Public Opinion Clustering for Hot Event Based on BR-LDA Model

Ningning Ni, Caili Guo, Zhimin Zeng

Public Opinion Clustering for Hot Event Based on BR-LDA Model

Ningning Ni*, Caili Guo, and Zhimin Zeng

Beijing University of Posts and Telecommunications,
NO.10 Xitucheng Road, Haidian District, Beijing 100876,China
{niningning,guocaili,zengzm}@bupt.edu.cn
http://www.springer.com/lncs

Abstract. With the rapid development of web2.0, there is more and more content on social media, and information is widely spread in people's lives through social media. Public often make vast opinions on hot Events on social media platforms, such as Sina Weibo and Twitter. Clustering these opinions can increase understanding of the semantics of public opinions. Mining these opinions thoroughly can help companies and management make better decisions. The challenge of opinion clustering for hot events is that most of opinions contain background information of event. The background information could reduce opinion clustering performance. In this paper, we propose a topic model named background removal LDA(BR-LDA) model for opinion clustering. The model adds the idea of removing background to the LDA model so it can separate opinion words from background words. First, we remove some words with high frequency in the corpus. Then the model applies BR-LDA model to automatically cluster public opinions. Experimental results on two real-world datasets of two languages, Chinese and English, verify the efficiency of the proposed model.

Keywords: public opinion, clustering, hot events, social media, topic model

1 Introduction

With the rapid development of web2.0, there is more and more content on social media (such as Twitter, Sina Weibo, etc.), and information is widely spread in people's lives through social media. On social media, people create and spread a lot of interesting content, interact with others, and gain more knowledge. People discuss hot events on social media, publish and exchange their opinions [1]. Mining these data thoroughly can help companies understand the needs of users and make better user-oriented products. The management can track peoples

* Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

reactions to policies and provide more informed advice for implementing future policies.

Opinion clustering can be seen as a kind of text clustering. There are many studies on text clustering. Traditional text analysis methods such as latent Dirichlet distribution (LDA) have also been widely used and have achieved good results. However, traditional text clustering methods mainly focus on event-based clustering, and the clustering granularity is relatively large so there are big differences between clusters. And Since almost all opinions related to the same hot event have a similar background, the background information will reduce opinion clustering performance if they are not removed. Using the traditional methods do not completely subdivide the information between background and opinion. Some documents with different opinions may have common background and make their differences submerge. There are relatively few researches on opinion clustering, but this is the challenge we must face.

This paper presents a background removal LDA (BR-LDA) model. The model adds the idea of removing background to the LDA model. Experimental results on two real-world datasets of two languages verify the better performance of the proposed model. The contributions of this article mainly include the following points:

–This paper proposes an opinion clustering model based on the BR-LDA model to solve the problem of opinion clustering on the texts with the same event background. The BR-LDA model can remove background for better opinion clustering.

–The experiments in this paper were conducted on datasets of two languages, Chinese and English, and proved that our model does not have language dependence.

The paper is organized as follows. We begin with a discussion of related work in the areas of opinion clustering and in Section 2. Then, the proposed model is described in Section 3. The experiments for the evaluation of the proposed model is reported in Section 4. Finally, we conclude this paper in Section 5.

2 Related Work

In text clustering, LDA [2] is widely used and has achieved good performance. LDA expresses documents and words as probability distributions on the subject, and obtains the relationship between documents and topics, and words and topics. Zhao et al. proposed TwitterLDA [3]was considered to be the first topic model designed specifically for tweet data. Unlike traditional official documents, tweets are short and noisy. TwitterLDA made two major contributions to the tweet data. First, because tweets are relatively short in length, they believe that each tweet maps to only one topic, rather than the document as a distribution of topics. Second, they divide words into background words and topic words. Background words are frequently used words in all tweets, and topic words are meaningful words related to topics. Llewellyn et al. is focused on the clustering of news reviews [4]. Like many social media data sets, comment data contains very

short documents. The number of words in the document is a limiting factor in the performance of LDA clustering. They propose that they can combine annotations to form larger documents to improve clustering quality. Llewellyn et al. used LDA and k-means, as well as some simple metrics such as cosine distances, clustered the comments of the most single news article, and demonstrated that LDA performs best [5]. They also use LDA to cluster news commentary and use the resulting class information to generate comment summaries [6].

