Content And Featured Based Document Clustering

[1] Zete Sonali, [2]Rajole Rohini, [3]Deshmukh Shital,[4]Tile Swati
[1]BE Student,[2]BE Student,[3]BE Student,[4]BE Student
[1,2,3,4]*Computer Engg. Dept. Of Matoshri college of Engineering and Research Center Eklahare, Nasik,India,Pin:422105*

***Abstract*—**Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users. Finding the appropriate number of clusters to which documents should be partitioned is crucial in document clustering. In this will focus on various clustering techniques and our proposed system to discover the cluster structure without requiring the number of clusters as input. Document features or even we can say that the various attributes will be automatically partitioned into two groups, in particular, discriminative words and non-discriminative words, and contribute differently to document clustering. There is one variation inference algorithm which have studied to infer the document collection structure as well as the partition of document words at the same time. Will justify at the end in the conclusion how our approach will perform well on the data set. Then will justify our systems accuracy and efficiency by describing what we have proposed with the predicted modules and features for the system.

***Index Terms*—**Database management, database applications-text mining, pattern recognition,ss
Clustering, document clustering, feature partition

1.INTRODUCTION

Clustering is the process of organizing given objects into certain number of groups whose members are similar in some way therefore cluster contain same and different type of object. The proposed approach is handled document clustering and feature partition simultaneously. In existing document clustering approach the number of cluster K is known before the document clustering. K is the predefined parameter resolve by user. However in reality resolve the appropriate value of K is difficult task. In which first given set of documents, user have to browse the whole document collection to resolve the value of K. This is very time consuming and unrealistic when work on large document. In proposed system it attempt group the documents into optimal number of cluster while the number cluster K is discovered automatically. For this uses Dirichlet Process Mixture (DPM) model to partition documents. It shows the promising results for the clustering problem when the number of cluster is unknown.

1.1 **Justification**

Document clustering is the process of grouping the different text document into meaningful cluster. One assumption taken by existing process or methodology is that the number of cluster K is known before the process of document clustering. K is predefined parameter resolve by user. However in reality, resolve the appropriate value of K is difficult task. First, users have to browse whole document collection in order to find K. This is very time consuming especially when dealing with large document data set. The improper finding the value of K might easily misguide the clustering process. Therefore, it is very useful if document clustering method could be design relaxing the assumption of predefined K.

1.2 Scopes

The scope of system is to decrease amount of data by grouping similar data item and present them collectively. The user start at the top of the list and follows it down examining one result at a time, until the sought information has been found. Third method is search results clustering, which consist of grouping the results return by a search engine into hierarchy of labeled clusters. Mainly the Document clustering is used for document classification and document searching. It also using marketing, biology, libraries, and insurance for the data sorting and searching process.

2. EXISTING SYSTEM

In existing document clustering process the number of cluster K is known before the process of document clustering. The K is predefined parameter resolve by user. In reality resolving appropriate value of K is difficult task. To resolve appropriate value of K the user will browse the whole document collection. In exiting system used the Gibbs Sampling Algorithm for document clustering. The limitation for this algorithm, it required more time to form cluster and value of cluster K known before clustering.

3. PROPOSED SYSTEM

As earlier, we discuss in proposed approach handles document clustering an featured partition simultaneously. It attempts to group documents into optical number of cluster while the number of cluster K is discover automatically. It uses Dirichlet

Process Model to partition documents. The goal of document clustering is to minimize intra cluster distance between documents, while maximizing inter cluster distance. In this first the developed Dirichlet Process Model to partition document which is show result for clustering problem when number of cluster unknown. We will adopt our proposed approach for semisuperwised document cluster.

