Survey on Term Weighting Using Coherent Clustering in Topic Modeling

Manisha N. Amnerkar[1], Ashwini Tikle[2]
[1]*M-Tech Student, Department of Computer Science & Engineering, Wainganga Collage of Engineering & Management, Nagpur, India*
[1]*Assistant Professor, Department of Computer Science & Engineering, Wainganga Collage of Engineering & Management, Nagpur, India*

***Abstract*- Topic models often produce uncountable topics that are filled with noisy words. The reason is that words in topic modelling have same weights. More frequency words dominate the top topic word lists, but most of them are meaningless words, e.g., domain-specific stopwords. To address this issue, in this paper we aim to investigate how to weight words, and then develop a straightforward but effective term weighting scheme, namely entropy weighting (EW). The proposed EW scheme is based on conditional entropy measured by word co-occurrences. Compared with existing term weighting schemes, the highlight of EW is that it can automatically reward informative words. For more robust word weight, we further suggest a integrated form of EW (CEW) with two existing weighting schemes. Basically, our CEW assigns unmeaning words lower weights and informative words higher weights, leading to more coherent topics during topic modelling inference. We apply CEW to DMM and LDA, and evaluate it by topic quality, document clustering and classification tasks on 8 real world data sets. Exploratory results show that weighting words can effectively improve the topic modelling performance over both short texts and normal long texts. More importantly, the proposed CEW significantly outperforms the existing term weighting schemes, since it further considers which words are informative.**

Index Terms- Topic modeling, Term weighting, Informative word, Conditional entropy

1. INTRODUCTION

With the rapid increase in the amount of electronic information on internet web pages and modern applications, text analysis in the domain of text mining requires complex techniques to deal with numerous text documents. Topic models are widely used to uncover the latent semantic structure from text corpus. The effort of mining the semantic structure in a text collection can be dated from latent semantic analysis (LSA) [17], which employs the singular value decomposition to project documents into a lower dimensional space, called latent semantic space. Probabilistic latent semantic analysis (PLSA) [6] improves LSA with a sound probabilistic model based on a mixture decomposition derived from a latent class model. In PLSA, a document is represented as a mixture of topics, while a topic is a probability distribution over words. Extending PLSA, Latent Dirichlet Allocation (LDA)[7] adds Dirichlet priors for the document-specific topic mixtures, making it possible to generate unseen documents. Due to its nice generalization ability and extensibility, LDA has achieved huge success in text mining. In the last decade, topic models have been extensively studied. Many complicated variants and extensions of the standard LDA model have been proposed, which can be found in the comprehensive survey [18]. Here we only list some work closely related to us. Wallach [19] proposed the bigram topic model extending LDA by incorporating bigram statistics into topic modeling, but its detail is quite different from ours. The bigram topic model aims to capture ordinal dependencies between words (in normal texts) by exploiting document-level sequential patterns, while our model is designed specifically for short texts and tries to capture the semantic dependencies between words by exploiting corpus-level word co-occurrence patterns. Besides, two recently proposed models, i.e., the regularized topic model [20] and the generalized P_olya model [21], share the same idea of utilizing word co-occurrence (i.e., biterm) statistics to enhance topic learning. However, both of them only use word co-

occurrence information as prior to guide the generation of words, rather than directly modeling the co-occurrences.

The basic assumption of topic modeling is that there exists a latent topic level beyond the observable word level, where each topic is a multinomial distribution over the vocabulary. Given topics learnt by topic models, we can deeply explore text documents in a variety of tasks, such as sentiment analysis [6, 7] and classification applications. The past decade has witnessed an explosive development of topic modeling algorithms[2]. Topic models, such as Dirichlet multinomial mixture (DMM) and latent Dirichlet allocation (LDA)[4], are nowadays tools for text analysis. They can effectively uncover the hidden structures of short text and normal long texts.

2. LITERATURE REVIEW

Text categorization (TC) is the task of automatically classifying unlabelled natural language documents into a predefined set of semantic categories. As the first and a vital step, text representation converts the content of a textual document into a compact format so that the document can be recognized and classified by a computer or a classifier. Recently, the study of term weighting methods for TC has gained increasing attention. In contrast to Information retrieval (IR), TC is a supervised learning task as it makes use of prior information on the membership of training documents in predefined categories. This known information is effective and has been widely used for the feature selection [1] and the construction of text classifier to improve the performance of the system. In this study, we group the term weighting methods into two categories according to whether the method involves this prior information, i.e., supervised term weighting method (if it uses this known membership information) and unsupervised term weighting method (if it does not use this information).The present study builds on earlier research that has examined issues of dimensionality reduction in information retrieval environments.

3. PROPOSED METHODOLOGY

The goal of this paper is to investigate a novel term weighing scheme that can automatically reward informative words. To improve the issues mentioned above, a straightforward way is to weight words, i.e., term weighting, during topic inference. Basically, one designs term weighting schemes following two principles:

Principle I: assigning meaningless words, e.g., domain-specific stopwords, lower weights;

Principle II: assigning informative words higher weights.

