Research on text mining model

Peiyao Zhang

School of Economics and Management, Xidian University, Xi'an 710126, China

zhangpeiyao366@163.com

Abstract

The number of comments and information in the Internet is huge and rapid, and the huge amount of comments expresses the emotional tendency of the information publishers. Topic model can be automatically extracted from a large-scale discrete data set implied semantic information to generate probability model, this paper introduces the common improvement and extension model of LDA, finally analyzes the topic model based emotional hybrid model, the theme of the information can be extracted without supervision and the theme of the corpus and the corresponding emotion tendencies.

Keywords

Thematic model, thematic emotional hybrid model, LDA, BTM.

1. Introduction

Language and writing play an important role in the history of human evolution, and the content of the text records the development of politics, economy and culture around the world. People have long recognized the importance of text messages, trying to preserve, organize and disseminate them. In the age of information carries, and its quantity is increasing at an alarming rate. In July 2017, according to China Internet network information center (CNNIC) in Beijing issued a 40 times the China Internet network development state statistic report shows that as of June 2017, the scale of Chinese Internet users reached 751 million, China`s Internet penetration rate of 54.3%, a 1.1% in increase at the end of 2016; The number of mobile Internet users is 724 million, the proportion of Internet users is 96.3%, and the proportion of mobile Internet users continues to increase[1]. However, the vast amount of text information services and people also create new problems: for example, it is difficult to find useful information accurately and comprehensively when text distribution is scattered.

Text mining is the method of analyzing the results of data mining to analyze the text described in natural language. The thematic model represented by the model of LDA[2] is a hot research direction in the field of text mining in recent years. Topic model has good ability of dimension reduction, the use of topic modeling unearthed can help people understand the theme of the massive amounts of text hidden semantics, can also act as other input text analysis method, finish the text classification, topic detection, automatic text summarization and relation of text mining tasks. LDA topic model has been a great success in traditional network text mining based on news data.

2. The Basic Principles of the LDA Theme Model

The main idea of LDA is that documents are a mixed distribution of several topics, and each topic is a probability distribution of words. So it can be implied theme as a probability distribution of vocabulary (Topic, word), a single document can be expressed as the probability distribution of these implied themes(Doc-Topic), this assumption is also conducive to large-scale space dimension reduction in data

processing, namely the documents onto the Topic space. Figure 1 shows the schematic diagram of the LDA implicit theme topology.

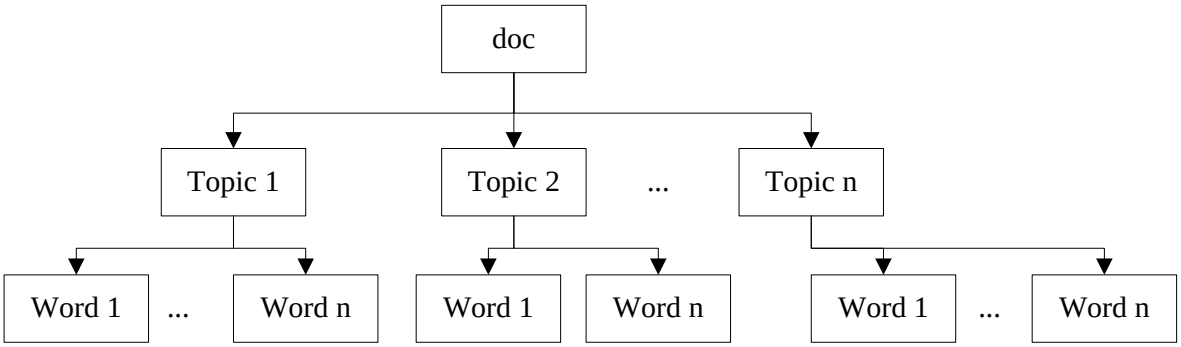

Figure 1. Schematic diagram of LDA implicit theme topology

3. The Application of the Extended Model Based on LDA in Micro-Scale

(1). ATM(author-topic model)[3] to the author of the article information is introduced to guide the LDA theme generation, different from the LDA is that of "text-theme" distribution in the model replaced by "the author-theme" distribution. ATM modeling gets a mix of topics at the user level, rather than the topic at the post level.