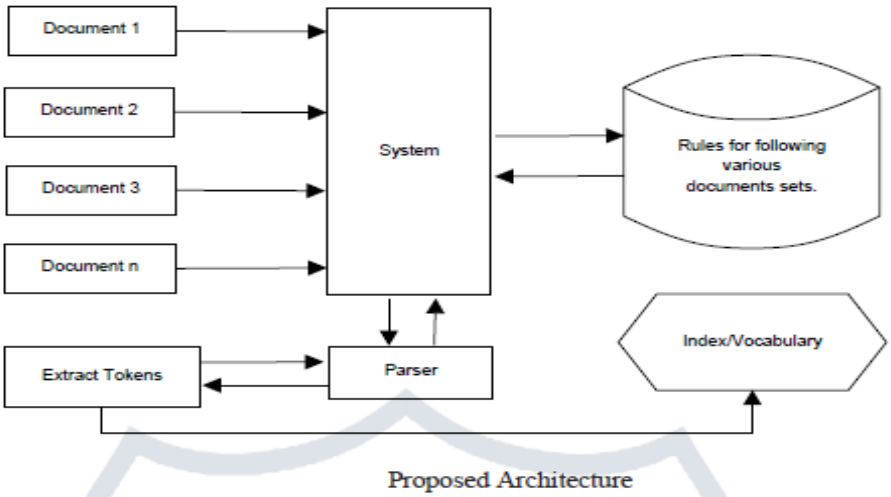

Figure 1.Proposed Architecture

The figure 1 shows n numbers of documents are gives inputs to the system, then system can provide this inputs to the parser. The parser will extract the tokens from it. And it forwarded to the system and extracted tokens are stored in the index. The system can provide the rules for various documents set that apply on the documents.

4. SYSTEM ARCHITECTURE

The figure 2 shows that system architecture, in that different types of files are documents as input to the system. To identify the file type then it can generate the tokens of it. From the tokens extract the similar keyword and calculate the frequency from it. From the frequency count cluster can be form and from it mapping of documents to clusters can perform.

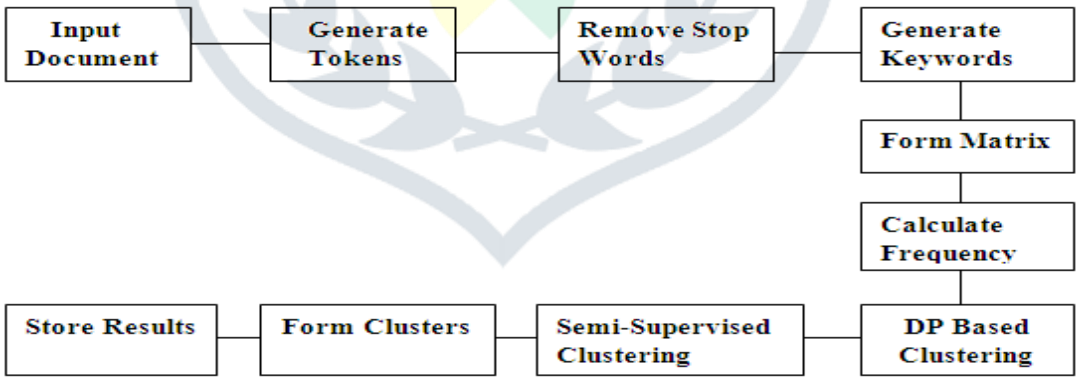

Figure 2. System Architecture

5. ALGORITHMS

5.1 Blocked Gibbs Sampling Algorithm

Another effective inference algorithm for our proposed model is the blocked Gibbs sampling algorithm. Many authors have used this method to infer various model based on the DP prior. For the DMAFP model, the state of the Markov chain is W=$\{\gamma, P, \eta_0, \eta_1, \ldots, \eta_N, z_1, \ldots, z_D\}$. Let $\{z_1^*, \ldots, z_M^*\}$ denote the set of distinct values of $\{z_1, z_2, \ldots, z_D\}$ and the hyper parameter θ, the blocked Gibbs sampling procedure iterate between the following steps:

1. Update the latent discriminative words indicator γ by repeating the following Metropolis step R times: A new candidate γ new which adds or deletes a discriminative word is generated by randomly picking one of the W indices in γold and changing its value. The new candidate is accepted with the probability

$$\min\left\{1, \frac{f(\gamma_{new}|\chi, z)}{f(\gamma_{old}|\chi, z)}\right\}, \tag{12}$$

where $f(\gamma|\chi, z) \propto f(\chi|z, \gamma)p(\gamma)$ and $f(\chi|z, \gamma)$ is given by (11).

