Knowledge-Based Systems 82 (2015) 102–114

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Dynamic non-parametric joint sentiment topic mixture model

Xianghua Fu, Kun Yang, Joshua Zhexue Huang, Laizhong Cui *

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China

ARTICLE INFO

Article history:
Received 13 August 2014
Received in revised form 31 January 2015
Accepted 23 February 2015
Available online 10 March 2015

Keywords:
Topic sentiment analysis
Dynamic topic analysis
Non-parametric topic model
Social media
Hierarchical Dirichlet Process
Text mining

ABSTRACT

The reviews in social media are produced continuously by a large and uncontrolled number of users. To capture the mixture of sentiment and topics simultaneously in reviews is still a challenging task. In this paper, we present a novel probabilistic model framework based on the non-parametric hierarchical Dirichlet process (HDP) topic model, called non-parametric joint sentiment topic mixture model (NJST), which adds a sentiment level to the HDP topic model and detects sentiment and topics simultaneously from reviews. Then considered the dynamic nature of social media data, we propose dynamic NJST (dNJST) which adds time decay dependencies of historical epochs to the current epochs. Compared with the existing sentiment topic mixture models which are based on latent Dirichlet allocation (LDA), the biggest difference of NJST and dNJST is that they can determine topic number automatically. We implement NJST and dNJST with online variational inference algorithms, and incorporate the sentiment priors of words into NJST and dNJST with HowNet lexicon. The experiment results in some Chinese social media dataset show that dNJST can effectively detect and track dynamic sentiment and topics.

1. Introduction

The rise of the social media motivates people to express their sentiment and opinions about anything more freely and frequently than ever before. For most commercial organizations and government departments, the user generated reviews represent invaluable source of information. Although many works have been done to extract information from reviews, summarize user's opinions, and categorize reviews according to opinion polarities [15,18,20,22,26,30,32,45], it is still a challenge for users to easily digest and exploit the large number of reviews due to the inadequate supports for understanding individual reviewer's opinions at the fine-grained level of topical aspects [9,41,48]. In fact, most existing works detect sentiment in isolation of topic detection. To address this issue, topic models are introduced for simultaneous analysis of topics and sentiment in a document. These studies, which jointly model topic and sentiment, take the advantage of the relationship between topics and sentiment, and are shown to be superior to traditional sentiment analysis tools [17,18,21,37,38,53]. But these methods only consider the static dataset, Bollen's [5] and Connor's [24] have shown that sentiment dynamics of online contents have a strong correlation with the fluctuations of macroscopic social and economic indicators in the same time period. Furthermore, social media data are produced continuously by many uncontrolled users, so the dynamic nature of such data requires the sentiment and topic analysis model to be updated dynamically.

To the best of our knowledge, TSM [21] and dJST [11] are the very few studies to detect and track dynamic topic and sentiment based on probability topic model, where TSM is based on probability latent semantic analysis (pLSA) model [12], and dJST is based on latent Dirichlet allocation (LDA) [4]. Since pLSA and LDA are parametric probabilistic model, both of them require to determine the topic number beforehand. It is insufficient for the dynamic and massive social media data. Furthermore, dJST is implemented with the Gibbs sampling algorithm, the drawbacks of which include: they are often hard to access convergence of the Markov chains, and they are not sufficient to deal with massive corpus [29,40].

In this paper, we propose a dynamic non-parametric joint sentiment topic model (dNJST) for detecting and tracking dynamic sentiment and topics of social reviews. We introduce a non-parametric joint sentiment topic model (NJST) through adding a sentiment level to the hierarchical Dirichlet process (HDP) topic model, and then present dynamic NJST (dNJST) which adds time decay dependencies of historical epochs to the current epochs. Compared with the existing sentiment-topic models, the biggest difference of dNJST is that dNJST can determine topic number automatically. Furthermore, we implement dNJST with an online variational inference algorithm, and improve the sentiment identification by HowNet lexicon. The experiment results show that dNJST can effectively detect and track dynamic sentiment and

* Corresponding author.
E-mail address: cuilz@szu.edu.cn (L. Cui).

http://dx.doi.org/10.1016/j.knosys.2015.02.021

topic of Chinese social media. The main contributions of this paper are four-folds:

(1) We propose the non-parametric joint sentiment and topic model (NJST) and its dynamic version dNJST. Different with the existing sentiment topic mixture models, NJST and dNJST add sentiment levels to the non-parametric HDP topic model, which can determine the topic number of each epoch automatically. To the best of our knowledge, both NJST and dNJST are the first work to attempt dynamic sentiment and topic detection based on the non-parametric HDP topic model.

(2) We implement online variational algorithms for NJST and dNJST, which can compute quickly for large corpus. The existing JST and dJST models are implemented with the standard Gibbs sampling algorithms, and they are difficult to deal with massive data because they have to repeatedly sample from the posterior topic assignment for each word token through the entire corpus at each iteration.

(3) The main purpose of the sentiment and topic mixture model is to extract the sentiment and topics from social media reviews. We apply our dNJST model to discover the dynamic sentiment and topics of Chinese social media with real social media data from the biggest Chinese Web online forum "Tianya Forum". We compare the performance of dNJST with NJST, JST and dJST. The experimental results show that dNJST outperform NJST, JST and dJST in extracting topics of specific sentiment orientation, which indicates the effectiveness of our dynamic non-parametric model.

