A Non-Parametric Topic Model for Short Texts Incorporating Word Coherence Knowledge

Yuhao Zhang[1] Wenji Mao[1,2] Daniel Zeng[1,2]

[1]State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Computer and Control Engineering, University of Chinese Academy of Sciences
{zhangyuhao2012, wenji.mao, dajun.zeng}@ia.ac.cn

ABSTRACT

Mining topics in short texts (e.g. tweets, instant messages) can help people grasp essential information and understand key contents, and is widely used in many applications related to social media and text analysis. The sparsity and noise of short texts often restrict the performance of traditional topic models like LDA. Recently proposed Biterm Topic Model (BTM) which models word co-occurrence patterns directly, is revealed effective for topic detection in short texts. However, BTM has two main drawbacks. It needs to manually specify topic number, which is difficult to accurately determine when facing new corpora. Besides, BTM assumes that two words in same term should belong to the same topic, which is often too strong as it does not differentiate two types of words (i.e. general words and topical words). To tackle these problems, in this paper, we propose a non-parametric topic model npCTM with the above distinction. Our model incorporates the Chinese restaurant process (CRP) into the BTM model to determine topic number automatically. Our model also distinguishes general words from topical words by jointly considering the distribution of these two word types for each word as well as word coherence information as prior knowledge. We carry out experimental studies on real-world twitter dataset. The results demonstrate the effectiveness of our method to discover coherent topics compared with the baseline methods.

Keywords

Text Mining; Topic Model; Bayesian Nonparametric Model.

1. INTRODUCTION

Short texts such as tweets, instant messages and advertisements, are widespread in Web environment. Mining topics in these texts can help us understand the key contents implied and facilitate many applications like user profiling [8], recommendation [1, 10], information diffusion and influence analysis [13, 16] and so on. It is more challenging to mining topics in short texts than that in normal long documents as the data is sparse and usually noisy as well. Traditional topic models like PLSI [5], LDA [2] and HDP [6] represent documents as a mixtures of topics and capture the document-level word co-occurrence patterns to reveal topics. These methods often suffer from data sparsity when facing short texts.

CIKM'16 , *October 24-28, 2016, Indianapolis, IN, USA*

DOI: http://dx.doi.org/10.1145/2983323.2983898

Several models have been proposed specifically for mining topics in short texts. Traditional methods [9, 16] adopt the "aggregation strategies" which combine short texts together as pseudo documents and then use conventional topic model like LDA for topic detection. These strategies make conventional topic models perform better than using them directly in short texts. As these methods did not really model the short texts, their performances vary with datasets and aggregation strategies. To model short texts, non-negative matrix factorization is used to exploit global word co-occurrence, which is proved useful in practice [18]. The Biterm Topic Model (BTM) [17] further extends it to a more principle approach by modeling the generative process of word co-occurrence patterns in corpus, which avoids the problem of data sparsity at document-level. The BTM model performs well in both short and normal texts. There are also some other models designed specifically for tweets. Twitter-LDA [19] and Twitter-BTM [4] incorporate user information into a LDA variation and the BTM respectively. However, all the above models need to specify the topic number manually, which is difficult to accurately determine when facing new corpora.

As one of the state of the art methods focusing on general-domain short texts modeling, BTM lacks the ability to differentiate general words and topical words, which may affect the detection of true topical words. The distinction is both important and effective in practice for topic modeling [4, 19]. Twitter-BTM attempts to address this issue by considering user's preference between the two word types in twitter environment. However, it models the two words within same biterm separately without considering their coherence information. Twitter-LDA only differentiates the two word types at the corpus level, and lacks a more fine-grained distinction of them for different words. Besides, both Twiter-LDA and Twitter-BTM are not applicable to topic detection in short texts without user information.

