Physica A 416 (2014) 331–339

Contents lists available at ScienceDirect

Physica A

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

Burst topic discovery and trend tracing based on Storm

Shihang Huang, Ying Liu, Depeng Dang*

College of Information Science and Technology, Beijing Normal University, Beijing 100875, China

HIGHLIGHTS

- A non-homogeneous Poisson process model was proposed to fit the trend of dataset.
- Heat degree factor and trend degree factor were presented to characterise the trend.
- Real time stream computing was used to trace burst topics.
- The effects of window size and trend degree threshold were both considered.

ARTICLE INFO

Article history:
Received 29 May 2014
Received in revised form 23 August 2014
Available online 8 September 2014

Keywords:
Non-homogeneous Poisson process
Storm
Burst topic
Trend

ABSTRACT

With the rapid development of the Internet and the promotion of mobile Internet, microblogs have become a major source and route of transmission for public opinion, including burst topics that are caused by emergencies. To facilitate real time mining of a large range of burst topics, in this paper, we proposed a method to discover burst topics in real time and trace their trends based on the variation trends of word frequencies. First, for the variation trend of the words in microblogs, we adopt a non-homogeneous Poisson process model to fit the data. To represent the heat and trend of the words, we introduce heat degree factor and trend degree factor and realise the real time discovery and trend tracing of the burst topics based on these two factors. Second, to improve the computing performance, this paper was based on the Storm stream computing framework for real time computing. Finally, the experimental results indicate that by adjusting the observation window size and trend degree threshold, topics with different cycles and different burst strengths can be discovered.

1. Introduction

With the development of the Internet, microblogs have become an important information-sharing platform [1]. Through microblogs, users can not only share their own fresh news around them but can also learn about the community current concerns and publish their opinions about hot issues that they are interested in [2]. The large number of the users in microblogs enables the formation of a complex social network, into which, as users continually release new microblogs relating to different topics, the microblogs have gradually become a barometer of public opinion. Among these topics, burst topics often relate to social emergencies. The spread of burst topics in microblogs will cause a widespread concern about the emergency in the community, and the degree of community concern will in turn be reflected by the spread of the burst topics. Therefore, burst topic discovery and trend tracing are helpful in discovering social emergencies and monitoring the development trend of such emergencies in real time.

* Corresponding author. Tel.: +86 13521534496.
E-mail address: ddepeng@bnu.edu.cn (D. Dang).

http://dx.doi.org/10.1016/j.physa.2014.08.059
0378-4371/

A microblog is usually short and has arbitrary language, which has presented challenges in the study of microblog topics [3]. Based on the features of microblogs, in this paper, by studying the regular occurrence of the words in microblogs, we first propose using a non-homogeneous Poisson process for fitting the variation trend of the words. For a better representation of word frequency changes with time, we have proposed heat degree factor and trend degree factor and have realised topic discovery and trend tracing based on these two factors. Additionally, because microblogs have a large number of users and there are a huge number of microblogs every day, traditional methods can hardly address such large-scale data, especially for real time analysis. Therefore, to address the challenges of big data [4–7] and meet the needs of real time analysis, we adopt the Storm stream computing framework to apply real time computing to the data accessed through microblog APIs. Finally, we conduct several experiments to observe and compare the topic characteristics that were discovered under different time window sizes. The results indicate that by adjusting the time window size, we can discover the burst topics with different cycles and by adjusting the trend degree threshold, we can discover the burst topics with different burst strengths. The research method and conclusion of this paper will serve as a reference to the real time mining of burst topics across a large number of topics; the research also has practical significance and value for other applications.

2. Related work

The discovery and analysis of network topics is an important research area in the study of Internet-based public opinion. In earlier studies, topic discovery algorithms only analysed static data. Ref. [8] focused on segmenting and clustering HashTags in Twitter to discover topics. Ref. [9] proposed a topic keyword extraction method for Twitter and ranked keywords through a topic PageRank algorithm. Unlike traditional Vector Space Model (VSM), Ref. [10] used Latent Dirichlet Allocation (LDA) topic model to extract the hidden topic information. Although the studies above can effectively discover topics, they do not consider the effect of temporal information on the topics.

In recent years, increasing numbers of topic discovery methods consider temporal information and discover and analyse the topics in streaming data. Additionally, there have been several studies on topic prediction and real time discovery. Ref. [11] proposed a text representation model based on emergencies. Compared with the traditional vector space model, this method considers temporal information, but still mines the topics from historical data. Ref. [12] established a topic propagation model and analysed the features of microblog data that can affect the spread of the topic, to predict the heat degree of short-term topics. However, this study needs to build a more complex community network model to simulate the spread of the topic. The Twitter Monitor shown in Refs. [13,14] can discover burst topics in Twitter stream; the burst topics are the words whose reach frequency goes beyond the usual time period. Twitter Monitor is able to discover those burst topics that have already received widespread attention, but does not trace the dynamic trend of the burst topics. Instead of calculating the average arrival rate of each word, the method proposed in Ref. [15] discovers hot topics by ranking the heat degree of the words. Compared with the Twitter Monitor, this method does not require the initialisation process, but it can only discover current hot topics.

Unlike the above studies, this paper is not only able to discover burst topics in real-time, but is also able to trace in real-time the variation trend of the burst topic. In this paper, we have proposed to use the method of process fitting to keep tracing the variation trend of the words, and as soon as possible, discover the burst topic and trace the development trend of the topic in a timely fashion in its growth period. Thus, we can discover social emergencies at the earliest available opportunity, and in real-time, acquire the community concerned about such emergencies and the impact of these emergencies on the community, and organise the relevant information to provide an objective basis for decision-making to the appropriate departments.

3. Word frequency trend calculation model

In this section, we first propose that if without the impact of the emergencies, the frequency of a word coincides with the Poisson distribution; next, we validate the assumption. In this study, the word frequency refers to the number of times that a word appears in the latest microblogs within one unit of time. In addition, to further represent the variation trend of the word frequency with respect to time, we have introduced the non-homogeneous Poisson process model and given the solution to the two parameters in the model, and also given the fitting efficiency of this model.

3.1. Word frequency process model

Homogeneous Poisson distribution is a commonly used discrete probability distribution [16–18]. The number of calls received by a telephone exchange, the number of particles emitted by radioactive substances and the number of passengers in a bus station and many other examples are often represented by a Poisson distribution. Because the occurrence number of a certain word within a unit time is similar to the above examples, we make the following Poisson assumption.

Assume that during a short period, without the impact of the emergencies, for any $t > s \geqslant 0$, the frequency of one word in the microblog $N_{s,t}(N_{s,t} = N_t - N_s)$ obeys a Poisson distribution with parameter $\lambda(t - s)$; i.e., for $k = 1, 2, \ldots$, the probability that the word frequency $N_{s,t} = k$ is as follows.

$$P\left(N_{s,t} = k\right) = \mathrm{e}^{-\lambda(t-s)}\left[\lambda\left(t - s\right)\right]^k / k!. \tag{1}$$