ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Partitioning The Documents Based On Semi-supervised Clustering Method.

Madhuri V. Malode[1], Prof. J.V.Shinde[2]

M. E. Student, Dept. of Computer Engg.Late G.N.Sapkal COE,Nashik, Maharashtra,India[1]

Assistant Professor, Dept. of Computer Engg.Late G.N.Sapkal COE,Nashik,Maharashtra,India[2]

ABSTRACT: One crucial role of document clustering is to examine the number of clusters in an appropriate way from the given dataset to which documents should be partitioned. In this paper, we propose a novel approach, namely Semi-supervised method of document clustering, to address this issue. 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.Our research indicates that our proposed approach performs fine on the man-made data set as well as actual data sets. The comparison between our approach and Dirichlet process mixture model document clustering approaches shows that our approach is vigorous and operative for document clustering.

KEYWORDS: Database management, pattern recognition,stemming, Semi-supervised Clustering, feature partition.

I. INTRODUCTION

Document clustering, means combinationof unlabeled text documentsinto significant cluster, is of considerableinterest in numerous applications.One assumption, taken bycustomary document clustering approaches, as in [1], [2], [3], is that the number of clusters Nwhich is to be generated in the process of document clustering is user-defined. Nis 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 impracticableespecially when document data sets are bulky.Besides, an incorrect assessment of Nmight deceived the clustering process. Clustering accuracy reducesconsiderably if a greater or a lesser number of clusters are used.

Semi-supervised clustering lies in between automatictagging and auto-organization. It is assumed that it is not essential for themanager is to specify a set of modules, butonly to make available a set of texts grouped by the criteria to beused to form the group.Thus if properly prepared, thealgorithm is able to remove the noisy terms and to increasethe parting among the documents in the different clustersusing the consistencies available in the large unlabeledcollection. In the experiments the algorithm showed very good performance even when only few starting topics aredesignated.

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 algorithmis based on the two metrics: i) minimize total processing time. ii) Maximizingefficiency. 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 revealedmechanically. 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 relatesto 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 burstinessoccurrence that if a word arises once in a document, it is likely to occur frequently.

Madsen et al. [2] used theDCM model to capture burstiness well. Its researchdisclosed that the performance of DCM was equivalent tothat obtained with multiple experimentaldeviations to themultinomial model. However, DCM model lacks perceptivenessand the restrictions in that model cannot beassessedrapidly.

Elkan[1] derived the EDCM distributionwhich belongs to the exponential family. It is a well-intentionedcalculation to the DCM distribution. The EM algorithmwith the EDCM distributions is much quicker than thecorresponding algorithm with DCM distributions offeredin [2]. It also achieves high clustering accuracy. Inrecent years, EM algorithm with EDCM distribution is themost viable algorithm for document clustering if thenumber of clusters is predefined. If the number of clusters K is unknown before the clustering process, one solution is to estimate N first and use this estimation as the input parameter for those document clustering algorithms requiring N predefined. Many methods have been introduced to find an estimation of N. The most straightforward method is the likelihood cross-validation technique [6], which trains the model with different values of K and picks the one with the highest likelihood on some held-out data. Another method is to assign a prior to K and then calculate the posterior distribution of K to determine its value [7].

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 difficultto diagnose. Furthermore, it's difficult for us to developeffective variational inference method for the DPMFSmodel.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. PROBLEM STATEMENT

When more and morelabeled documents are available in real life, processing capacity of variational inference algorithm and gibb's sampling theorem degrades. It becomes a time consuming process. To overcome this, and to improve the performance of our approach of semi-supervised document clustering came into existence. With this approach input documents many vary from hundreds to thousands, also it is a time saving process. Not only clustering but also some additional information has been generated in this technique. Additional information are as follows: 1) We can search a particular file among given dataset by giving related keyword as input.2) We can estimate the time(in milliseconds) needed to generate the clusters very easily.3) We can compare our proposed results with variational inference algorithm and gibb's sampling theorem approach in a graphical manner. From this we can evaluate that our approachis more effective and faster.