2. Conditioned on other latent variables, for i =1, 2 . . . N, if i is not in $\{z_1^*, z_2^*, \ldots, z_M^*\}$, draw ηi from a Dirichlet distribution with parameterλ. Otherwise, update ηi by sampling a value from a Dirichlet distribution with parameter:

$$\left\{\lambda_1 + \sum_{\{d: z_d = i\}} x_{d1}\gamma_1, \ldots, \lambda_W + \sum_{\{d: z_d = i\}} x_{dW}\gamma_W\right\}. \tag{13}$$

3. Update η0 by sampling a value from a Dirichlet distribution with parameter:

$$\left\{\beta_1 + \sum_{d=1}^{D} x_{d1}(1 - \gamma_1), \ldots, \beta_W + \sum_{d=1}^{D} x_{dW}(1 - \gamma_W)\right\} \tag{14}$$

4. Update P by sampling a value from a Dirichlet distribution with parameter:

$$\left\{\frac{\alpha}{N} + \sum_{d=1}^{D} I(z_d = 1), \ldots, \frac{\alpha}{N} + \sum_{d=1}^{D} I(z_d = N)\right\}, \tag{15}$$

Where *I*(zd =*i*) is an indicator function which equals to 1 if *zd =i*.

5. Conditioned on other latent variables, for d = 1, 2. . . D, update zd by sampling a value from a discrete distribution with parameter {sd1, sd2 . . ., sdN}, where

$$\sum_{i=1}^{N} s_{di} = 1 \quad \text{and} \quad s_{di} \propto p_i f(x_d|\eta_i, \eta_0, \gamma), i = 1, \ldots, N. \tag{16}$$

After the Markov chain has reached its stationary distribution, we collect H samples of {z1 . . ., zD} and {γ1, . . . ,γW}. Latent document labels and the partition of document words are then estimated as follows:

1. The estimated label of document xd is the most frequent value of zd in the last H samples.
2. The word wj is discriminative if the average value of the last H sample of γj is bigger than a threshold "(We set ε as 0.7 in our experiments.) Otherwise, wj is regarded as no discriminative.

5.2 Variational Inference Algorithm

We use the mean field variational inference algorithm to approximate posterior distribution of the latent variables $\bar{W} = \{\gamma, P, \eta_0, \eta_1, ..., \eta_N, z_1, z_2, ..., z_D\}$ in the DMAFP model. In this setting, the hyper parameter is $\theta = \{\alpha, \omega, \beta, \lambda\}$. It is very natural to choose the mean field variational approximations Q as the following family of distributions:

$$q_\nu(\bar{W}) = q_\sigma(P) q_{\tau_0}(\eta_0) \prod_{j=1}^{W} q_{\omega_j}(\gamma_j) \prod_{i=1}^{N} q_{\tau_i}(\eta_i) \prod_{d=1}^{D} q_{\phi_d}(z_d) \tag{1}$$

Where $q_\sigma(P)$ a Dirichlet distribution with parameter is $(\sigma_1, \ldots, \sigma_N)$. $q_{\tau_i}(\eta_i)$ is a Dirichlet distribution with parameter $(\tau_{i1}, \ldots, \tau_{iW}), i = 1, 2, \ldots, N.$ $q_{\omega_j}(\gamma_j)$ is a Bernoulli distribution with parameter

$\omega_j, j = 1, 2, \ldots, W.$ $q_{\phi_d}(z_d)$ is a multinomial distribution with parameter $(\phi_{d,1}, \ldots, \phi_{d,N}), d$ = 1, 2…D. In this case, the free variational parameters are

$$\nu = \{\sigma, \omega_1, \ldots, \omega_W, \tau_0, \tau_1, \ldots, \tau_N, \phi_1, \ldots, \phi_D\}. \tag{2}$$

In order to acquire a good approximation for the posterior distribution of the latent variables $\bar{W}$, we need to iteratively update the free variational parameters ν and maximize the lower bound on the log marginal likelihood as follows:

$$\log p(\chi|\theta) \geq E_{q_\nu}[\log f(\bar{W}, \chi|\theta)] - E_{q_\nu}[\log q_\nu(\bar{W})]. \tag{3}$$

Therefore, the lower bound of the log marginal likelihood is as follows:

$$\begin{aligned} L &= E_{q_\nu}[\log f(\bar{W}, \chi|\theta)] - E_{q_\nu}[\log q_\nu(\bar{W})] \\ &= E_{q_\nu}[\log f(\chi|\bar{W}, \theta)] + E_{q_\nu}[\log f(\bar{W}|\theta)] - E_{q_\nu}[\log q_\nu(\bar{W})]. \end{aligned} \tag{4}$$

