A Semi-Supervised Text Clustering Algorithm with Word Distribution Weights

Ping Zhou[1, a], Jiayin Wei[1, b], Yongbin Qin[1, c]

[1]College of Computer Science and Information, Guizhou University, Guiyang, 550000, China

[a]email: pingzhou329@163.com, [b]email: weijiayin05@sina.com, [c]email: ybqin@foxmail.com

Keywords: Text Clustering; Semi-supervised Clustering; LDA; Word Distribution

Abstract. Semi-supervised text clustering, as a research branch of the text clustering, aims at employing limited priori knowledge to aid unsupervised text clustering process, and helping users get improved clustering results. Because labeled data are difficult, expensive and time-consuming to obtain, it is important to use the supervised information effectively to improve the performance of clustering significantly. This paper proposes a semi-supervised LDA text clustering algorithm based on the weights of word distribution (WWDLDA). By introducing the coefficients of word distribution obtained from labeled data, LDA model can be used in the field of semi-supervised clustering. In the process of clustering, coefficients always adjust the word distribution to change the clustering results. Our experimental results on real data sets show that the proposed semi-supervised text clustering algorithm can get better clustering results than constrained mixmnl, where mixmnl stands for multinomial model-based EM algorithm.

Introduction

Text clustering, as an important method of knowledge discovery, is a procedure and an unsupervised method of automatic text classification. By analyzing the relationship between documents, text clustering makes the same theme articles classified as a class. Without the training process and prior category label, text clustering is provided with higher ability of automatic processing and flexibility, which is widely used in data mining, information retrieval and theme testing. Research on text clustering is demonstrated in [1-3]. Traditional document clustering algorithm is an unsupervised learning method that processes unlabeled documents. In practical applications, however, people can get limited priori knowledge of the data, including class labels and documentation division of constraints conditions (such as pairwise constraints) [4]. Semi-supervised text clustering is a text clustering research branch. It utilizes priori labeled data to guide unsupervised text clustering process on the basis of the traditional text clustering method, and gets better clustering results. Semi-supervised text clustering has recently become a topic of significant interest.

The complexity of document corpora has led to considerable interest in applying hierarchical statistical models based on what are called topic. Topic model could reduce data dimension by changing the document representation from by words to by topic, and achieve new document representation. Among topic models, Latent Dirichlet allocation (LDA) [5] is one of the simplest, most popular models and arguably most important probabilistic models in widespread use today. While cluster documents according to topic, the obtained distribution of topic helps us get clustering results. Therefore LDA can be applied on text clustering.

LDA is a unsupervised learning algorithm. This paper puts forward a new semi-supervised text clustering algorithm, which embed weights of words distribution to LDA. The coefficient guides the clustering process by updating the word item distribution, and then enhances the clustering performance. The semi-supervised LDA text clustering algorithm based on the weights of word distribution (WWDLDA) is experimented on real data sets. The experimental results show that WWDLDA has a better performance than the constrained mixmnl algorithm [6].

Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) presented by Blei is a topic model and a generative probabilistic model of a corpus. A document consisting of a large number of words might be concisely modeled as deriving from a smaller number of topics. A topic is a probability distribution over words. The basic idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

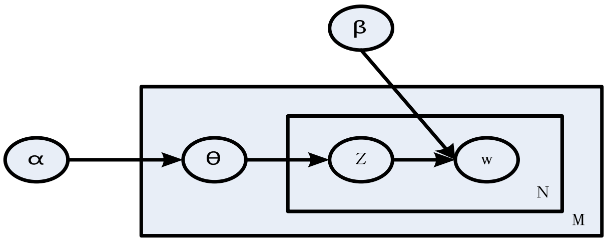


Fig.1. Graphical model representation of LDA.

