A Critical Review on Probablistic Topic Model

Dehao Li and Qi Shao

Abstract

Social micro blogging service has been one of the most important web services to publish and share different kinds of information. The social information is mostly stored as unstructured model. Retrieval and extraction of the information is essential work and important in the semantic web area. Natural language Processing (NLP) is very efficient and intelligent for processing amounts of textual information. Probabilistic Topic Model (PTM) is one kind of statistical human language processing methodologies which needs sets of corpus as if indicating the desirable relations and dependencies. In this review, we will discuss several related studies using PTM to demonstrate the feasibility of applying PTM to extract valuable information from social information.

Keyword: **Social Information, Natural Language Processing, Probabilistic Topic Model**.

1. Introduction

Social network is mostly referred to as a combination of Information Technologies to make communication and conversations into an interactive format. As a part of social network, social micro blogging service has been one of the most important web services to publish and share different kinds of information. The trend of utilizing social information is to leverage different features which reflect valuable information, including both textual contents and social structure. Social text contains rich information about author of the information, such as what the user is talking about, what the user is going through and so on, so social text is potentially useful in determining specific and common characteristics of an individual user. In this review, we focus on a specific model (PTM) for investigating how to solve problems by using it. In the next section, different usage of different aspects of PTM will be introduced by taking several related studies. At last, the review will be summarized and our assessment will be proposed.

2. Body of the review

2.1 LDA used in topic extraction

“The social network is a term which is in use to represent a large group of activities implementing the web and mobile technologies(Dolatabadi, Soon, Shirazi, & Mohammadi,

2012) .” Dolatabadi et al. (2012) set up a research about clustering Twitter users using Twitter - Latent Dirichlet Allocation(LDA) methodology. LDA is one of the topic modeling algorithms which is used to extract the topics from electronic documents. Twitter-LDA is a modified methodology for topic extraction based on LDA algorithm which is also an unsupervised machine learning method to identify latent topics from large document collections. The authors aim to cluster Twitter users by the tweets which include latent information, which can find groups of users based on similarities on their shared interests. They take a number of Malaysian Twitter users as a sample to find the similar topics in these users and cluster the users in different groups according to their interested topics. Park, Lee, Jung, and Lee (2012) also do their researches by using a probabilistic generative model based on LDA. Their research is about richness evaluation of Blogs on its topics. The difference is that they did not directly apply LDA on the blogs. They introduce a model to evaluate the richness for given keywords by applying LDA to calculate the probability distribution of each given keyword. They achieve their goal of richness evaluation by calculating with overall probabilities. LDA is applied to a set of given background documents about a keyword. The background documents come from a background corpus which consists of documents covered most of the content about the given topic. They evaluate the richness of blogs on its topic keywords and identify the important keywords by the richness. These two previous researches show the feasibility of applying probabilistic topic model to analyze social information. LDA is a statistical method to identify latent semantics from observable data from given context. It is often used in Information Retrieval (IR) area for topical analysis of specific corpus.

Wu, Wang, Ding, Xu, and Guo (2010) present a web news recommendation method which is topic based and combining Affinity Propagation (AP) with Latent Dirichlet Allocation (LDA). LDA could automatically find the topics exist in the web pages and recommend the topic based news to Internet users. The topic distance is defined using LDA, which is used to generate the topic distance matrix. AP clustering is used to cluster the web page collections into different topic clusters. In this paper, the researchers propose a topic based web news recommendation method combining AP and LDA. In order to prove the effect of combining AP and LDA, they sample web page collections with different topics and web pages from a web news corpus containing 30 topics and 5033 web pages. A series of experiments are implemented on these web page collections. A comparison of clustering results of AP with information distance and AP

with LDA are presented. The experiments show that our method combining AP and LDA is effective in topic based news recommendation system.

2.2 pLSA used in ontologies learning

Wang, Barnaghi, and Bargiela (2010) propose a new approach for automatic learning of terminological ontologies from text corpus. They present two algorithms for learning terminological ontologies using the principle of topic relationship and exploiting information theory with the probabilistic topic models learned. The probabilistic topic models have been primarily used in document modeling, topic extraction, and classification(Hofmann, 1999). In their research, Latent Semantic Analysis (LSA) is introduced to overcome the shortcomings of conventional IR. Parameter estimation behind pLSA is based on the likelihood principle and is approximated using Expectation-Maximization algorithm. The expectation step computes the posterior probabilities of the latent variables given the current estimate of the parameters(Wang et al., 2010). The methodology behind the LSA is to transform document representation in high-dimension word space to low-dimension semantic space to capture implicit structure in the association of terms with documents. Hofmann (1999) proposes Probabilistic Latent Semantic Analysis (pLSA) for analyzing co-occurrence of data. The pLSA is differentiated from LDA for its incompleting in that it provides no probabilistic model at the level of documents. The contribution of the research of Wang et al. (2010) is the development of a new method for learning terminological ontologies with both pLSA and LDA. These topic models are used for the purpose of ontology learning. Ontology learning is a joint research area of the semantic Web and knowledge acquisition. The idea of pLSA is also a widely used methodology on organizing topics into terminological ontologies. The results of their study show that ontologies learned using the LDA are superior to those learned using the pLSA in terms of recall.