Graph-based methods are also clustered in user texts, such as Aker et al. based on the similarity features and the weights trained using automatically derived training data, proposes a linear regression model for the similarity between graph nodes (comments). To mark the cluster, the author's graph-based approach uses DBPedia to abstract topics extracted from the cluster [7]. Chen et al. built topics using a topic graph, where the topics were represented as concept nodes and their semantic relationships using WordNet. Then, the author extracted each topic from the topic graph to obtain a corpus by community discovery. In order to find the optimal topic to describe the related corpus, they defined a topic pruning process using Markov decision processes [8].

These methods are not suitable for opinion clustering for hot event because almost all opinions related to the same hot event have a similar background. These methods dont have the ability to remove the background words from opinion words. Our BR-LDA model can effectively remove backgrounds and achieve better opinion clustering results.

3 Model

In this section we elaborate on our opinion clustering model. When data (both Chinese and English are suitable) are fed into our model, first, our preprocessing module is used to preprocess the data, such as removing punctuation, stop words, links, etc. Then remove the high frequency (HF) words. These HF words are usually background information related to the event and have nothing to do with the opinion. Then, we input the preprocessed data into our BR-LDA model to cluster opinions. The BR-LDA model can further separate opinion words from background words.

3.1 Data Preprocessing

Before the opinion clustering, the data must be preprocessed, because the original data usually includes many useless information, such as punctuation, stop words, links and so on. The task of this research is to cluster opinions of hot events, and the opinions about the same hot event usually include some event related background words. According to our observations, high frequency words are usually background words. At this stage, the top K high-frequency words are removed and so some background words are filtered out. The frequency of a word (term frequency, tf) is calculated as follow:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

Where $n_{i,j}$ is number of occurrences of the word in the document d_j, and the denominator is the sum of the occurrences of all words in the file d_j.

3.2 Opinion Clustering

Notations and Definitions. Table 1 summarizes the notations used in this paper for our proposed BR-LDA model and the corresponding descriptions. Opinion: Every tweet or microblog is viewed as an opinion. Each document fed into the model is an opinion. Word type: The words from corpus are divided into two types: background words and opinion words. For example, in the event of Saudi Arabia grants citizenship to a robot, Saudi was a background word, and scared is an opinion word. In our model, general word in the model is the background word, specific word is the opinion word. Opinion cluster: A collection of opinions that express similar views. Each topic in our BR-LDA model is an opinion cluster.

Table 1. Notation

Notation	Description
D	total number of documents
T	total number of topics
N_d	total number of words in *d-th* document
W	total number of words
z, w	label for topic,word
x	indicator of general or specific for word
φ^G	general word distribution
φ^S	specific word distribution
π	document-specific Bernoulli distribution
θ	topic distribution
$\alpha, \gamma, \beta^S, \beta^G$	Dirichlet priors

BR-LDA Model. The graphical representation of BR-LDA model is illustrated in Figure 1. Formally, we assume that there are a total of Z topics in the corpus. The original LDA assumes that each document has a topic probability distribution, but because the length of these documents is short, we assume that each document belongs to only one topic. TwitterLDA assumes that each user has a specific topic distribution, but we don't think users have a specific topic distribution because users have very different opinions about different events. TwitterLDA assumes that the indicators of all documents come from the same Bernoulli distribution, but different opinions have different degree of expression of the background, so in our model, each document has its own unique Bernoulli distribution. Our model is more suitable for the clustering of opinions on hot events.

The document generation process is as follows, where $Dir()$ and $Multi()$ represent Dirichlet and Multinomial distributions respectively.

1. Draw $\varphi^G \sim Dir(\beta^G)$ indicating the general word distribution. Draw $\theta \sim Dir(\alpha)$ indicating the topic distribution.

2. For each topic $z = 1;;T$, draw $\varphi^S \sim Dir(\beta^S)$, denoting the specific word distribution for topic z.

3. For the *d-th* document:

a. Draw $z \sim Multi(\theta)$, corresponding to the topic assigned for each document. Draw $\pi \sim Dir(\gamma)$ the Bernoulli distributions that determine the selections between the general words and specific words.

b. For the *n-th* word in the document, $n = 1;;N_d$:

i. Draw a variable $x \sim Bernoulli(\pi)$ as an indicator for general or specific word;

ii. Draw $w \sim Multi(\varphi^G)$ if $x = 0$, and $w \sim Multi(\varphi^S|z)$ if $x = 1$.