According to the graphical model representation shown in Figure 1, LDA assumes the following generative process for a document: first, choose a variable θ ,where θ is the random variable parameter of a multinomial over topics and θ follows Dirichlet distribution; secondly choose a topic z_n and then choose a word w_n from a multinomial probability conditioned on the topic z_n ; last repeated choosing topic and word N times. The probability of a corpus is obtained.

$$\mathrm{p}(D\,|\,\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d\,|\,\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{d_n}}p\left(z_{d_n}|\theta_d\right)p\left(w_{dn}\,|\,z_{d_n},\beta\right)\right)d\theta_d \tag{1}$$

Because the posterior distribution of the hidden variables θ is intractable to compute, variational inference could be considered and the free variational parameters γ and φ be added. Due to variational EM, the following pair of update equations is obtained:

$$\gamma_i=\alpha_i+\frac{|V|}{K} \tag{2}$$

$$\phi_{mj}=\beta_{jw_n}\exp(\Psi(\gamma_j)) \tag{3}$$

Compute the Weights of Word Distribution

Given a set of labeled texts D , where d_i represents a text in D . $p_j(w_m)$ represent the word distribution in cluster j and can be estimated by counting the number of documents in each cluster and the number of times w_m occurs in all documents in the cluster j [7].

$$p_j(w_m)=\frac{1+\sum_i p(j\,|\,d_i,D)n_{im}}{|V\,|+\sum_m\sum_i p(j\,|\,d_i,D)n_{im}} \tag{4}$$

Where $|V|$ is the size of the word vocabulary; n_{im} represents the number of times w_m occurring in d_i . Basis on the word distribution over cluster, the weights of word distribution can be computed and normalized as following.

$$\omega_{m,j}=p_j(w_m)/\sum_j p_j(w_m) \tag{5}$$

Semi-supervised LDA Text Clustering Algorithm with Word Distribution Weights

According to the above theory formula, let’s take the part of determined and the other institutions into the condition, and let the rest of the adjustable parameters for quantitative analysis of the design. It can be found to pick the ball with the organization of the contact point has a best value. The point of the height and pick the ball institutions can achieve maximum distance, and the position of the ball fulcrum makes the lift distance and the lift height maximization produce different places, so we can't thought set lift ball point to satisfy the high altitude and the maximize the farthest distance at the same time. Electromagnet for height and the distance of the influence is linear, so we should as far as possible to improve the average output power electromagnets. According to the front of the quantitative analysis, set to pick the ball distance as far as the objective function is as follows:

The semi-supervised text clustering algorithm based on LDA is improved from the perspective of vocabulary. In the new algorithm, the value of parameters should be computed during the iteration. After adding the weights coefficient, according to the variable inference in LDA, the update equation of the multinomial parameter φ is given by

$$\phi_{mj} = \beta_{jw_m} \exp(\Psi(\gamma_j)) * \omega_{m,j} \tag{6}$$

The WWDLDA algorithm's framework is shown in Figure 2.

```
Algorithm: WWDLDA
Input:  A set of N data object $D$ containing $N_l$ labeled data objects $D^l = \{d_1,...,d_{N_l}\}$, and $N_u$ unlabeled data
        objects $D^u = \{d_{N_L+1},...,d_N\}$, the number of clusters $K$.
Output: the data objects given by the cluster identity vector,
        $Y = \{y_1,...,y_N\}$ , $y_n \in \{1,...,K\}$ ,.
Steps:
  1. Initialization: randomly initialize the model parameters $\alpha$ and $\beta$, and set all document in $D$ be unlabeled.
  2. Compute the weight of word distribution over classes in $D^l$ according to (5).
  3. Initialize the parameter $\gamma$ and $\varphi$.
  4. Optimization: optimize (1) by iterating between the following two steps until convergence
       (1)E-step: compute (2), (6);
       (2)M-step: update $\alpha$ and $\beta$ ;
  5. For each data object $d_i$ ,Set $y_{d_i} = \arg\max_j p(\gamma_j, d_i | \alpha, \beta)$.
```

The difference between WWDLDA and LDA is that the update equation of parameter φ changed from (3) to (6). Guiding the clustering process makes the probability of word in unlabeled texts same as in labeled texts, which is the goal of adding the weights coefficient. In other words if the probability of w_m in cluster j is the biggest, the weights guarantee the probability of w_m occurred in unlabeled data in cluster j also is the biggest.

Experiment

A. Datasets

We used the 20-Newsgroups data and several datasets from the CLUTO toolkit. These datasets provide a good representation of different characteristics. A summary of all the datasets used in this paper is shown in Table 1.

