Microblog Topic Detection Based on LDA Model and Single-Pass Clustering*

Bo Huang, Yan Yang, Amjad Mahmood, and Hongjun Wang

School of Information Science & Technology, Southwest Jiaotong University, Chengdu, 610031, P.R China
Key Lab of Cloud Computing and Intelligent Technology, Chengdu, Sichuan Province, 610031, P.R China
huangbo1582@163.com, {yyang,wanghongjun}@swjtu.edu.cn, amjad.pu@gmail.com

Abstract. Microblogging is a recent social phenomenon of Web2.0 technology, having applications in many domains. It is another form of social media, recognized as Real-Time Web Publishing, which has won an impressive audience acceptance and surprisingly changed online expression and interaction for millions of users. It is observed that clustering by topic can be very helpful for the quick retrieval of desired information. We propose a novel topic detection technique that permits to retrieve in real-time the most emergent topics expressed by the community. Traditional text mining techniques have no special considerations for short and sparse microblog data. Keeping in view these special characteristics of data, we adopt Single-pass Clustering technique by using Latent Dirichlet Allocation (LDA) Model in place of traditional VSM model, to extract the hidden microblog topics information. Experiments on actual dataset results showed that the proposed method decreased the probabilities of miss and false alarm, as well as reduced the normalized detection cost.

Keywords: Microblog, topic detection, LDA model, Single-pass clustering.

1 Introduction

Microblogging has become a primary channel by which people not only share information, but also search for information. It fills a gap between blogging and instant messaging, allowing people to publish short messages on the web about what they are currently doing. First Microblog was launched by Evan William in 2006. According to Twitter, there were 175 million registered users by the end of 2010. This rapid adoption has generated interest in gathering information from microblogging about real time news and opinions on specific topics. This interest, in turn, has led to a proliferation of microblog search services from

* This work is partially supported by the National Science Foundation of China (Nos. 61170111, 61003142 and 61152001) and the Fundamental Research Funds for the Central Universities (No. SWJTU11ZT08).
both microblogging service providers (like Twitter) and general purpose search engines (like Bing and Google). However, compared with traditional document retrieval and web search, microblog search is still in its infancy.

In a typical microblog search scenario using Twitter, around 1500 tweets that contain the query terms, will be returned, ranked by their creation time. Although, other presentation formats are also available (e.g., ordering results by author popularity, or by hyperlinks referenced), presentation formats optimized for topic monitoring are not yet widely available. The goal of this paper is to explore the potential for topic organization of microblog search results.

This is a challenging problem because microblog posts are short and sparse, so traditional topical clustering technique based on lexical overlap is necessarily weak. We use single-pass clustering method with Latent Dirichlet Allocation (LDA) Model instead of traditional VSM model [1]. The experimental results has proved the effectiveness of LDA model over VSM.

The rest of this paper is organized as follows: Section 2 presents the current state of topic detection. Section 3 explains the Latent Dirichlet Allocation (LDA) model and the MCMC method with Gibbs sampling for LDA. Section 4 covers the methodology of Single-pass clustering algorithm. Section 5 describes the experiments and analysis of results. Finally, section 6 discusses the conclusion and future work.

2 Related Work

Yang et al. [2] investigates the use and extension of text retrieval and clustering techniques for event detection using hierarchical and non-hierarchical document clustering algorithm. They found that resulting clustering hierarchies are highly informative for retrospective detection of previously unidentified events. Trieschnigg and Kraaij [3] proposed an incremental hierarchical clustering algorithm. They take a sample from the corpus to build a hierarchical cluster structure, then optimize the resulting binary tree for the minimal cost metric, finally assign the remaining documents from the corpus to clusters in the structure obtained from the sample. Papka and Allen [4] detect topic by using a Single-pass clustering algorithm and a novel thresholding model. This model incorporates the properties of events as major component, but the priori report sparse will lead to the topic model is not accurate. Finally, explored that the probabilities of miss alarm and false alarm may increase with the Single-Pass Clustering. Cataldi et al. [5] proposed the new hot topic detection methods based on the relationship between the timing and the social evaluation Twitter. In an appropriate period of time, if a topic has been widely detected, but before this rarely occurs, then you can think that this topic is the new hot topic at this particular moment. Phuvipadawat and Murate [6] put forward a collection of breaking news on Twitter, He designed a program called “Hotstream” to provide users breaking news.

In the topic detection process, building Model is a basic challenge. The vector space model (VSM) is the most common model. For the short and sparse