Detection of Food Safety Topics Based on SPLDAs

Jinshuo Liu[1(✉)], Yabo Li[1], Yingyue Peng[2], Juan Deng[2], and Xin Chen[1]

[1] Computer School, Wuhan University, Wuhan 430072, China
liujinshuo@whu.edu.cn, {921834021, 459617701}@qq.com
[2] International School of Software, Wuhan University, Wuhan 430072, China
706357455@qq.com, dengjuan@whu.edu.cn

Abstract. Nowadays, the problems of food safety are more and more serious. This paper focuses on network topic detection of food safety problems, which is difficult because of several reasons, such as various description of a same problem and sparseness of the data. In this paper, a novel method based on Single-pass in LDA space is proposed to detect the food safety problems from various sources, such as microblog and news reports. The experiments show that the method could detect food safety topics efficiently. The F-measure value of clustering almost increases from 56.03 % to 87.21 %, compared with Single-Pass based on traditional VSM. In addition, experiments about the influence of similarity parameter to models' performance demonstrate that our method has a better robustness.

Keywords: Food safety · Topic detection · LDA space · Single-Pass

1 Introduction

Unfortunately, food safety incidents occur frequently, such as the inferior milk powder, the Sudan red events, etc. To effectively detect such topics about food safety from the vast number of Internet data is difficult for several reasons. One reason is that sometimes, people discuss the same problem with different descriptions. For example, "Melamine incidents" and "*Sanlu* milk powder incidents" are in fact the same topic, but the two descriptions have big difference on the vocabulary level. Another reason is sparseness of the data. The data from Microblog and BBS is relatively short while the length of traditional news is long. As the amount of data rises, the sparseness of short texts representation will become more and more serious.

This paper explores an approach aimed to detect food safety topics effectively. In Sect. 2 we briefly introduces the Latent Dirichlet Allocation (LDA) and Single-Pass. Section 3 presents the modeling of topic detection of food safety problems in this paper. Section 4 describes experiments. Finally, Sect. 5 concludes the paper.

J. Tian et al. (Eds.): SecureComm 2014, Part I, LNICST 152, pp. 551–555, 2015.
DOI: 10.1007/978-3-319-23829-6_42

2 Related Work

Topic detection is a task of the Topic Detection and Tracking (TDT), which defined to automatically detect new topics in the news stream and associating incoming stories with topics created so far [1]. There are two main representation models: Vector Space Model (VSM) based on feature terms and semantic topic model [2].

In VSM, the representative algorithm Single-Pass [3] is a widely used method for topic detection, which is based on the VSM space. In Single-Pass, each document is mapped as a feature term vector and will be calculated the similarity between the existing topics. Single-Pass is easy to understand but the loss of semantic is serious. LDA [4] is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. LDA steers a new direction about semantic topic modeling in natural language processing. But it does not work well for topic detection.

3 Our Modal: SPLDAs

The traditional Single-Pass in the VSM is based on the feature terms. However, the feature selection and weighting is difficult and there is no authoritative method. Besides, if the number of food safety data is larger, the dimension of document vector will be higher. What's worse, VSM model may lead to the loss of semantic. In order to detect topics of food safety effectively, the semantic information should be analyzed. Meanwhile the difficulties in VSM ought to be solved or avoided. A method: Single-Pass in LDA space (noted as SPLDAs) is proposed in this paper. SPLDAs is based on the LDA space. As a result the feature term selection is avoided and the semantic information is warranted. Besides with the improving of the data size, the dimension of the vector space is fixed according to the number of latent topics.

SPLDAs can divided into two phrases, the mapping of food safety data sets to LDA space and the processing of Single-pass in LDA space, which is given as Fig. 1.

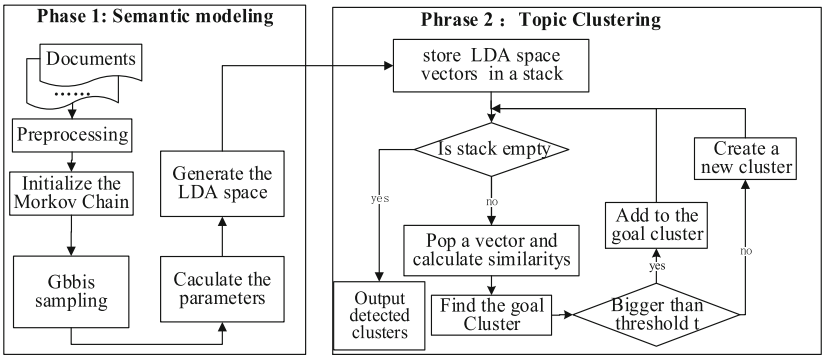

Fig. 1. Single-pass in LDA space

3.1 Mapping of Data Sets to the LDA Space

The paper uses Gibbs sampling [5] to estimate the parameters in LDA. Firstly our model defines LDA space as below,

$$\theta = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{21} & \cdots & p_{2K} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ p_{M1} & p_{M2} & \cdots & p_{MK} \end{bmatrix}$$

θ is a M∗K matrix, where M is the total number of documents, while k is the number of latent food safety topics. Element p_{ij} of the matrix indicates the probability of the ith document in data set to generate the j^{th} topic.

LDA Space is a new vector space, where the i^{th} document could be viewed as a vector $(p_{i1}, p_{i2}, \ldots\ldots, p_{iK})$, which meets the condition $\sum_{j=1}^{K} p_{ij} = 1$. The document set is mapped into the LDA Space after Gibbs sampling.