TABLE I. SUMMARY OF TEXT DATASETS

Data	Scource	N	|V|	K
News-diff300	20 Newsgroups	300	3991	3
News-sim300	20 Newsgroups	300	2971	3
NG20	20 Newsgroups	2000	13278	20
K1B	WebACE	2340	21839	6
classic	CACM/CISI/CRANFIELD/MEDLINE	7094	41681	4

The 20-Newsgroups dataset contain 20 different newsgroups, 1000 messages form each. The NG20 dataset is constituted by randomly choosing 100 messages form all categories. Additionally we respectively choose 3 classes to make up 2 datasets: News-sim300 and News-diff300. News-sim300 includes 3 similar classes that more overlap within such as comp.graphics, comp.os.ms-windows, and comp.windows.x. The number of documents in datasets is 300. There are 3 different classes in News-diff300 dataset. The boundary of each class is clear. News-diff300 dataset has 300 documents. The classic dataset is obtained by combing the CACM, CISI, CRANFIELD, and MEDLIN abstracts that are used in the past to evaluate various information retrieval systems. The K1B dataset is from the WebACE project.

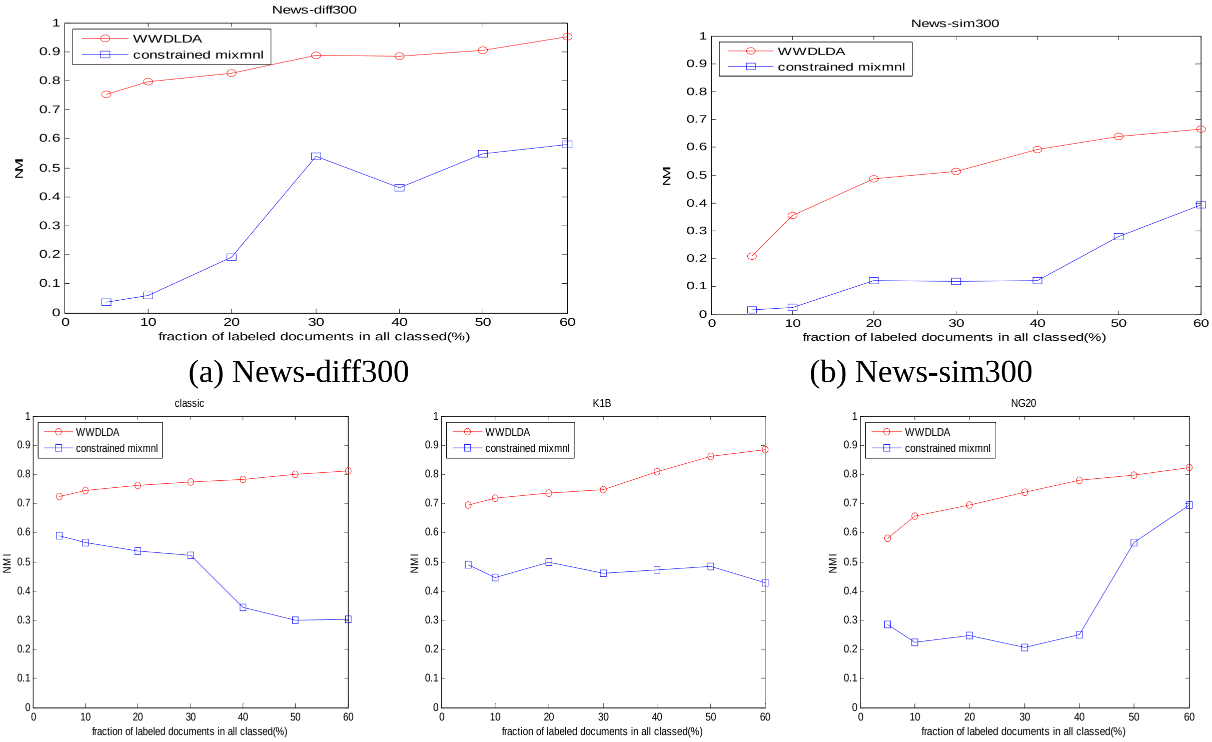
B. *Evaluation Criteria*

Normalized mutual information that refers to NMI [8] can be used as clustering evaluation criteria. NMI is an external measure, mainly used to evaluate the effect of clustering on a data set and the degree of similarity of the real division of the data set. The NMI value is between 0 and 1, the higher the NMI value is, the more perfectly the clustering results match.

