# 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).

J.T. Yao et al. (Eds.): RSCTC 2012, LNAI 7413, pp. 166–171, 2012.

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 contains 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. Sectiona 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

microbolgging text, VSM (using words or terms as characters) cannot perform accurate calculation of the text Similarity. In order to reduce the date scarcity and make it more topic-focused, we propose the LDA model [6] to the data modeling, extracting the hidden microblog topics information. High-dimensional sparse text vector is mapped to low-dimensional hidden topic space, combined with the classic single-pass clustering algorithm for text clustering different topic.

## 3 The Method of Microblog Text Modeling

### 3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA)[7] is a generative probabilistic model for a corpus of discrete data. It models the words in documents under the "bag-of-words" assumption, ignoring the orders of the words. Following this "exchangeability", the distribution of the words would be independent and identically distributed under some given conditioned of parameters. This conditionally independence allows us to build a hierarchical Bayesian model for a corpus of documents and words. This process can be described graphically as shown in Fig.1.

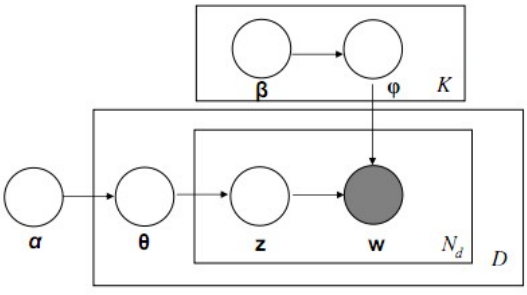


**Fig. 1.** Graphical model representation of LDA

For each document d in the corpus, the LDA model first picks a multinomial distribution $\theta_d = [\theta_{d1}...\theta_{dk}]^T$ from the Dirichlet distribution $\alpha_d = [\alpha_{d1}...\alpha_{dk}]^T$, and then the model assigns a topic $z_{id} = k$ to the $i$th word in the document according to the multinomial distribution $\theta_d$. Given the topic $z_{id} = k$,the model then pick a word $w_{id}$ from the vocabulary of $V$ words according to the multinomoial distribution $[\phi_{k1}...\phi_{kV}]^T$ which is generated from the Dirichlet distribution $[\beta_{k1}...\beta_{kV}]^T$ for each topic $k$.

Markov Chain Monte Carlo (MCMC)[8] is a general method to obtain samples from complex distribution. We have to construct a Markov chain that is irreducible, a periodic, and reversible in order to make the chain have a unique stationary distribution. Such properties are guaranteed if we apply the Gibbs sampling for the state transitions [9].The algorithm is detailed as follows:

We first consider the joint distribution of $z$ and $w$

$$p(w, z|\alpha, \beta) = \int_{\theta} \int_{\phi} p(w, z, \theta, \phi|\alpha, \beta) d\theta d\phi \tag{1}$$

Given the joint distribution of $w$ and $z$ under LDA, we can compute the conditional probability for the Gibbs sampler by

$$p(z_{id} = k|z^{\neg id}, x, \alpha, \beta) = \frac{p(z^{\neg id}, z_{id} = k|x, \alpha, \beta)}{\sum_{k`=1}^{K} p(z^{\neg id}, z_{id} = k`|x, \alpha, \beta)} \tag{2}$$

After the Markov chain reach the stationary distribution, we can start drawing samples from the chain. As shown in [8], given a sampled $z$, we can estimate the values of the other latent variables by

$$\theta_{dk} = \frac{\alpha_k + n_{dk}}{\alpha + n_d}, \phi_{kv} = \frac{\beta_{kv} + n_{kv}}{\beta_k + n_k} \tag{3}$$

where the counts are obtained from the assignment $z$. The above two equations are derived by computing the expectance of the Dirichlet distribution in the posterior form.

## 4 Topic Detection by Single-Pass Clustering

As a result of Gibbs sampling for LDA, $\theta$ is a $d * k$ matrix,where $d$ is the total number of microblog texts, $k$ is the number of latent topics. Matrix element value indicates the probability of each text data set to generate implicit topic, can also be seen as the document-topic vectors.

The proposed Single-pass clustering[3] algorithm is as following:

For each document $d$ in the sequence loop;
  Find a cluster $c$ that maximizes $\cos(c, d)$;
    If $\cos(c, d) > t$ then
      Include $d$ in $c$;
    Else create a new cluster whose only document is $d$;
  End loop.

## 5 Experiment and Results

### 5.1 Evaluation Criteria

Detection performance is characterized in terms of the probability of Miss and False alarm errors ($P_{Miss}$ and $P_{FA}$). These error probabilities are then combined into a single detection cost $C_{Det}$, by assigning costs to miss and false alarm errors[10]:

$$(C_{Det}) = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{non-target} \tag{4}$$

According to the TDT standards,we set $C_{Miss} = 1.0, C_{FA} = 0.1, P_{target} = 0.02$.