To tackle the above issues in topic mining for short texts, in this paper we propose a non-parametric coherent topic model (npCTM) that differentiates general words and topical words at the word level and incorporates word coherence information as prior knowledge for differentiation. On the basis of the BTM model, we first incorporate the Chinese restaurant process (CRP) [7] to determine topic number automatically in our model. As different words may prefer different word types, we directly model the distribution of the word types for each word. We also incorporate the word coherence information at the corpus level using point mutual information (PMI). We then distinguish general words from topical words by jointly considering the distribution of the two word types for each word and using word coherence information as prior knowledge. Experimental results on real world twitter dataset demonstrate that our method performs better in discovering coherent topics than the baseline methods.

2. Proposed Method

In this section, we first give a brief introduction of the non-parametric statistical method CRP. Then we propose how to acquire word coherence knowledge from corpus. Finally, we present our non-parametric topic model npCTM and the parameters inference of the model.

2.1 Chinese Restaurant Process

The Chinese restaurant process is a distribution on partitions of integers. The CRP works according to the following metaphor: imagine a sequence of customers entering a restaurant with infinite tables. When the *n*th costumer enters, she sits at an occupied table with probability proportional to the number of previous customers sitting there, and at an unoccupied table with probability proportional to α. The cluster assignments under the CRP distribution are exchangeable, which means the probability of a particular seating configuration does not depend on the customers' arriving order.

Let z_i be the index of the table where the ith customer sits and z_{-i} be all the other customers' seats. Let m_k denotes the number of customers sitting at table k and m be the total number of customers. Let K be the number of tables where $m_k > 0$ and α be the parameter of CRP. The probability of customer i sitting at each table is represented as follows:

$$p(z_i = k \mid z_{-i}, \alpha) = \begin{cases} \dfrac{m_k}{m-1+\alpha} & k \le K \\ \dfrac{\alpha}{m-1+\alpha} & k = K+1 \end{cases} \quad (1)$$

2.2 Acquiring Word Coherence Knowledge

We follow the idea of BTM [17] which models the word co-occurrence patterns directly by turning original texts into biterms, and represents biterms as unordered word-pairs co-occurring in the same context. In addition, to differentiate general words (G) and topical words (T) in biterms, we extract word coherence information from corpus, use this information as prior knowledge and incorporate it into our model.

A document containing n words will be converted into C_n^2 biterms. There are three types of biterms G-G, G-T (or T-G), and T-T in our method. Intuitively, the representative words of each topic tend to co-occur more frequently, indicating that biterms of the T-T type could have higher coherence values than those of the G-T, T-G and G-G types. For two words w_a and w_b in a biterm, we use PMI as the coherence measure and compute their PMI value from the corpus as follows:

$$PMI(w_a, w_b) = \log \frac{P(w_a, w_b)}{P(w_a)P(w_b)} \quad (2)$$

where $P(w_a, w_b)$ is the probability of w_a and w_b co-occurring in the corpus, and $P(w_a)$ and $P(w_b)$ are probabilities of w_a and w_b occurring in the corpus respectively.

Then we use sigmoid function to convert the PMI of words within each biterm into a probability value which reflects the word coherence in our method. For biterm b_i containing words w_a and w_b, this probability η_{b_i} can be acquired as follows:

$$\eta_{b_i} = \frac{1}{1+e^{-PMI(w_a,w_b)}} = \frac{1}{1+\dfrac{P(w_a)P(w_b)}{P(w_a,w_b)}} \quad (3)$$

2.3 The npCTM Model

We now present our non-parametric topic model npCTM. Our method uses CRP as a non-parametric prior to determine the topic number. To distinguish general words from topical words, our model captures both the global coherence information of a biterm and the distributions of the word type information within the biterm. For each biterm b_i, we consider both the probability η_{b_i} and the distribution of the word types for each word in b_i. Each word in b_i either belongs to a certain topic or the background topic reflecting the distribution of general words. Specifically, given a biterm b_i ,we choose its topic z_i based on CRP. As η_{b_i} reflects the coherence of words within b_i and higher value of η_{b_i} indicates b_i is more likely to belong to the T-T type. So with probability η_{b_i} we set the topic of the words in b_i to z_i. With probability $1-\eta_{b_i}$, we determine the word types of words in b_i separately. We use the word type distribution $Bernoulli(\mu_{w_{i,j}})$ to determine whether a word in b_i belong to the topic z_i or to the background topic. The generative process of our model is as follows:

1. Draw background word distribution $\phi_B \sim Dirichlet(\beta)$
2. For each word w, draw $\mu_w \sim Beta(\lambda_1, \lambda_2)$
3. For each biterm b_i , $i \in [1, \mathrm{N}]$
   a) Draw topic $z_i \sim CRP(\alpha)$
   b) If z_i is a new topic,
      - draw topic word distribution $\phi_{z_i} \sim Dirichlet(\beta)$
   c) Draw $\rho_i \sim Bernoulli(\eta_{b_i})$
   d) If $\rho_i = 1$,
      - draw $w_{i,1}$, $w_{i,2} \sim Multinomial(\phi_{z_i})$

      If $\rho_i = 0$,
      - For each word $w_{i,j}$, $j = 1, 2$
        i. $\kappa_{w_{i,j}} \sim Bernoulli(\mu_{w_{i,j}})$
        ii. If $\kappa_{w_{i,j}} = 1$, draw $w_{i,j} \sim Multinomial(\phi_{z_i})$

            If $\kappa_{w_{i,j}} = 0$, draw $w_{i,j} \sim Multinomial(\phi_B)$

2.4 Parameters Inference

We adopt collapsed Gibbs Sampling, a widely used inference method based on Markov chain Monte Carlo algorithm, to estimate the parameters of our model. Gibbs sampling cycles through the variables and estimates them alternatively. Each time the method replaces the value of one variable by a value drawn from the distribution of that variable conditioned on the values of the remaining variables. As conjugate prior is used in our model, we adopt collapsed Gibbs sampling to integrate out some variables which can simplify the sampling procedure.

In our model, we need to sample topics $z = \{z_1, \ldots, z_N\}$ for each biterm $b_1, \ldots, b_N$. The topic-word distributions ϕ and word type parameter μ can be updated based on the sampling results. To perform Gibbs sampling, we first initialize these variables for the Markov chain randomly. For simplicity, let Θ be the variables except z and m_{z_i} be the number of biterms belonging to z_i. We can sample z as follows:

$$P(z_i = t \mid b_i, z_{-i}, \Theta) \propto \begin{cases} m_{z_i} P(b_i \mid z_i, z_{-i}, \Theta) & (t \text{ is an existing topic}) \\ \alpha P(b_i \mid z_i, z_{-i}, \Theta) & (t \text{ is a new topic}) \end{cases} \quad (4)$$

Based on the generative process of our model, $P(b_i \mid z_i, z_{-i}, \Theta)$ can be split into four parts, with each part listing in one line below:

$$P(b_i \mid z_i, z_{-i}, \Theta) \propto \eta_{b_i}\phi_{w_{i,1}|z_i}\phi_{w_{i,2}|z_i} + (1-\eta_{b_i})\mu_{w_{i,1}}\phi_{w_{i,1}|z_i}\mu_{w_{i,2}}\phi_{w_{i,2}|z_i} + (1-\eta_{b_i})\mu_{w_{i,1}}\phi_{w_{i,1}|z_i}(1-\mu_{w_{i,2}})\phi_{w_{i,2}|B} + (1-\eta_{b_i})(1-\mu_{w_{i,1}})\phi_{w_{i,1}|B}\mu_{w_{i,2}}\phi_{w_{i,2}|z_i} + (1-\eta_{b_i})(1-\mu_{w_{i,1}})\phi_{w_{i,1}|B}(1-\mu_{w_{i,2}})\phi_{w_{i,2}|B} \quad (5)$$