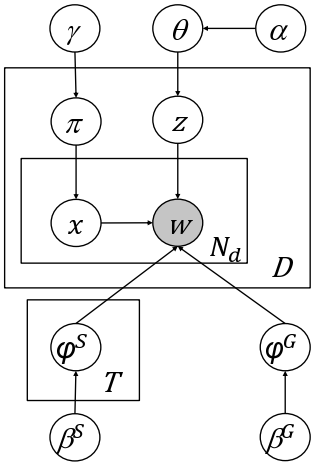

Fig. 1. Graphical model of BR-LDA

Inference. In our BR-LDA model, the topic assignment Z as well as general-specific indicator X are latent variables to be inferred from the observations. We use Gibbs sampling to achieve the inference due to its efficiency and effectiveness.

the probability of assigning a topic z to t for *d-th* document can be estimated as follows:

$$p(z_d = t|Z_{\neg d}, W, X) \propto \frac{n^t_{\neg d} + \alpha}{\sum_{t=1}^{T} n^t_{\neg d} + T\alpha} \times \frac{\prod_{w=1}^{W} \prod_{p=1}^{n_i} (n^w_{\neg d,t,x=1} + \beta^S)}{\prod_{q=1}^{N_i} (\sum_{w=1}^{W} n^w_{\neg d,t,x=1} + W\beta^S)} \quad (2)$$

Where t represents the topic of the current document and $n^{t}_{\neg d}$ is the number of times topic t occurs in the documents not including to the current document. w denotes the current sample word, and $n^{w}_{\neg d,t,x=1}$ denotes the number of specific word w is sampled as topic t not including to the current document. n_i denotes the number of times word w occurs in the current document. N_i denotes the number of words sampled with $x = 1$ occurs in the current document.

Next, we sample the general-specific indicators x. Let i be $\{d,n\}$. The probability of assigning a binary label 1 to x_i as a specific word indicator is estimated as below:

$$p(x_i = 1|X_{\neg i}, W, Z) \propto \frac{n^{x=1}_{\neg i,d} + \gamma}{\sum^{1}_{x=0} n^{x}_{\neg i,d} + 2\gamma} \times \frac{n^{w}_{\neg i,t,x=1} + \beta^{S}}{\sum^{W}_{w=1} n^{w}_{\neg i,t,x=1} + W\beta^{S}} \tag{3}$$

The probability of assigning a binary label 0 to x_i as a general word indicator is estimated as follow:

$$p(x_i = 0|X_{\neg i}, W, Z) \propto \frac{n^{x=0}_{\neg i,d} + \gamma}{\sum^{1}_{x=0} n^{x}_{\neg i,d} + 2\gamma} \times \frac{n^{w}_{\neg i,x=0} + \beta^{G}}{\sum^{W}_{w=1} n^{w}_{\neg i,x=0} + W\beta^{G}} \tag{4}$$

Where $n^{x=1}_{\neg i,d}$ ($n^{x=0}_{\neg i,d}$) denotes the number of times $x = 1(x = 0)$ occurs for the corpus not including the *n-th* word in the *d-th* document. $n^{w}_{\neg i,t,x=1}$ denotes the number of times the specific word w is sampled as topic t for the corpus not including the *n-th* word in the *d-th* document. $n^{w}_{\neg i,x=0}$ denotes the number of times $x = 0$ occurs and the number of times word w is sampled with label 0 for the corpus not including the *n-th* word in the *d-th* document.

After the algorithm converges, general word distribution and specific word-topic distribution can be estimated according to the following two formulas:

$$\varphi^{G}_{w} = \frac{n^{w}_{x=0} + \beta^{G}}{\sum^{W}_{w=1} n^{w}_{x=0} + W\beta^{G}} \tag{5}$$

$$\varphi^{S}_{wt} = \frac{n^{w}_{t,x=1} + \beta^{S}}{\sum^{W}_{w=1} n^{w}_{t,x=1} + W\beta^{S}} \tag{6}$$

4 Experiments

In this section, we conduct a systematic analysis to evaluate our proposed opinion clustering model. First introduce the datasets used for the experiment. Then, introduce the evaluation metrics. Finally, we compare the performance of our model with the other three models.

4.1 Dataset Description

The dataset used in this experiment contains both English and Chinese. The English data comes from Twitter, and the Chinese data comes from Sina Weibo. The data comes from hot events, "Saudi Arabia grants citizenship to a robot". Finally, we got 4,430 tweets and 4,310 microblogs.