To maximize the lower bound L, the update equations for ν are as follows:

$$\sigma_i = \frac{\alpha}{N} + \sum_{d=1}^{D} \phi_{d,i} \tag{5}$$

$$\tau_{0j} = \beta_j + \sum_{d=1}^{D} x_{dj}(1 - \omega_j) \quad (6)$$

$$\tau_{ij} = \lambda_j + \sum_{d=1}^{D} x_{dj}\omega_j\phi_{d,i} \quad (7)$$

$$\phi_{d,i} = \exp\{M_{d,i}\} \quad (8)$$

$$\omega_j = \frac{\exp(A_j)}{1 + \exp(A_j)}, \quad (9)$$

where $i \in \{1,\ldots,N\}, j \in \{1,\ldots,W\}, d \in \{1,\ldots,D\}$ and

$$M_{d,i} = \sum_{j=1}^{W} x_{dj}\omega_j\left(\psi(\tau_{ij}) - \psi\left(\sum_{j=1}^{W}\tau_{ij}\right)\right) + \psi(\sigma_i) - \psi\left(\sum_{i=1}^{N}\sigma_i\right) - 1 \quad (10)$$

$$A_j = \log\frac{\omega}{1-\omega} - \sum_{d=1}^{D} x_{dj}\left(\psi(\tau_{0j}) - \psi\left(\sum_{j=1}^{W}\tau_{0j}\right)\right) + \sum_{d=1}^{D}\sum_{i=1}^{N} x_{dj}\phi_{d,i}\left(\psi(\tau_{ij}) - \psi\left(\sum_{j=1}^{W}\tau_{ij}\right)\right). \quad (11)$$

The digamma function is denoted by which arises from the derivative of the log normalization factor in the Dirichlet distribution. Repeatedly updating the variational parameter through (5) to (9) would increase the low bound L and finally acquire local maxima of L. When the improvement of L is less than a threshold, we estimate the latent clustering structure and the partition of document words by the variational parameter $\{\phi_{d,i}, d = 1,\ldots,D, i = 1,\ldots,N\}$ and $\{\omega_1,\ldots,\omega_W\}$, respectively. The cluster to which the document Xd belongs is determined by the value of $\phi_{d,i}$. In particular, let ϕ_{d,i^*} is the largest value acquired by the document Xd, Xd will then be assigned to the cluster labeled by i. The word wj is discriminative when j is larger than a threshold. Otherwise, wj is regarded as no discriminative. Note that as the traditional EM algorithm, our proposed mean field variational inference method yields local maxima. For practical applications, we run the algorithm multiple times with different initial values. Some authors suggested that the inference should base on the result which acquires the largest L as shown in (5) [2]. However, for our proposed variational inference algorithm designed for the document clustering, our large amounts of experiments indicate that an effective way to choose a good result is to choose the one which acquires the largest value of $E_{q_v}[\log f(\bar{W}, \chi|\theta)]$.

Since $E_{q_v}[\log f(\bar{W}, \chi|\theta)] = E_{q_v}[\log f(\bar{W}|\chi, \theta)] + \log f(\chi|\theta)$,

This method tends to choose the result which reaches the highest posterior likelihood of the latent variables. The calculation of $E_{q_v}[\log f(\bar{W}, \chi|\theta)]$ is shown in available in the online supplemental material.

6. CONCLUSION

In this paper, we proposed the system which handled document clustering and featured partition simultaneously. Document clustering process is based on DPM model which groups document into number of clusters. Both the variational inference algorithm and blocked Gibbs sampling algorithm are proposed to conclude the cluster structure as well as different data set. The comparison between our proposed system and existing method indicate that our system is robust and effective for document clustering. This will improve the document clustering quality.

REFERENCES

[1] "Dirichlet Process Mixture Model for Document Clustering with Feature Partition", Ruiz hang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi.

[2] **"**Recent trend in hierarchical document clustering: critical", Peter Willett.

[3] "Clustering Web Documents using Hierarchical Method for Efficient Cluster Formation", I. Ceema, M. Kavitha, G. Renukadevi, G. sripriya, S. Rajesh Kumar.

[4] "Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution", Charles Elkan.

[5] "Incremental Hierarchical Clustering of Text Documents ",Jamie Callan, Ramayya Krishnan

[6] "Document Clustering", Jajoo.