(2). Twitter-LDA[4] can not only model the topic at the user level of weibo, but also model the topic of a single weibo post.

(3). Labeled LDA[5] considers weibo as a kind of network text, and many data have been Labeled as tags by readers, which can help to better carry out topic mining by utilizing the existing tag resources. Labeled LDA USES a simple constraint on the theme model to introduce monitoring, which is to use only those topics that correspond to the collection of tags that can be observed with a single document. The theme of the model learning is directly associated with each tag. Labeled LDA`s advantage is by providing a Set of Label(Label Set) to the learning process, make its have the explanation, the theme of learning to get more can also be used to solve the credible belonging problem in text categorization.

(4). Mb-LDA has been extended on the basis of the original LDA, and the text association relationship and contact relation of weibo are introduced into the topic modeling of microblog.

4. This Topic Model is BTM

BTM topic model[6] broke the document topic layer of the traditional theme model. By translating the document into word pairs, the word pair refers to two words of any co-occurrence after the document is preprocessed. For the whole corpus, the topic of modeling learning is used to overcome the sparse problem and semantic relation between the words is taken into consideration, and the information of micro-blog is better understood than the traditional thematic model. The theme learning process of this model does not need any external data, which is also the first general topic model.

Expand the study of the model that also includes text sentiment analysis aspect, for example, Lin Chenghua joint emotional topic model put forward by JST(be sentiment/topic model) in the absence of supervision, extract text theme, and at the same time realize document level of emotional detection; Mei Qiaozhu theme emotional mixture model put forward by the TSM(topic-sentiment mixture) extract theme words can be divided into three classes of neutral, positive and negative, support the topic level detection of emotion and emotional analysis of the dynamic evolution over time.

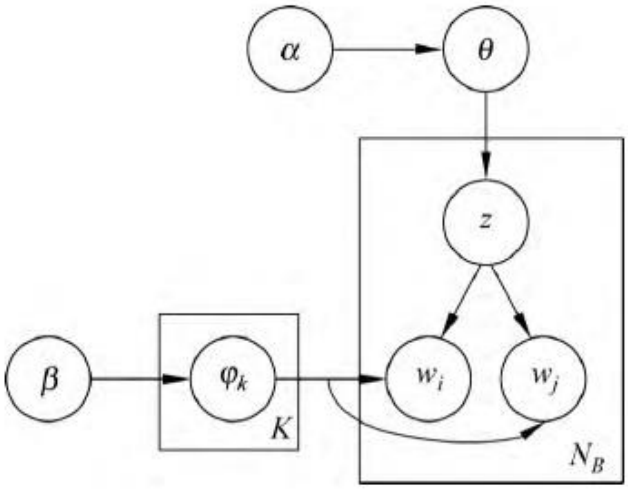

Figure 2. BTM model

5. Thematic Emotional Hybrid Model

Emotional analysis refers to analyzing and sorting out the emotional tendency information expressed by users through the mining of text with personal views, and text emotion analysis is an important part of emotional calculation.

There are two ways to express thematic emotion mixed model in language model. The first is to depict themes and emotions as a single language model, in which a word may be associated with both subject and emotion, such as ASUM model and JST model. The other is to use emotion and theme as separate language models, a word that is either an emotional word or a subject word, and only one, such as the TSM model.

(1). The LSM model [7] (Latent Sentiment Model), the model will be emotional as a special case of the theme, think that the distribution of document vocabulary related to emotion, so as to realize the document of unsupervised classification of emotion, but unable to identify the theme of the more granular information.

(2). The TSM model [8] (Topic Sentiment Mixture Model) is able to extract unsupervised document theme and emotional information, but the TSM model with PLSA algorithm (aim-listed Probability Latent Semantic Analysis) as the foundation, restricted by its prone to fitting.