To achieve this goal, we develop an entropy-based term weighting scheme of topic modeling using information theory, namely entropy weighting (EW). We suppose that a word is more important if it has more influence to the occurrence of any other word, and then quantize this “influence” using conditional entropy values computed by word co-occurrences. Following this mechanism, EW will prefer assigning informative words higher weights. We further combine the EW weight with two existing weighting schemes (i.e., the log and BDC weights), and then obtain a more robust combination weight (CEW). The CEW can simultaneously meet the Principle I and Principle II, assigning meaningless words lower weights and informative words higher weights. In this work, we apply CEW to DMM and LDA topic modeling for short texts and normal long texts, respectively. To evaluate the proposed CEW with the help of following below mentioned flow diagram.

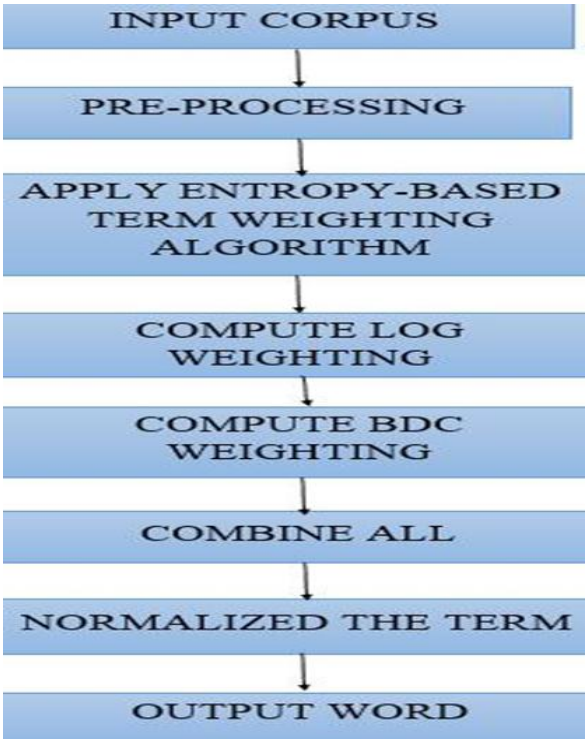

Fig: Flow Diagram

The contributions of this project are summarized as follows:

1. We develop a novel entropy-based term weighting scheme for topic models, namely EW. The highlight of EW is that it can automatically reward informative words.
2. We further combine the EW weight with two existing weighing schemes, leading to a more robust combination weighting scheme, namely CEW.
3. We apply CEW to two topic models, i.e., DMM and LDA, which are used for modeling short texts and normal long texts, respectively.
4. The empirical results show that CEW significantly outperforms the existing term weighting baselines on topic quality, document clustering and classification tasks.

4. CONCLUSION

We present a novel term weighting scheme to improve topic modeling, namely CEW. The CEW weight is a combination weight of three types of word weights, i.e., EW, log, and BDC weights. The EW weight is used to reward informative words that frequently co-occur with a small set of semantically related words, and the log and BDC weights are used to punish the words that are frequently occurring and likely to scatter most of topics. We conducted a number of experiments to evaluate CEW on topic quality and document clustering. Experimental results indicate that CEW outperforms the existing term weighting schemes of topic modeling on both short texts and normal long texts.

This paper gives an empirical implication that weighting words can (straightforwardly) improve topic models, where all term weighting schemes improve the basic models in some degree. Our CEW uses word co-occurrences to detect informative words, and it performs the best in most settings. This implies us that the word co-occurrence information is helpful for scoring words, and more fortunately, it is model-independent. That is, one can also use word co-occurrence based term weighting for a wider range of algorithms.

REFERENCES

[1] Andrzejewski, D., Zhu, X., & Craven, M. (2009). Incorporating domain knowledge into topic modeling via Dirichlet forest priors. International conference on machine learning25–32.

[2] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.

[3] Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of scinece. The Annals of Applied Statistics, 1(1), 17–35.

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3,993–1022.

[5] Bougouin, A., Boudin, F., & Daille, B. (2013). Topicrank: Graph-based topic ranking for keyphrase extraction. International joint conference on natural language processing5430–5551.

[6] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. Neural information processing systems288–296.

[7] Chen, E., Lin, Y., Xiong, H., Luo, Q., & Ma, H. (2011). Exploiting probabilistic topic models to improve text categorization under class imbalance. Information Processing and Management, 47(2), 202–214.

[8] mChen, Z., Mukherjee, A., & Liu, B. (2014). Aspect extraction with automated prior knowledge learning. Annual meeting of the association for computational linguistics347–358.

[9] Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Discovering coherent topics using general knowledge. International conference on information and knowledge management209–218.

[10] Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 26(12), 2928–2941.

[11] Chew, P. A., Bader, B. W., Helmreich, S., & Abdelali, A. (2011). An information-theoretic, vector-space-model approach to cross-language information retrieval. Journal of Natural Language Engineering, 17(1), 37–70.

[12] Graph-based term weighting scheme for topic modeling (2016). International conference on

data mining workshops. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. The National Acsdemy of Sciences volume, 101, 5228–5235.

[13] Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. Annual meeting of the association for computational linguistics1262–1273.

[14] Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. Annual meeting of the association for computational linguistics216–223.

[15] Jiang, X., Hu, Y., & Li, H. (2009). A ranking approach to keyphrase extraction. Internatioal acm sigir conference on research and development in information retrieval756–757.

[16] K. R. Canini, L. Shi, and T. L. Griffiths, "Online inference of topics with latent Dirichlet allocation," in Proc. Int. Conf. Artif. Intell. Statist., 2009, pp. 65–72.