C. *Experimental Results and Analysis*

We construct a series of training datasets by randomly sampling 5%, 10%, … , and 60% of all documents as the labeled set and the rest as the unlabeled set. For each algorithm and each percentage setting, we repeat the random sampling process ten times and report the average and standard deviation of NMI values for clustering results.

WWDLDA guides the clustering process by constraining the term distribution, which is similar to the idea of constrained mixmnl algorithm. Therefore, to verify the algorithm, WWDLDA is compared to constrained mixmnl. From comparison results shown in Figure.3 we find that the NMI values of WWDLDA are more stable and increase as the percentage of labeled instances grows. Experiment results show that this method was efficient and feasible, it raised the NMI values by 8% to 20% on five datasets.



(a) News-diff300

(b) News-sim300

(c) Classic

(d) K1B

(e) NG20

Fig.3. Comparing NMI results for WWDLDA and constrained mixmnl algorithms on five datasets in Table 1.

Conclusion

Semi-supervised clustering exploits labeled data to enhance clustering results on unlabeled data. There are many methods and algorithms to exploit labeled data. This paper will choose to analysis the word distribution, applying it in the LDA clustering process. The reason choosing LDA is that documents can be associated with multiple topics under this model. We embed LDA with the weights coefficient to form WWDLDA. After adding the weights coefficient, the word distribution in every iteration is adjusted to decrease the change of the probability of word over each cluster in labeled and unlabeled documents. The adjustment could get good clustering results.

Semi-supervised clustering can be used to discover new classes in unlabeled data in addition to assigning appropriate unlabeled data instances to existing categories. In this paper, our labeled data are sampled from all document classes. Next we will consider the situation that some classes are not available in labeled documents.

Acknowledgement

The authors would like to thank the anonymous reviewers for their comments and kindly suggestions. This paper is supported by the National Natural Science Foundation of China (NSFC, NO. 60863005; No.61262006), the Science and Technology Foundation of Guizhou Province (NO.20122125), the Scientific Research Project of Introduce Talents of Guizhou University (NO.201114), the Foundation of Informatization of Manufacturing Industry of Guizhou Province (NO. GY (2011)3074) and Graduate Innovated Foundation of Guizhou University.

References

[1] XH. Hu, XD. Zhang, CM. Lu, EK. Park, and XH. Zhou, “Exploiting wikipedia as external knowledge for document clustering, ”Proc. the 15th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining, ACM Press, Jun, 2009. 389-396.

[2] HT. Zheng, BY. Kang, and HG. Kim. Exploiting noun phrases and semantic relationships for text document clustering [J]. Information Sciences, 2009 179(13) 2249-2262,.

[3] M. Mahdavi and H. Abolhassani. Harmony *K*-means algorithm for document clustering [J]. Data Mining and Knowledge Discovery, 2009 18(3) 370-391.

[4] WZ. Zhao, HF. Ma, ZQ. Li, and ZZ. Shi. Efficiently active learning for semi-supervised document clustering [J]. Journal of Software, 2012 23(6) 1486−1499,.

[5] D. Blei and M. Jordan. Modeling annotated data. Proc. SIGIR, ACM Press, Jul, 2003. 127–134.

[6] Shi Zhong. Semi-supervised Model-based Document Clustering: A Comparative Study [J]. Machine Learning, 2006 (65) 3-29.

[7] S. Zhong and J. Ghosh, “A comparative study of generative models for document clustering,”Proc. SDM Workshop on Clustering High Dimensional Data and Its Applications, SDM Press, May. 2003.

[8] A. Strehl, J. Ghosh, and R. Mooney, “Impact of similarity measures on Web-page clustering,” Proc. the Workshop on Artificial Intelligence for Web Search, AAAI Press, Jul. 2000. 58-64.