The 2015 Conference on Computational Linguistics and Speech Processing
ROCLING 2015, pp. 164-166
© The Association for Computational Linguistics and Chinese Language Processing

Word Co-occurrence Augmented Topic Model in Short Text

陳冠斌　Guan-Bin Chen
國立成功大學資訊工程學系
Department of Computer Science and Information Engineering
National Cheng Kung University
gbchen@ikmlab.csie.ncku.edu.tw

高宏宇　Hung-Yu Kao
國立成功大學資訊工程學系
Department of Computer Science and Information Engineering
National Cheng Kung University
hykao@mail.ncku.edu.tw

摘要

在網際網路上，大量的文字使得人們難以在有限的短時間內加以吸收並了解，主題模型（如 pLSA 與 LDA）被提出來試圖對這些長文件做摘要與總結成幾個代表性的主題字。近年來，隨著社群網路的興起（如 Twitter），使得短文件的數量也隨之變大，在為數眾多的短文本中如何良好地做摘要與整理也變成一大課題，因而有了應用主題模型於短文本的想法。然而直接應用主題模型到這些短文本上，由於短文本中字數不足以用來良好地統計該主題的字詞共現特性，所以經常會得到一些相干度低的主題。根據我們回顧的文獻，雙詞主題模型（Bi-term topic model, BTM）透過整個資料集中的雙詞（Bi-term），直接對字詞共現特性做建模，能有效改善單一文件中字數不足的問題。然而 BTM 於統計過程中只考慮雙詞的共現頻率，導致產生的主題很容易會被單一高頻字所主導。

本研究提出基於字詞共現性的主題模型來改善 BTM 中主題被高頻字所主導的問題。對於 BTM 的問題，我們提出的 PMI-β-BTM 方法導入點對點交互資訊（pointwise mutual information, PMI）分數於其主題字的事前機率分布中，來降低單一高頻字的影響。實驗結果顯示，我們的 PMI-β-BTM 無論是在正規的新聞標題上或是在雜訊高的 tweet 上皆有較好的主題性。另外，我們所提出的方法不需修改原始主題模型，因此可直接應用於 BTM 的衍生模型上。

關鍵詞：短文本，主題模型，文件分類，文件分群

Keywords: Short Text, Topic Model, Document Clustering, Document Classification.

Topic models learn topics base on the amount of the word co-occurrence in the documents. The word co-occurrence is a degree which describes how often the two words appear together. BTM, discovers topics from bi-terms in the whole corpus to overcome the lack of local word co-occurrence information. However, BTM will make the common words be performed excessively because BTM identifies the word co-occurrence information by the bi-term frequency in corpus-level. Thus, we propose a PMI-β priors methods on BTM. Our PMI-β priors method can adjust the co-occurrence score to prevent the common words problem. Next, we will describe the detail of our method of PMI-β priors.

However, just consider the frequency of bi-term in corpus-level will generate the topics which contain too many common words. To solve this problem, we consider the Pointwise Mutual Information (PMI) [9]. Since the PMI score not only considers the co-occurrence frequency of the two words, but also normalizes by the single word frequency. Thus, we want to apply PMI score in the original BTM. A suitable way to apply PMI scores is modifying the priors in the BTM. The reason is that the priors modifying will not increase the complexity in the generation model and very intuitive. Clearly, there are two kinds of priors in BTM which are β-prior and β-priors. The β-prior is a corpus-topic bias without the data. While the β-priors are topic-word biases without the data. Applying the PMI score to the β-priors is the only one choice because we can adjust the degree of the word co-occurrence by modifying the distributions in the β-priors. For example, we assume that a topic contains three words "pen", "apple" and "banana". In the symmetric priors, we set <0.1, 0.1, 0.1> which means no bias of these three words, while we can apply <0.1, 0.5, 0.5> to enhance the word co-occurrence of "apple" and "banana". Thus the topic will prefer to put the "apple" and "banana" together in the topic sampling step.

Table 1 shows the clustering results on the Twitter2011 dataset, when we set the number of topic to 50. As expected, BTM is better than Mixture of unigram and LDA got the worst result when we adopt the symmetric priors <0.1>. When apply the PMI-β priors, we get the better result than BTM with symmetric priors. Otherwise, our baseline method, PCA-β, is better than the original LDA because the PCA-β prior can make up the lack of the global word co-occurrence information in the original LDA.

Table 1. The Clustering Results on Twitter2011 dataset

Model	β priors	Purity	NMI	RI
LDA	<0.100>	0.4174	0.3217	0.9127
	PCA-β	0.4348	0.3325	0.9266
Mix	<0.100>	0.4217	0.3358	0.8687
	PCA-β	0.3748	0.3305	0.7550
BTM	<0.100>	0.4318	0.3429	0.9092
	PCA-β	0.4367	0.4000	0.8665
	PMI-β	0.4427	0.3927	0.9284

In this paper, we propose a solution for topic model to enhance the amount of the word co-occurrence relation in the short text corpus. First, we find the BTM identifies the word co-occurrence by considering the bi-term frequency in the corpus-level. BTM will make the common words be performed excessively because the frequency of bi-term comes from the whole corpus instead of a short document. We propose a PMI-β priors method to overcome this problem. The experimental results show our PMI-β-BTM get the best results in the regular short news title text.

Acknowledgement

The work reported in this paper was partially supported by the National NEP-II Project MOST 104-3113-F-260-001, 2015.

References

[1] T. Hofmann, "Probabilistic latent semantic analysis," in Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 289-296, 1999.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[3] M. Divya, K. Thendral, and S. Chitrakala, "A Survey on Topic Modeling," International Journal of Recent Advances in Engineering & Technology (IJRAET), vol. 1, pp. 57-61, 2013.

[4] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262-272, 2011.

[5] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro, Brazil, pp. 1445-1456, 2013.

[6] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," Machine learning, vol. 39, pp. 103-134, 2000.

[7] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, et al., "Comparing twitter and traditional media using topic models," in Advances in Information Retrieval, ed: Springer, pp. 338-349, 2011.

[8] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: Topic Modeling over Short Texts," Knowledge and Data Engineering, IEEE Transactions on, vol. 26, pp. 2928-2941, 2014.

[9] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," Computational linguistics, vol. 16, pp. 22-29, 1990.

[10] H. M. Wallach, D. Minmo, and A. McCallum, "Rethinking LDA: Why priors matter," 2009.