2.3 Other utilizes of PTM

Jayagopi and Gatica-Perez (2010) argue that how to automatically classify conversational actions is a significant issue in today's social communication. They introduce a methodology to handle the classification by defining bag of group-nonverbal-patterns (NVPs) which is a novel descriptor and defined to briefly observe the connections and actions in the group, finding themes through employing principled probabilistic topic models. This introduced bag of group NVPs is able to blend every single piece of information and fosters the final comparison among

groups with an unfixed dimension. With the help of topic models, the group connections and actions can be clustered and the difference among them can be quantified probabilistic conceptually. It is helpful to comprehend and model connections between the members and a group as an intact unit by descripting the features of smalls groups with nonverbal behavior. In the paper, an unsupervised discovery approach is proposed to automatically mine group behavior patterns in conversation, in a robust and data-driven way. Their work has presented a method of discovering conversational group behavior in a data-driven approach. The method used to describe group behavior by defining group descriptors and then mining them using topic models is promising, allowing for the possibility of learning models to analyze group behavior on large meeting corpora in an unsupervised way, and therefore saving a potentially huge annotation effort (compared to supervised approaches).

Griffiths, Steyvers, and Tenenbaum (2007) analyze an abstract problem underlying the extraction and use of the kernel in a sentence or document, forming this problem as a rational statistical inference. In order to process language, the retrieval of knowledge in answer to an unremitting stream of information is required. This retrieval is facilitated if one can reason the kernel of a sentence or document. Moreover, the kernel can be employed to predict relevant concepts and make clear understanding of words. Consequentially, this leads to a new approach of semantic representation which represents word meanings in terms of a set of probabilistic topics. It is demonstrated that the topic model performs well in reasoning word association and the effects of semantic association and ambiguity on a variety of language-processing and memory tasks. It also provides a foundation for developing more richly structured statistical models of language, as the generative process assumed in the topic model can easily be extended to incorporate other kinds of semantic and syntactic structure. Identify the latent semantic structure to generate a set of words is part of learning and using language. By implementing probabilistic generative models, it is possible to use powerful statistical learning to infer structured representations. The topic model performs better than LSA, a major model for acquiring semantic knowledge, in predicting word association and other linguistic processing. The structured representation of the model that it assumes is the key to its success on these tasks. The model can identify their different meanings and senses through expressing the meaning of words in terms of different topics. Beyond the topic model, generative models provide a path toward a more comprehensive exploration of the role of structured representations and statistical learning in the acquisition and

application of semantic knowledge. The researchers have sketched some of the ways in which the topic model can be extended to bring it closer to capturing the richness of human language.

Watanabe, Iwata, Hori, Sako, and Ariki (2010) focuses on changes in the language environment, and applies a topic tracking model to language model adaptation for speech recognition and topic word extraction for meeting analysis. The topic tracking model can adaptively track changes in topics based on current text information and previously estimated topic models in an online manner. The effectiveness of the proposed method is shown experimentally by the improvement in speech recognition performance achieved with the Corpus of Spontaneous Japanese and by providing appropriate topic information in an automatic meeting analyzer. Their paper introduces the Topic Tracking Model, and extended it to a Topic Tracking Language Model for application to the unsupervised incremental adaptation of language models.

Interpretation and conclusion

In this review we present previous researches which aim to extract valuable information from specific data set. PTM is widely used in information engineering and the studies adapts PTM to measure an external variable of interest, a difficult task for unsupervised learning that must be carefully validated. The information extraction problem is well addressed by teaming computer scientists with other scholars to use topic models to help explore, visualize, and draw hypotheses from their data. By working with scholars in diverse fields, PTM can be developed as a new interdisciplinary computational methodology for working with and drawing conclusions from archives of texts.

References

Dolatabadi, Hossein, Soon, Lay-Ki, Shirazi, Mahdi Negahi, & Mohammadi, Mohammad. (2012). *Clustering Users in Micro Blogging Social Networks Using Probabilistic Topic Modeling-A Framework.* Paper presented at the Computational Science and Its Applications (ICCSA), 2012 12th International Conference on.

Griffiths, Thomas L, Steyvers, Mark, & Tenenbaum, Joshua B. (2007). Topics in semantic representation. *Psychological review, 114*(2), 211.

Hofmann, Thomas. (1999). *Probabilistic latent semantic analysis.* Paper presented at the Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence.

Jayagopi, Dinesh Babu, & Gatica-Perez, Daniel. (2010). Mining group nonverbal conversational patterns using probabilistic topic models. *Multimedia, IEEE Transactions on, 12*(8), 790-802.

Park, Jinhee, Lee, Jaedong, Jung, Hye-Wuk, & Lee, Jee-Hyong. (2012). *Richness evaluation of blogs on its topics using a generative model and probabilistic analysis.* Paper presented at the Soft

Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on.

Wang, Wei, Barnaghi, Payam Mamaani, & Bargiela, Andrzej. (2010). Probabilistic topic models for learning terminological ontologies. *Knowledge and Data Engineering, IEEE Transactions on, 22*(7), 1028-1040.

Watanabe, Shinji, Iwata, Tomoharu, Hori, Takaaki, Sako, Atsushi, & Ariki, Yasuo. (2010). *Application of topic tracking model to language model adaptation and meeting analysis.* Paper presented at the Spoken Language Technology Workshop (SLT), 2010 IEEE.

Wu, Yong-Hui, Wang, Xiao-Long, Ding, Yu-Xin, Xu, Jun, & Guo, Hong-Zhi. (2010). Adaptive on-line web topic detection method for web news recommendation system. *Dianzi Xuebao(Acta Electronica Sinica), 38*(11), 2620-2624.