterministic annealing procedures [5] are proposed to allow his algorithm to find better local goals of the likelihood function. Though multinomial distribution is often used to model text document, it fails to account for the burstiness occurrence that if a word arises once in a document, it is likely to occur frequently.

In our preliminary work, we proposed the DPMFS approach [8] using the DPM model to model the documents. A Gibbs Sampling algorithm was provided to infer the cluster structure. However, as the other MCMC methods, the Gibbs sampling method for the DPMFS model is slow to converge and its convergence is difficult to diagnose. Furthermore, it's difficult for us to develop effective variational inference method for the DPMFS model.In [9] author's novel algorithm for clustering text documents which exploits the EM algorithm

together with a feature selection technique based on Information Gain. The experimental results show that only very few documents are needed to initialize the clusters and that the algorithm is able to properly extract the regularities hidden in a huge unlabeled collection.

III. System Architecture

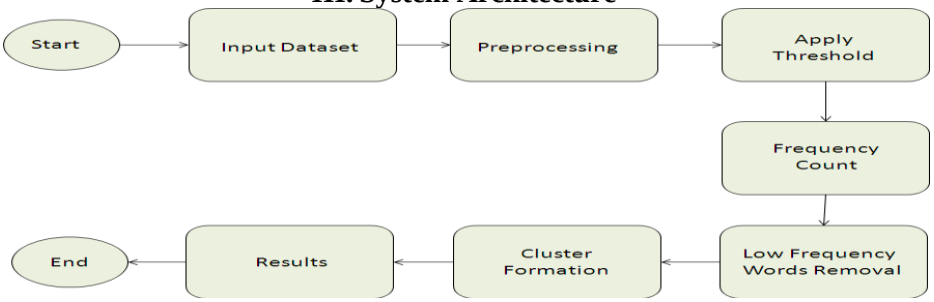

Fig 1: Block diagram of system

Following Steps need to be performed for clustering.
Step 1: Get user defined path for input files.
Step 2: Sort input files according to their mime class.
Step 3: Read all words from ignore file and store it in an array. Here, Ignore file is a file which consist of list of stop- words that are used to remove noisy words from given input file.
Step 4: Read all words from all input files and store it in an array. All words are stored in an array format A[1….N].
Step 5: Remove stop words from Ignore array and perform stemming operation.
Step 6: Remove distinct keywords from array i.e those words those have frequency count as 1.
Step 7: Calculate frequency of remaining words.

$$f(t,D) = \log \frac{N}{|\{d \in D : t \in D\}|} \quad (1)$$

where,
N= Number of documents.
$\{d \in D : t \in D\}$ = number of documents where the term tappears

Step 8: Calculate DMAF value which is frequency vector of discriminative words which is given by.

$$E_q = [\log f(W,X/\Theta)] \quad (2)$$

Step 9 :Check threshold frequency Θ and create clusters.

IV. Results

In this approach, we have added two more features to semi- supervised technique.

- *Search operation*

In this we can search any particular documents by giving a particular keyword as input file.

- *Time taken*

Here time taken by this technique to generate the clusters is shown in milliseconds of time. From this we can easily prove that time taken by semi-supervised technique to generate the clusters is much less as compared to gibb's samplings theorem.
The following fig shows the graphical result for semi-supervised clustering algorithm. Where X-axis represents number of input files and Y-axis represents numbers of clusters generated.

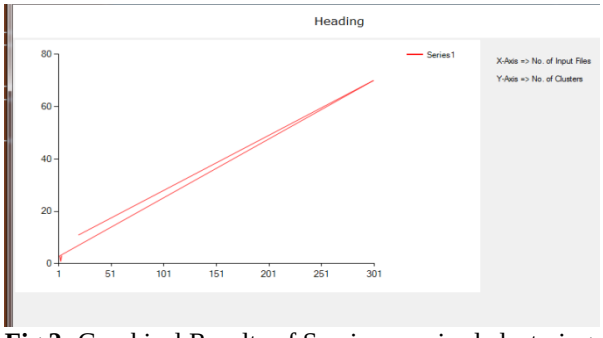

Fig 2: Graphical Results of Semi-supervised clustering

The following fig shows the graphical result for comparison between semi-supervised clustering algorithm and DMAFP. Where X-axis represents number of input files and Y-axis represents numbers of clusters generated.

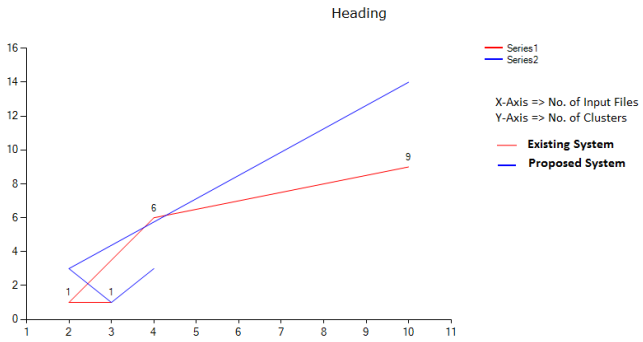

Fig 3. Comparison between Semi-supervised system and DMAFP system

V. Conclusion

We have seen that following targets will definitely achieve as follows if we will form a set or clusters of given documents.So it will very useful to have clusters of data based on some similarity. In our proposed system we will use Dirichlet Process Mixture Model, mean variance algorithm and blocked gibbs sampling algorithm. Our proposed system with semi-supervised clustering technique tells us that time taken by semi-supervised technique to generate the clusters is much less as compared to DMAFP algorithm. Also here we have added two more features i.e we can apply searching operation to search a particular document by giving a keyword as input. And also we have shown time taken by different documents to generate the clusters in milliseconds. Hence we can conclude that semi-supervised technique is much faster to form clusters.

References

[1] C. Elkan, "Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution," Proc. Int'l Conf. Machine Learning, pp. 289-296, 2006.

[2] R. Madsen, D. Kauchak, and C. Elkan, "Modeling Word Burstiness Using the Dirichlet Distribution," Proc. Int'l Conf. Machine Learning, pp. 545-552, 2005.

[3] K. Nigam, A.K. McCallum, S. Thrun, and T.M. Mitchel, "Text Classification from Labeled and Unlabeled Documents Using Em," J. Machine Learning, vol. 39, no. 2, pp. 103-134, 2000.

[4] Ruizhang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi," Dirichlet Process Mixture Model for Document Clustering with Feature Partition", IEEE Trans. On knowledge and data engineering, vol. 25,no. 8, August 2013

[5] K. Rose, "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems," Proc. IEEE, vol. 86, no. 11, pp. 2210-2239, Nov. 1998.

[6] P. Smyth, "Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood," Statistics and Computing, vol. 10, no. 1, pp. 63-72, 2000.

[7] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freedman, "Autoclass: A Bayesian Classification System," Proc. Int'l Conf. Machine Learning, pp. 54-64, 1988.

[8] G. Yu, R. Huang, and Z. Wang, "Document Clustering via Dirichlet Process Mixture Model with Feature Selection," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining, pp. 763-772,2010.

[9] Leonardo Rigutini, Marco Maggini,"A Semi-supervised Document Clustering Algorithm based on EM", Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05) 0-7695-2415-X/05 $20.00 © 2005.

[10] The description of the 20-Newsgroups data set can be found at http://people.csail.mit.edu/jrennie/20Newsgroups.