MR-LDA:  
An Efficient Topic Model for Classification of Short Text in Big Social Data  

Xiongwen Pang, School of Computer, South China Normal University, Guangzhou, China  
Benshuai Wan, Information Technological Department, Guangdong Nanhai Rural Commercial Bank Company Limited, Guangzhou, China  
Huifang Li, School of Computer, South China Normal University, Guangzhou, China  
Weiwei Lin, School of Computer Science and Engineering, South China University of Technology, Guangzhou, China  

ABSTRACT  

Latent Dirichlet Allocation (LDA) is an efficient method of text mining, but applying LDA directly to Chinese micro-blog texts will not work well because micro-blogs are more social, brief, and closely related with each other. Based on LDA, this paper proposes a Micro-blog Relation LDA model (MR-LDA), which takes into consideration the relations between Chinese micro-blog documents and other Chinese micro-blog documents into consideration to help topic mining in micro-blog. The authors extend LDA in the following two points. First, they aggregate several Chinese micro-blogs as a single micro-blog document to solve the problem of short texts. Second, they model the generation process of Chinese micro-blogs more accurately by taking relationship between micro-blog documents into consideration. MR-LDA is more suitable to model Chinese micro-blog data. Gibbs sampling method is borrowed to inferences the model. Experimental results on actual datasets show that MR-LDA model can offer an effective solution to text mining for Chinese micro-blog.

KEYWORDS  
Big Data, Latent Dirichlet Allocation, Micro-Blog, Social Network, Topic Mining

1. INTRODUCTION  

In recent years, the text topic mining has been gaining extensive attention. Various algorithms emerge, among which topic model Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003) is widely used to mine the topics of large text corpora. 

In 2003, Blei et al. (2003) proposed LDA based on Latent Semantic Analysis and probabilistic latent semantic analysis. LDA is a topic model and a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. LDA is based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents. It specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics.
Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents.

Micro-blog platforms such as Tencent Micro-blog and Sina Micro-blog have become a part of peoples’ daily life, from which we can gain information timely to keep in touch with the world at any time. However, sometimes we may sink into the massive information. A lot of time can be saved for users in data mining if micro-blog can be classified efficiently. The purpose of this paper is to find an efficient topic model for classification of Chinese micro-blog. However, the problem is still more intractable than classification of any traditional documents, since micro-blog posts suffer from short texts, heavy noise and bad normalization, while traditional documents are usually in nice structure, long texts and clear semantic. To overcome the above difficulties, we propose a new method named Micro-blog Relation LDA model (MR-LDA) to classify micro-blog by aggregating several Chinese micro-blogs as a single micro-blog document to solve the problem of short texts and taking relationship between micro-blog documents into consideration to improve the performance of classification for Chinese micro-blog. Most existing micro-blog classification methods suffer from low performance. Unlike most existing methods, the authors’ new model can offer an effective solution to text mining for Chinese micro-blog. MR-LDA is more suitable for model Chinese micro-blog data.

The main contributions of the authors’ work can be summarized as follows. 1) This paper proposes a new model to classification of Chinese micro-blog by taking relationship of micro-blog documents into consideration. 2) The authors propose a way to overcome problem of short texts by using a sliding window to expand the texts. According to the size of the sliding window (MN), MN micro-blogs combined into a text. 3) We propose a new method to compute relational value between Chinese micro-blogs. 4) This paper proposes new methods to compute the parameter $\alpha$ of Dirichlet distribution on MR-LDA model. Therefore, the MR-LDA method is more suitable to model Chinese micro-blog data. Compared with classification of Chinese micro-blog based on LDA, the experimental results on actual dataset demonstrate the superiority of the authors’ method.

The outline of this paper is as follows. The authors briefly review the related work in Section 2, and then describe the authors’ proposed MR-LDA method in details in Section 3. Experimental results on actual Chinese micro-blog dataset are shown in Section 4, followed by a conclusion.

2. RELATED WORK

Data mining in social media has recently received attention from many researchers. Giovanni et al presented a novel method for clustering words in micro-blogs, based on the similarity of the related temporal series (Stilo & Velardi, 2016). Lazard et al. (2015) applied a textual analytics method to reveal insights from these tweets that can inform communication strategies. Yang, Mu and Shen (2015) applied text mining and geospatial methods to detect tweets related to depression and their spatiotemporal patterns at the scale of Metropolitan Statistical Area. Xiao, Ai, Ching-Hsien and Wang (2015) proposed a time-ordered collaborative filtering recommendation algorithm (TOCF), which takes the time sequence characteristic of user behaviors into account. Zhang, Hsu, Chen and Chen proposed bi-clustering and fusion (BiFu)- a newly-fashioned scheme for the cold-start problem based on the BiFu techniques under a cloud computing setting. Wang, Huang, Hsu and Yang (2016) proposed collaboration reputation to enhance the trustworthiness of Web service selection. Huang, Wang, Hsu, Zhang and Yang (2015) introduced an accurate reputation measurement approach, which uses both virgin and non-virgin reputation scores to shield services against malicious feedback ratings. Bouazizi and Ohtsuki (2015) proposed a method that makes use of a minimal set of features, yet, efficiently
12 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:

www.igi-global.com/article/mr-lda/172507?camid=4v1


www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

One Anchor Distance and Angle Based Multi - Hop Adaptive Iterative Localization Algorithm for Wireless Sensor Networks
www.igi-global.com/chapter/one-anchor-distance-angle-based/61996?camid=4v1a

Scalable Fault Tolerance for Large-Scale Parallel and Distributed Computing
www.igi-global.com/chapter/scalable-fault-tolerance-large-scale/36433?camid=4v1a

Event-Driven Mobile Computing with Objects
www.igi-global.com/chapter/event-driven-mobile-computing-objects/44406?camid=4v1a
Web Services in Distributed Information Systems: Availability, Performance and Composition
[www.igi-global.com/article/web-services-distributed-information-systems/39039?camid=4v1a](www.igi-global.com/article/web-services-distributed-information-systems/39039?camid=4v1a)