Information Processing and Management xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

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

Exploring coherent topics by topic modeling with term weighting

Ximing Li*[a,b], Ang Zhang[a,b], Changchun Li[a,b], Jihong Ouyang[a,b], Yi Cai[c]

[a] College of Computer Science and Technology, Jilin University, China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[c] School Of Software Engineering, South China University of Technology, China

ARTICLE INFO

Keywords:
Topic modeling
Term weighting
Informative word
Conditional entropy

ABSTRACT

Topic models often produce unexplainable topics that are filled with noisy words. The reason is that words in topic modeling have equal weights. High 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 combination form of EW (CEW) with two existing weighting schemes. Basically, our CEW assigns meaningless words lower weights and informative words higher weights, leading to more coherent topics during topic modeling inference. We apply CEW to Dirichlet multinomial mixture and latent Dirichlet allocation, and evaluate it by topic quality, document clustering and classification tasks on 8 real world data sets. Experimental results show that weighting words can effectively improve the topic modeling 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.

1. Introduction

The past decade has witnessed an explosive development of topic modeling algorithms (Blei, 2012). Topic models, such as Dirichlet multinomial mixture (DMM) (Nigam, Mccallum, Thrun, & Mitchell, 2000) and latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003), are nowadays tools for text analysis. They can effectively uncover the hidden structures of short texts (Cheng, Yan, Lan, & Guo, 2014; Zuo, Zhao, & Xu, 2016) and normal long texts (Blei & Lafferty, 2007; Xie & Xing, 2013).

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 (Lin & He, 2009; Lu, Ott, Cardie, & Tsou, 2011) and classification applications (Chen, Lin, Xiong, Luo, & Ma, 2011; Yang, Gao, Tan, & Wong, 2013). However, they often produce low-quality topics that are unexplainable (Mimno, Wallach, Talley, Leenders, & McCallum, 2011). To briefly explain such topics, we list five example topics learnt by LDA across the *Newsgroup*[1] data set in Table 1. It can be seen that two issues exist. First, all top word lists contain the word “article”. We thus conclude that the word “article” is much less discriminative among different topics. We call this kind of general word domain-

* Corresponding author.
E-mail address: liximing86@gmail.com (X. Li).

[1] *Newsgroup* is a commonly used training corpus. Statistics of *Newsgroup* are shown in the experiment section.

https://doi.org/10.1016/j.ipm.2018.05.009
Received 17 October 2017; Received in revised form 8 April 2018; Accepted 20 May 2018

Table 1
The top 5 words of example topics.

Topic	Top 5 words
1	nasa **article** space research satellite
2	jews war jewish world **article**
3	food **article** good *eat taste*
4	gun guns fbi fire **article**
5	software graphics **article** ftp data

specific stopword. Second, the third topic is about the subject of food. The words “eat” and “taste” are obviously more semantically associated with this subject, however, they rank behind the more general words such as “article” and “good”.

The underlying cause of the two issues is that with statistical inference algorithms, topic inference is dominated by high frequency words in some degree. First, high frequency words are more likely to be the top words in topics. Second, a semantics-specific word, i.e., informative word, may be underestimated if it is relatively less frequently occurring.

To alleviate 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 our knowledge, there are some previous term weighting schemes of topic modeling (Chew, Bader, Helmreich, & Abdelali, 2011; Malliaros & Skianis, 2015; Wilson & Chew, 2010; Yang, Cai, Chen, fung Leung, & Lau, 2016). For example, the log weighting LDA (Wilson & Chew, 2010) computes the word occurrence weight following the axiom of information theory, punishing the high frequency words; and the balanced distributional concentration (BDC) weighting LDA (Yang et al., 2016) concerns whether a word tends to scatter across most of topics. However, the existing models mainly focus on **Principle I**. How to effectively detect informative words using limited information, i.e., meeting **Principle II**, is still challenging.

Table 2 presents the words with highest weights, measured by the log weighting and BDC weighting schemes across the *Newsgroup* data set. We observe that most of those words are informative by no means. These examples imply that the issue of underestimating the true informative words remains unchanged.

Motivation and contribution. The goal of this paper is to investigate a novel term weighing scheme that can automatically reward informative words, i.e., **Principle II**.

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. The example shown in Table 2 presents that EW can effectively find informative words to some extent.

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, we conducted a number of experiments on 8 real world data sets, including 4 short text collections and 4 normal long text collections. Experimental results indicate that our CEW can produce more coherent topics than the existing term weighting schemes. Besides, it can significantly improve the document clustering and classification performance of topic models.

The contributions of this paper are summarized as follows:

- 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.

Table 2
The example words with highest weight.

Weighting scheme	Words with highest weight
Log	hens xmas finou zaurak rescorla isolar ids brevity permutations esac
BDC	mov db compass geode cosmo bi angmar een cols bh
EW	padding median islander capita imply px makeup snail households couples