(3). ASUM model [9] (Aspect and Sentiment Unification Model) has established the theme-word "sentence" three layers model, ASUM on the basis of LDA, generalization ability is stronger, ASUM model, the same word in the sentence come from the same language model, the assumption is too strict.

(4). The JST model [10] (be Sentiment/Topic Model) is a kind of themes and emotional information can be extracted without supervision and document the four layers of Bayesian network, this model framework with a layer of more than LDA and between layer and layer in the document to join an additional emotional layer similar to these methods mentioned above, the main consideration within a single document word co-occurrence information, more suitable for the long text subject emotional conjoint analysis.

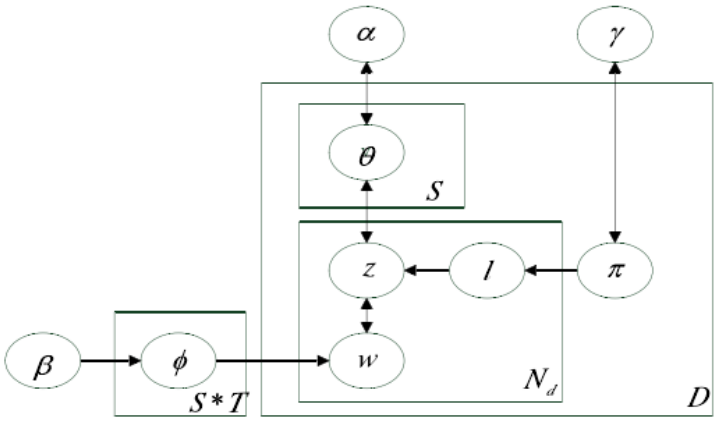

Figure 3. JST model

(5). SSTM model (Short-text Sentiment-topic model) to improve document emotional classification accuracy, but this method is based on each topic in both positive and negative two hypothesis of emotion, does not quite agree with the actual situation, the algorithm in inference document emotion, every word of emotional binarization also affect the emotion classification precision.

Subjective emotional tendency of text is usually correct the appreciation and affirmation or negative criticism and negation, if the two mixed emotions hypothesis of subjective hidden in the text of the two kinds of emotional topic, using a topic model, examining the relationship between the emotional theme and the emotion characteristic is a very new idea – and experiment validation for a plus or minus two emotions under each topic hypothesis.

(6). BJSTM model increases the setting of the emotional layer on the basis of BTM, thus forming the three-layer Bayesian model of "emotion-theme-word". The BJSTM model makes full use of the rich lexical co-occurrence and word frequency information at the corpus level, which reduces the influence of the short text feature sparse on the subject/emotion joint analysis.

6. Conclusion

In this paper, the principle technology and development of LDA model are expounded from the perspective of model representation. Through the analysis of the improvement of LDA in each text set and the application of emotion analysis fully explain that the theme model has a large development space in the field of data mining, which reflects the powerful modeling function of thematic model.

Reference

[1] China Internet network center.40 times the China Internet network development state statistic report.http://www.cac.gov.cn/2017-08/04/c_1121427728.htm,2017.

[2] Blei David, Ng Andrew and Jordan Michael. Latent Dirichlet Allocation[J].The Journal of Machine Learning Research,2003,(3): 993-1002.

[3] Michal Rosen-Zvi, Thomas Griffiths, etc. The Author-topic Model for Authors and Documents[J].

[4] Zhao W X, Jiang J, etc. Comparing twitter and traditional media using topic models. In : Proc. of ECIR , 2011:338-349.

[5] Ramage D, Hall D, Nallapati R , etc. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1-Volume 1.Association for Computational Linguistics , 2009:248-256.

[6] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C] //Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013: 1445－1456.
[7] Lin C,He Y,Everson R.A comparative study of Bayesian models for unsupervised sentiment detection.
[8] Mei Q,Ling X, Wondra M, etc. Topic sentiment mixture: modeling facets and opinions in weblogs.
[9] Yohan Jo,AH Oh.Aspect and Sentiment Unification Model for Online Review Analysis.
[10] Lin C,He Y,Joint sentiment/topic model for sentiment analysis.