Because these values vary with the application, $C_{Det}$ will be normalized so that $(C_{Det})_{Norm}$ can be no less than one without extracting information from the source data. This is done as follows:

$$(C_{Det})_{Norm} = \frac{C_{Det}}{min(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{non-target})} \tag{5}$$

The $(C_{Det})_{Norm}$ is smaller,the quality of topic detection is better.

### 5.2 Dataset

We collected 108122 texts of Sina-microblog August 2011 by Web crawler. All data covering 957 topics discussed by the microbolg usrers. Before the experiment, data preprocessed by ICTCLAS Segmentation system.

### 5.3 Results

After fixing different similarity threshold $t$, the Single-pass clustering based on VSM model and the Single-Pass clustering based on LDA model are executed. The corresponding experimental results are shown in Table 1 and 2. Followings observations are found in the results:

1. With increasing similarity threshold $t$, the missing rate increases gradually, and the fault detection rate decreases gradually, consuming function is down then up.
2. The Single-pass Clustering based on LDA model can reduce the $P_{Miss}$ ,$P_{FA}$ and$(C_{Det})_{Norm}$ to improve the topic detection accuracy.

**Table 1.** The results of Single-pass based on VSM model

| $t$ | 0.001 | 0.002 | 0.003 | 0.005 | 0.008 | 0.01 | 0.02 | 0.05 | 0.08 |
|---|---|---|---|---|---|---|---|---|---|
| $P_{Miss}$ | 0.2145 | 0.2386 | 0.3258 | 0.3322 | 0.3664 | 0.4615 | 0.4962 | 0.5138 | 0.5567 |
| $P_{FA}$ | 0.3360 | 0.3196 | 0.2862 | 0.2635 | 0.1179 | 0.1128 | 0.1012 | 0.1073 | 0.0943 |
| $(C_{Det})_{Norm}$ | 0.0372 | 0.0361 | 0.0346 | 0.0325 | 0.0189 | 0.0203 | 0.0216 | 0.0249 | 0.0313 |

**Table 2.** The results of Single-pass based on LDA model

| $t$ | 0.01 | 0.02 | 0.03 | 0.05 | 0.08 | 0.1 | 0.2 | 0.3 | 0.5 |
|---|---|---|---|---|---|---|---|---|---|
| $P_{Miss}$ | 0.0110 | 0.0126 | 0.0531 | 0.0639 | 0.1128 | 0.1532 | 0.2704 | 0.3001 | 0.3552 |
| $P_{FA}$ | 0.1538 | 0.1464 | 0.1052 | 0.0841 | 0.0533 | 0.0094 | 0.0002 | 0.0000 | 0.0000 |
| $(C_{Det})_{Norm}$ | 0.0152 | 0.0146 | 0.0133 | 0.0095 | 0.0074 | 0.0040 | 0.0054 | 0.0060 | 0.0071 |

## 6 Conclusion

Considering the in-built characteristics of large-scale and high-sparse microblog data, we purposed the Single-Pass Clustering algorithm based on LDA model to solve the data sparseness problem faced by the traditional VSM. The experimental results show that the algorithm could decrease the probabilities of Miss Alarm and False Alarm, and finally reducing the normalized detection cost. Future research will optimize the LDA model, and consider the real-time processing of larger data.

## References

1. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18, 613–620 (1995)
2. Yang, Y., Pierce, T., Carbonell, J.: A study on Retro-spective and On-Line Event detection. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, USA, pp. 28–36 (1998)
3. Trieschnigg, D., Kraaij, W.: TNO hierarchical topic detection report at TDT 2004. In: The 7th Topic Detection and Tracking Conf. (2004)
4. Papka, R., Allan, J.: On Line New Event Detection using Single Pass Clustering. UMass Computer Science (1998)
5. Cataldi, L., Caro, D., Schifanella, C.: Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation. In: MDMKDD 2010 Proceedings of the Tenth International Workshop on Multimedia Data Mining, Washington, pp. 1–10 (2010)
6. Phuvipadawat, S., Murata, T.: Breaking News Detection and Tracking in Twitter. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Toronto, pp. 120–123 (2010)
7. Blei, D., Ng, A., Jordan, M., et al.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
8. Stuart, G., Donald, G.: Stochastic relaxation gibbs distributions and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 7212–7411 (1984)
9. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Science 101, 5228–5235 (2004)
10. The Linguistic Data Consortium.: The 2004 Topic Detection and Tracking. Task Definition and Evaluation Plan (2004), `http://www.itl.nist.gov/iad/mig/tests/tdt/2004/TDT04.Eval.Plan.v1.2.compare.1.1c`