The remainder of the paper is organized as follows. In Section 2, we introduce some related works. The Non-parametric Joint Sentiment-Topic Model is proposed in Section 3. Section 4 introduces the Dynamic NJST model and its online variational algorithm. Experiments and evaluations are reported in Section 5. We conclude the paper in Section 6 with future researches.

2. Related work

In recent years, there have been many research works on sentiment analysis and opinion mining [9,25,33]. Sentiment analysis methods usually can be divided into two categories: (1) the first type is based on part-of-speech (POS) tagging of the words and sentiment lexicons. This type is proposed by Peter [39] at first. Qun Liu [19] established a believable vocabulary on Chinese semantic knowledge named HowNet lexicon, and then got the sentiment polarity of words through comparison with the similarity between the words. Yanlan Zhu [54] succeeded in judging semantic orientation of Chinese online reviews based on the HowNet lexicon. Some other works such as Linhong Xu [47] and Weifu Du [8] adopted this type of method for the sentiment analysis of Chinese online reviews. (2) the second type is based on machine learning algorithms. For example, Whitelaw used support vector machine (SVM) to classify the sentiment orientation of movie reviews [43], Socher [30] trained a sentiment Treebank with Recursive Neural Tensor Network to identify the sentence level sentiment. Some researchers also used machine learning algorithms to classify the sentiment orientation of Chinese online reviews, e.g., Jun Xu used Naive Bayesian (NB) and maximum entropy [46] to classify sentiment of Chinese news, Yi Hu [13] compared the qualities of sentiment classification between SVM and NB classifier.

Although researchers have got many achievements, most existing works focused on the sentiment classification of the product and service reviews. Only a few researchers pay attention to the sentiment analysis of the social reviews. For example, Mullen and Malouf [23] described preliminary statistical tests on a new dataset of political discussion group postings. Somasundaran [31] explored the utility of sentiment and arguing opinions for classifying stances in ideological debates. Mei [21] proposed Topic-Sentiment Mixture (TSM) model to reveal the latent topical facets in a Weblog collection. Bollen [5] performed a sentiment analysis of tweets, and found that the events in the social, political, cultural and economic sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood. In Chinese social reviews' sentiment analysis, Tao [35] proposed an approach for feature extraction of sentiment analysis of the news comments. Yang [50] attempted to construct a new sentiment lexicon with sentiment orientation extent based on existing HowNet and NTUSD, which was applied to a semi-automatic Web public opinion analysis system. Daifeng Li [16] proposed a Topic-level Opinion Influence Model (TOIM) to predict the users' future opinions on specific topics with a large dataset from Tencent Weibo. In addition, because of the figurative language such as irony, metaphor is generally used in social media, some researchers focus on identifying the figurative language and analyzing its polarity [27,28].

On the other hand, probabilistic topic models provide a principled and elegant way to discover hidden topics from large document collections [3]. pLSA [12] and LDA [4] are widely-used probabilistic topic models. One of the main advantages of LDA is that it can be easily used as a module in more complicated models for more complicated goals. A number of extensions to LDA have been proposed. For example, to detection the topics in social media, Yan [49] proposed a bitterm topic model (BTM) for modeling topics in short texts, and Diao [6] proposed a topic model that captured the similar relation among posts. When the models like LDA are used, the question usually arises that how many topics the estimated model should have, given the document collection [51]. The problem can be addressed by sharing a discrete base distribution among documents. A hierarchical Dirichlet process (HDP) [36] creates such a discrete base distribution for the document Dirichlet processes (DPs) by sampling from another DP. To learn evolutionary topics from a time varying corpus, some works have focused on extending LDA and HDP to dynamic topic models. Wang and McCallum [42] presented a LDA-style topic model called Topic Over Time (TOT) that explicitly modeled time jointly with word co-occurrence patterns. Tang [34] proposed a new generative model to simulate the generation process of both web contents and user's participation in a unified framework. Kawamae [14] presented the theme chronicle model (TCM) which divided traditional topics into temporal and stable topics to detect the change of each theme over time. Amr Ahmed [1] presented an infinite dynamic topic models (iDTM) based on HDP topic model, which allowed for unbounded number of topics: topics can die or be born at any epoch, and the representation of each topic can evolve according to a Markovian dynamics. Zhang [52] proposed an evolutionary hierarchical Dirichlet process (EvoHDP) model.

Recently, probabilistic topic models are also introduced for simultaneous analysis of topics and sentiment in a document. These studies, which jointly model topic and sentiment, take the advantage of the relationship between topics and sentiment, and are shown to be superior to traditional sentiment analysis tools. The first topic and sentiment model is the Topic-Sentiment Model (TSM) [21], which jointly models the mixture of topics and sentiment predictions for the entire document. Because TSM is essentially based on pLSA model with an extra background component and two additional sentiment subtopics, it suffers from the problems of inference on new document and overfitting the data. Titov and McDonal [37,38] proposed the Multi-grain Latent Dirichlet Allocation model (MG-LDA) to build topics that were