IV. System Architecture

Basic flow of our system is as follows:

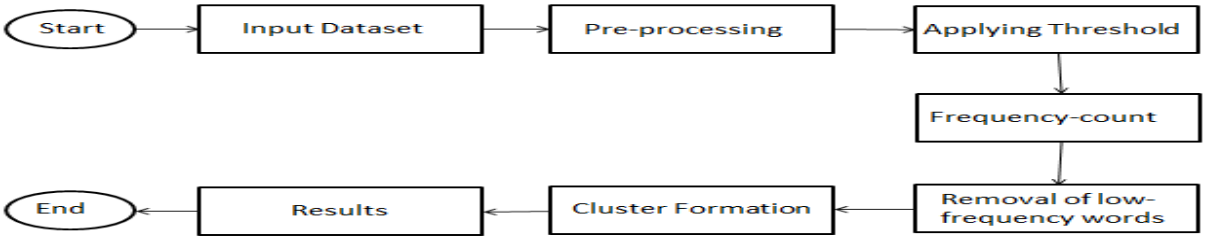

Fig 1.Flow diagram.

In proposed system we have evaluated our result from dataset which is taken from 20-Newsgroups.[10]. Two real document data sets were used for evaluating our proposed approach, in particular, News-different-3, Newssimilar- 3. News-different-3 consists of 300 messages from three newsgroups on relatively different topics (alt.atheism, rec.sport.baseball, sci.space) with well-separated clusters. News-similar-3 consists of 300 messages from three newsgroups on similar topics (comp.graphics, comp. os.ms-windows, comp.windows.x) where cross-posting often occurs.Once input is given to the system one step is to perform pre-processing in which we remove the stop-words and stemming of words takes place. After pre-processing, we calculate the frequency count of each discriminative words by applying some threshold value. Words with low frequency count than threshold are removed and those with high frequency are further proposed. Clusters are formed with words those have high frequency count.

V. Pseudo Code

```
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 arrayi.e those words those have frequency count as 1.
Step 7: Calculate frequency of remaining words.
        $f(t,D) = \log\frac{N}{|\{d \in D : t \in D\}|}$        (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/Θ)]        (2)
Step 9 :Check threshold frequency Θand create clusters.
```

VI. Simulation Results

In proposed system we have applied new technique to generate the clusters which is semi-supervised clustering. Here The supervisor onlyneeds to give a reasonable initialization for the cluster ”centres”without the need to define a set of explicit categories. The algorithm is able to remove the noisy terms i.e stop-words stand to improvethe separation among the documents (discriminative and non-discriminative) in the different clustersusing the regularities available in the large unlabelledcollection. In the experiments the algorithm showed verygood performance than gibb's sampling theorem.

Here , 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 are 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.

Thresholds forremoving high-frequency and low-frequency words forNews-different-3 and News-similar-3 data sets were set 100 to 150.We evaluate our proposed approach, namely Semi-supervised clustering algorithm with existing approach namely DMAFP,based on the variational inference algorithm. The settingof initial values of these hyper-parameters is arbitrarybecause all of them are updated during the clusteringprocess by the method proposed in Section 5.3. We set N to30 for the data sets News-different-3, News-similar-3.

Table no 1
Comparison of the Document Clustering Performance
on the News-Different-3, the News-Similar-3

Approach	News – different -3	News-similar-3
Semi-supervised Algorithm	0.93	0.91
DMAFP	0.80	0.56

For eachdata set, we conducted experiments 20 times and chose theresult which acquired the largest value of equation (2).Table 1 depicts the document clustering performance acquired by the Semi-supervised algorithm and DMAFP approaches on the News-different-3, the News-simlar-3data sets. The experimental results show that our proposed Semi-supervised approachachieves better performance.

