Expert Systems With Applications 115 (2019) 346–355

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Natural disaster topic extraction in Sina microblogging based on graph analysis

Tinghuai Ma[a,b,*], YuWei Zhao[a], Honghao Zhou[a], Yuan Tian[c], Abdullah Al-Dhelaan[c], Mznah Al-Rodhaan[c]

[a] School of Computer & Software, Nanjing University of information science & Technology, Jiangsu, Nanjing 210-044, China
[b] CICAEET, Jiangsu Engineering Center of Network Monitoring, Nanjing University of information science & Technology, Nanjing 210-044, China
[c] Computer Science Department, College of Computer and Information Science, King Saud University, Riyadh 11362, Saudi Arabia

ARTICLE INFO

Article history:
Received 4 April 2018
Revised 25 June 2018
Accepted 8 August 2018
Available online 9 August 2018

Keywords:
Topic detection
Community detection
Natural disaster
Sina microblogging
Graph analysis

ABSTRACT

In this paper, we will propose a novel approach based on graph analysis which will use community structure detection algorithm to detect topics in the keywords graph of micro-blogging data. Furthermore, considering the specificity of the Sina microblogging, we propose novel keywords filtering model and graph generation algorithm to meet the dual requirements of topic detection and community detection. We validate our approach on a big natural disaster dataset from Sina micro-blog, in which about 10^3 micro-blogging posts with about 10^4 distinct feature tags. The experimental results definitely revealed the relationship between the keywords and the natural disaster topics. Our methodology is a scalable method which can adapt to the changes in the amount of data. Especially, we can get abundant information about natural disasters in the topic detection and help the government guide the rescue of disasters.

1. Introduction

In recent years, the natural disasters happen frequently in the world which caused the majority of researchers' attention, such as debris flow, floods, earthquakes, and typhoons. How to find these disasters and lock the disasters' areas at the first time has become the focus of everyone's attention. With the development of Internet and the growing popularity of various communication devices, people can no longer obtain information and exchange information only confined to the traditional media. Social networking has become the open social media services based on the new network platform (Ma et al., 2018a; Ma, Shao, Hao, & Cao, 2018b). Due to the particularly rapid development of microblogging platform, it has not only become the important means of users to explore the news events, express their views and insights, but also become the important places to disseminate hot topics (Ma et al., 2016a). Taking Sina microblogging (China) as an example, the active users has reached 100 million and the daily number of microblogging posts up to more than 300 million as of June 2017. The topic detection of microblogging will help the community and the government to find those natural disasters and emergencies which was difficult to predict in time. Above all, it will help the government to keep abreast of the network public opinion and guide the public opinion correctly.

The previous hot topic detection (Benny & Philip, 2015; Huifang, Yugang, Xiaohong, & Zhou, 2016; Yu, Zhao, Chang, & He, 2014; Zhou, Zhong, & Li, 2014) is focused on how to improve the accuracy of the algorithm, enhance the topic of expression and improve hot topics sorting rules. It usually only consider the topic detection of the text content, but ignored the topic expression. In this paper, we propose a graph based method of using community detection algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) to detect topics which can find valuable topics in the massive and messy data in the form of a keywords graph. Further, we can find extra valuable information of the disaster, such as the location, the date and so on.

One of the key challenges in microblogging data is that due to the Chinese microblogging content is short, dispersed, sparse, noisy and complicated (Huang, Yang, Mahmood, & Wang, 2012; Lo, Chiong, & Cornforth, 2017; Yang, Lin, Lin, & Liu, 2016a; Yeh, Tan, & Lee, 2016), the correlation of the different texts is weak, so the latent topic is very difficult to detect or extract (Lv et al., 2016). For the general method, the dispersion and sparseness of the text are reduced by recalculation of decentralized data or considering other factors when building topic models (Chang, Hsieh, Chen, & Hsu, 2015). The researchers utilize the auxiliary information of the

[*] Corresponding author.
E-mail address: thma@nuist.edu.cn (T. Ma).

https://doi.org/10.1016/j.eswa.2018.08.010
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

text, such as considering the reprinting relationships to enhance the relevance of the text (Ma et al., 2016b). In this paper, a new keywords filtering model is proposed with the Natural Language Processing (NLP) word segmentation technique to remove the useless or redundant noise feature words. After the initial filtering, we continuously consider to select the feature words rigorously from three aspects which will make them to be the most representative keywords under both local and global conditions. We only pick those feature words with high weight to obtain a dense set of keywords. And then we use the graph generation model to transform the abstract content into the visible words graph. We utilize the relationship between the two keywords and calculate the association of each pair of keywords to obtain the kernel edges' value. To rule out the influence of the order of keywords, we use the two-by-two products to reduce the error. Finally, the community detection algorithm is proposed to divide closely related keywords into the same community which can extract the topics and show the content more intuitively. Specially, we utilize Gephi to obtain the graphical representation and we can know the distribution of each topic at a glance. Moreover, we focus on the detection and extraction of natural disaster topics. Our purpose is to find the emergency natural disasters in the big data, thereby to analysis the disasters and solve the disasters. As a result, we designed an algorithm that meet the requirement of topic extraction from the natural disaster data using the keyword extraction method mentioned above. We evaluate our proposed topic detection model on real Sina-microblogging dataset by comparing the topic detection performance. We also compared our method with other four topic detection methods in the experimental section.