3.2 Single-Pass Processing in LDA Space

Single-Pass is a widely used method for topic detection, which is based on the VSM space. Now the idea is used in LDA space. The process of Single-Pass in LDA space is described as below:

input: a stack $\{d_1, d_2, \ldots, d_M\}$ in LDA space **Output:** a set of clusters $\{c_1, c_2, \ldots, c_s\}$	
1)	Pop a vector d_i in the document stack
2)	Compute the similarity between d_i and each existing clusters and find the closest one, noted as C_{max}

3)	If $sim(C_{max}, d_i) > t$ then
	Include d_i in c
	Else
	Create a new cluster and add d_i to it.

4)	If the stack is empty, then
	Terminal the algorithm.
	Else
	Repeat step 1

4 Experiments

There is no public corpus of food safety problems. In order to evaluate the proposed method SPLDAs, 15,850 documents of food safety problem including 143 topics, are manually collected by Web crawler from "Xinhua" website (http://yuqing.news.cn/spaq.htm), Tencent Microblog (http://t.qq.com/zhangwuji9/), etc.

Based on the F-measure, the paper adopts size-weighted F-measure to evaluate our proposed method, which combines the precision P and recall R.

To verify the validity of the SPLDAs, we have done two groups of experiments. One is Single-pass in LDA space and the other is comparative experiments, Single-Pass on VSM (noted as SP). There are 143 topics included in our corpus. Thus, the number of detected clusters is expected to be close to 143. After fixing different similarity threshold t, we only consider the range from 100 to 200 in the experimental analysis. Table 1 shows the average results of experiments on the different similarity t.

Table 1. The average results of experiments on the similarity threshold level.

Method	t (frow, to)	P	R	F-measure	Number of clusters
SP	(0.0058,0.0118)	0.6039	0.5256	0.5603	146
SPLDAs	(0.15, 0.54)	0.8662	0.8789	0.8721	149

Figure 2 demonstrates that the number of detected clusters relates with the similarity threshold t. The slope of SP is much greater than of SPLADs, which means that a little change of t would have great influences on the detection result of SP. Thus, SPLDAs has a better robustness than SP.

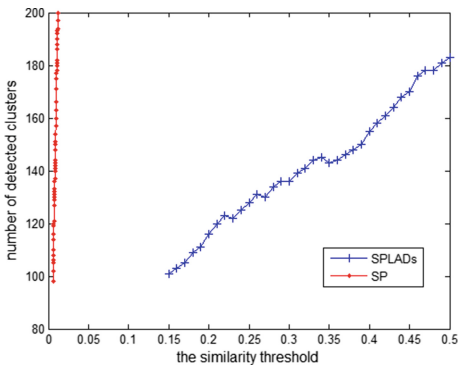

Fig. 2. Comparison of the number of detected clusters.

Figure 3 below shows that F-measure, R and P of SPLDAs are much higher than SP with different number of detected clusters. We can conclude from Table 2 that SPLDAs increase the average F-measure value by about 31 %, compared to SP. And it is proved that SPLADs is more efficient on topic detection of food safety problems.

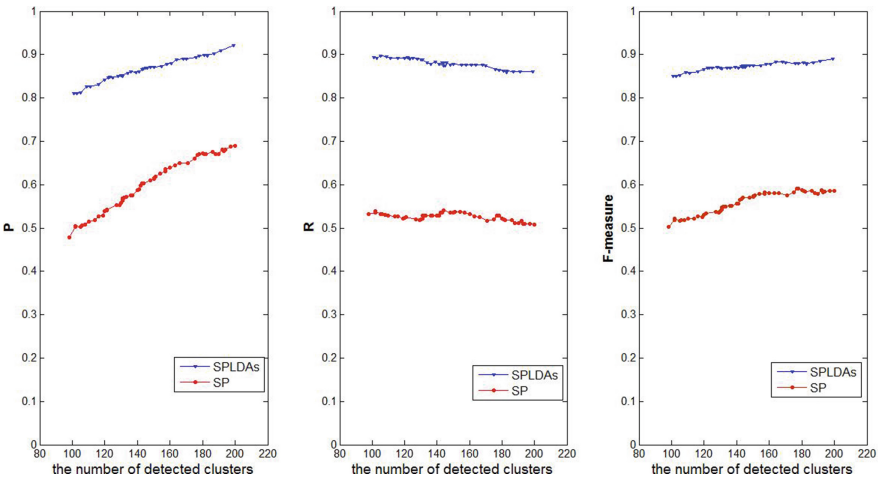

Fig. 3. Comparison of SPLDAs and SP from the P, R and F-measure value.

5 Conclusion

Considering the characteristics of rich semantic information and high sparseness in the food safety data from microblog or news reports, this paper proposes the Single-Pass Clustering algorithm in LDA space. The method has the advantage of solving the data sparseness problem and loss of semantic information, compared with the traditional VSM. According to the experimental results, the combined method could increase the precision and recall, and finally improve the clustering quality. Future research may consider and to do the real-time process of large food safety data and the representation of food safety problems.

References

1. Kamaldeep, K., Gupta, V.: A survey of topic tracking techniques. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5**(2), 383–392 (2012)
2. Lin, C., He, Y., et al.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 375–384. ACM (2009)
3. Papka, R., Allan. J.: On-line new event detection using single pass clustering. University of Massachusetts, Amherst (1998)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)