Table 2 shows the number of clusters estimated by ourproposed Semi-supervised approach. The DMAFP approach is alsoinvestigated for comparison analysis. From Table 2, itshows that our estimation for the number of clusters arerelatively bigger than the true one. The reason is that thereare a number of outlier documents in the real documentdata set. These outlier documents are dissimilar with otherdocuments belonging to the same cluster and are regardedas belonging to new clusters in the semi-supervised approach. Thesame effect could be easily achieved when documents aremanually partitioned into groups. The Semi-supervised approachacquires more precise estimation compared with the DMAFP approach. Therefore, partitioning discriminative words andnondiscriminative words is useful for estimating thenumber of clusters N.

TABLE 3
Estimated Number of Clusters on the News-Different-3, the
News-Similar-3

Approach	News – different -3	News-similar-3
Semi-supervised Algorithm	14	08
DMAFP	25	09

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.

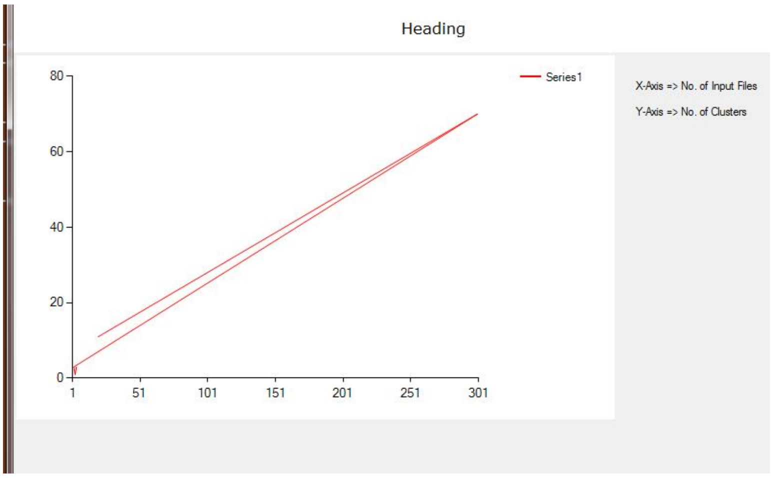

Fig 2 Graphical result of proposed system.

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.

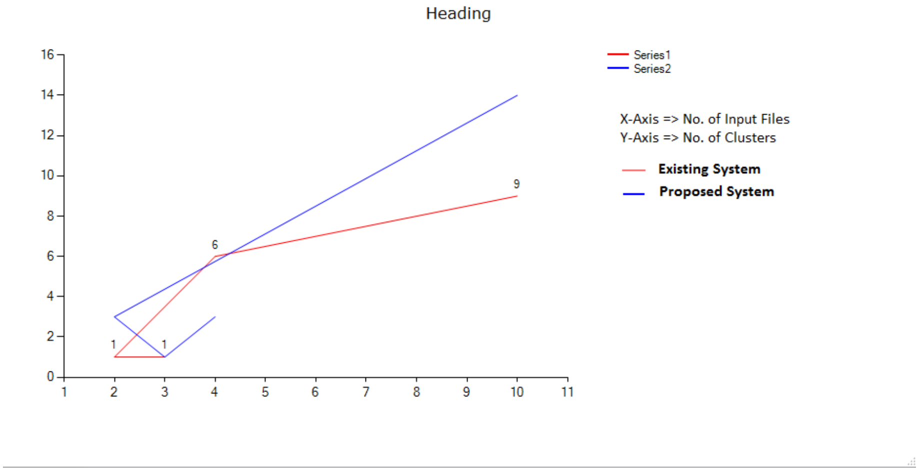

Fig 3. Comparison between proposed system and existing system

VII. Conclusion and Future Work

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.

DOI: 10.15680/ijircce.2015.0306076

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 WordBurstiness 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 forDocument 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.

BIOGRAPHY

Madhuri V. MalodeP.G. student: Department of Computer Engineering, Late G.N. Sapkal College of Engineering, Anjaneri, City-Nasik, Country-India. University: SavitribaiPhule Pune University.

Prof J.V.Shinde Associate Professor: Department of Computer Engineering, Late G.N.Sapkal College of Engineering, Anjaneri, City: Nasik, Country: India. University: SavitribaiPhule Pune University. She has presented papers at National and International conferences and also published paper in national and international journals on various aspects of the computer engineering.