The contribution of this work is threefold:

(1) We propose a novel keywords filtering model which gets representative and highly dense feature words from the complicated microblogging data. We not only considers some properties of the feature word itself, but also considers its importance in the local text as well as the global one.
(2) We utilize a graph generation algorithm transforming the abstract text data into a graph of the keywords. We also accurately calculate the value of association between two keywords to form a keywords map structure which can visualize the topics obviously.
(3) We utilize the heuristic community detection algorithm to group the disordered keyword graph in an orderly manner and proposed a topic extraction rule to describe the contents of each topics in detail which extract topic accurately and discover the relationship between topics.

The remainder of the paper is organized as follows: Section 2 reviews the related work; Section 3 describes the proposed topic detection method; Section 4 describes our experimental setup and the experimental results. We conclude the paper in Section 5.

2. Related works

The classic model of topic detection is proposed by Blei, Ng, and Jordan (2003) called the latent Dirichlet allocation (LDA) method, which is a Bayesian networks-based topic model widely used to identify topics from 2003. This method overcomes the shortcoming that the parameters are increased with the number of document set is increasing. Many researchers improve the LDA model according to different scenarios. Ye, Du, and Fu (2016) proposed a probabilistic generative model named Microblog Features Latent Dirichlet Allocation (MF-LDA) to extract microblog topics. They incorporate five microblog's unique features into the analysis of LDA model to improve the performance of the traditional one. Amoualian, Clausel, Gaussier, and Amini (2016) proposed two models for modeling topic and word-topic dependencies between consecutive documents in document streams. The first extension makes use of a Dirichlet distribution to balance the influence of the LDA prior (α and β) to topic and word-topic distribution of the previous document. The second extension makes use of copulas, which constitute a generic tool to model dependencies between random variables. Chen, Li, Guo, and Guo (2016) proposed a FSC-LDA model which combining the text clustering methods and feature selection methods. It can identify the number of topics adaptively, keep short micro-blog texts features better and make the result more stable. Most of the above research is based on offline data, but there are also approaches on the online data flow for topic detection and prediction. Dang, Gao, and Zhou (2016) proposed a new early detection method for emerging topics based on Dynamic Bayesian Networks in micro-blogging networks. They build a DBN-based model by the conditional dependencies between features to identify the emerging keywords and cluster the emerging keywords into emerging topics by the co-occurrence relations between keywords. Xie, Zhu, Jiang, Lim, and Wang (2016) proposed a sketch-based topic model with dimension reduction technique to achieve bursty topics from Twitter. They developed a "sketch of topic", which provides a "snapshot" of the current tweet stream and can be updated efficiently.

Although the probabilistic topic model has a wide range usage and dimensionality reduction, but it requires people to set the number of topics in advance and can't detect those new topics that haven't appear in the training data. Therefore, the unsupervised clustering method is considered to achieve automatic data analysis and topic detection which has become another focus of the researchers. Liang, Yilmaz, and Kanoulas (2016) proposed a dynamic clustering topic model method DCT for short-length streaming text. It can effectively model both the temporal nature of topics in streaming text and the sparsity problem of short text. Fitriyani and Murfi (2016) proposed a mini batch K-means method to detect topics in big datasets effectively. It is used to reduce the computational time and improve the accuracy of the algorithm. Further, although some approaches didn't use the classic clustering algorithm directly, it is developed based on the idea of clustering. Pang et al. (2015) proposed a clustering-like pattern across similarity cascades (SCs) which can truncate a similarity graph with a set of thresholds in a series of subgraphs to capture topics with maximal cliques. Then a topic-restricted similarity diffusion process is proposed to efficiently identify real topics from a large number of candidates.

In addition to the probabilistic topic model and the unsupervised clustering approaches above, more and more researchers use the graph analysis method to detect the words' association in the graph or networks. The method of graph analysis can add the topic extraction visualization and is suitable to utilize the community detection algorithm which seems more interesting and flexible (Rong, Ma, Tang, & Cao, 2018; Zhang, Ma, Cao, & Tang, 2016b). Cigarrn, ngel Castellanos, and Garca-Serrano (2016) proposed an approach based on Formal Concept Analysis (FCA), a fully unsupervised methodology to group similar content together in thematically-based topics and to organize them in the form of a concept lattice. Zhang, Wang, Cao, Wang, and Xu (2016a) proposed a hybrid relations analysis approach to integrate semantic relations and co-occurrence relations for topic detection. The approach fuses multiple relations into a term graph and detects topics from the graph using a graph analytical method. With the analysis of community detection methods, Sayyadi and Raschid (2013) proposed a graph analytical approach for topic detection. They used a KeyGraph algorithm to convert text data into a term graph based on co-occurrence relations between terms. Then they employed a community detection approach to partition the graph. Eventually, each community is regarded as a topic and terms within