4.2 Evaluation Metrics

In our experiments, we used the *precision*, *recall* and *F1-score* which are commonly used evaluation metrics in the clustering to evaluate the performance of the proposed model. Let *TP*, *FP*, *TN* and *FN* refer to the number of predictions falling into True Positive, False Positive, True Negative and False Negative categories.

$$precision = TP/(TP + FP) \tag{7}$$

$$recall = TP/(TP + FN) \tag{8}$$

$$F1\text{-}score = precision * recall/2(precision + recall) \tag{9}$$

F1-score can be viewed as a comprehensive indicator of *precison* and *recall*.

4.3 Experimental Results and Discussion

In order to prove that our BR-LDA model can separate background words from opinion words, we show the background words and opinion words respectively obtained from the BR-LDA model. The result shown in Table 2.

Table 2. Background words and opinion words

Dataset	Background Words	Opinion Words
Microblog	索菲亚(Sophia)	可怕(scary)
	沙特阿拉伯(Saudi Arabia)	细思极恐 (fridge horror)
	机器人(robot)	背后发凉(creepy)
	公民身份(citizenship)	恐怖(horrific)
	第一个(first)	美女(beauty)
Twitter	Sophia	outrage
	Saudi Arabia	cool
	Robot	intelligent
	citizenship	rights
	bestows	scary

In this section, we compared our opinion clustering model BR-LDA to three other models in order to verify its effectiveness. The three models are: LDA, k-means, and TwitterLDA [3]. In the experiment, the parameters $\alpha = 1$, $\beta^S = 0.25$, $\beta^G = 0.01$, $\gamma = 0.05$ are set empirically. The comparison of proposed BR-LDA model with the other models are shown in Table 3 and Table 4.

As seen from the results, the proposed BR-LDA model outperforms other models on both datasets in all metrics. The good performance of BR-LDA benefits from that our model separates opinion words from background words, and each document has its own unique Bernoulli distribution. Our model performs better than other models.

Table 3. Performance on Microblog dataset

Models	Precision	Recall	F1-score
K-means	0.76	0.645	0.698
LDA	0.814	0.659	0.728
TwitterLDA	0.773	0.703	0.736
BR-LDA	**0.843**	**0.71**	**0.77**

Table 4. Performance on Twitter dataset

Models	Precision	Recall	F1-score
K-means	0.76	0.663	0.708
LDA	0.748	0.682	0.713
TwitterLDA	0.772	0.684	0.726
BR-LDA	**0.782**	**0.72**	**0.75**

5 Conclusion

In this article, we propose a model called BR-LDA model for opinion clustering for hot events on social media. Since most of opinions contain background information of event. The background information could reduce opinion clustering performance. The BR-LDA model can effectively separate background words from opinion words. A large number of experiments on two datasets in Chinese and English in real life have demonstrated the effectiveness of our model and proved that our model does not have language dependence. Our model is used for offline opinion clustering. For future work, we plan to improve our model for real-time opinion clustering task so we can obtain dynamic clusters and realize the trend of opinions in real time.

References

1. Li, Q., Jin, Z., Wang, C., Zeng, D.D.: Mining opinion summarizations using convolutional neural networks in chinese microblogging systems. Knowledge-Based Systems 107(C), 289–300 (2016)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J Machine Learning Research Archive 3, 993–1022 (2003)
3. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: European Conference on Advances in Information Retrieval. pp. 338–349 (2011)
4. Llewellyn, C., Grover, C., Oberlander, J.: Improving topic model clustering of newspaper comments for summarisation. In: ACL 2016 Student Research Workshop. pp. 43–50 (2016)
5. Llewellyn, C., Grover, C., Oberlander, J.: Summarizing newspaper comments (2014)
6. Ma, Z., Sun, A., Yuan, Q., Cong, G.: Topic-driven reader comments summarization. In: ACM International Conference on Information and Knowledge Management. pp. 265–274 (2012)

7. Aker, A., Kurtic, E., Balamurali, A.R., Paramita, M., Barker, E., Hepple, M., Gaizauskas, R.: A graph-based approach to topic clustering for online comments to news. In: European Conference on Information Retrieval. pp. 15–29 (2016)
8. Chen, Q., Guo, X., Bai, H.: Semantic-based topic detection using Markov decision processes. Elsevier Science Publishers B. V. (2017)