The four parts correspond to the four word type combinations T-T, T-G, G-T and G-G respectively. To simplify the computation, the above formula (5) can be represented as:

$$P(b_i \mid z_i, z_{-i}, \Theta) \propto \{\eta_{b_i}\phi_{w_{i,1}|z_i}\phi_{w_{i,2}|z_i} + (1-\eta_{b_i})[\mu_{w_{i,1}}\phi_{w_{i,1}|z_i} + (1-\mu_{w_{i,1}})\phi_{w_{i,1}|B}][\mu_{w_{i,2}}\phi_{w_{i,2}|z_i} + (1-\mu_{w_{i,2}})\phi_{w_{i,2}|B}]\} \quad (6)$$

So we can sample the corresponding topic of each biterm according to formula (4) and (6). Then based on the values of the four parts in formula (5), we can sample the word type information for each biterm. With the word type information and the topic information of each biterm, we can update ϕ and μ as follows:

$$\phi_{w_{i,j}|z_i} = \frac{n_{w_{i,j}|z_i} + \beta}{\sum_w n_{w_{i,j}|z_i} + V\beta} \quad (7)$$

$$\phi_{w_{i,j}|B} = \frac{n_{w_{i,j}|B} + \beta}{\sum_w n_{w_{i,j}|B} + V\beta} \quad (8)$$

$$\mu_{w_{i,j}} = \frac{l_{w_{i,j},1} + \lambda_1}{l_{w_{i,j},1} + l_{w_{i,j},2} + \lambda_1 + \lambda_2} \quad (9)$$

where $n_{w_{i,j}|z_i}$ and $n_{w_{i,j}|B}$ denote the numbers of $w_{i,j}$ in topic z_i and background B respectively, and V denotes the vocabulary size, and $l_{w_{i,j},1}$ and $l_{w_{i,j},2}$ denote the numbers of $w_{i,j}$ belong to the topical word type and general word type respectively.

3. EXPERIMENTS

We evaluate the performance of our proposed model for short texts.by comparing it with the baseline methods under the typical topic coherence measures.

3.1 Experimental Setup

Dataset. We use a public twitter dataset collected in June 2011 through twitter API [14]. We only keep English tweets. We remove stop words as well as words with extremely high frequency, since similar to stop words, these words will influence the performance of topic models. We also remove words occurring less than 20 times as they can hardly represent the topic information. We then filter out tweets with one or two words and convert the letters of all tweets into lower case. We randomly sample 300,000 tweets used for training topic models and use 3,101,271 tweets left as reference dataset for the topic coherence evaluation task .

Baseline Methods. We compare our npCTM model with four baseline methods. They are LDA, BTM, HDP, and npCTM-UB. LDA is a classic topic model which has been widely used. BTM is a representative topic mining method for short texts. HDP is a widely used non-parametric topic model based on Dirichlet process. We also compare our model with npCTM-UB, which is a simplified version of our npCTM model without the background topic.

Parameter Setting. In our model, we set the parameter values $\alpha = 10$, $\beta = 1$ and $\lambda_1 = \lambda_2 = 1$. For npCTM-UB, we use $\alpha = 10$ and $\beta = 1$. The initial topic number of both models is set to 300. We choose a relative large initial topic number and the model will change it to a proper value by the sampling process. As LDA and BTM could not determine the topic number automatically, we choose three topic numbers for them. Their topic numbers K are set to 20, 50 and 80 standing for relatively small, medium and large topic numbers respectively. For LDA, we set $\alpha = 1/K$ and $\beta = 1/K$. For BTM, we use the settings in [17], that is $\alpha = 50/K$ and $\beta = 0.01$. For each model above, we run Gibbs sampling 300 times. We use the HDP model with a variational inference algorithm. We set the first level concentration $\alpha = 1$ and the truncation number K=150. We set the second level concentration $\beta = 1$ the second level truncation T = 15.

3.2 Evaluation of Topic Quality

A traditional way to evaluate topic models is comparing the perplexity or marginal likelihood on a held-out test [2, 15]. Recent study [3] shows that these evaluation metrics could not test the interpretability of topics very well. Besides, our model optimizes the likelihood of word occurrences directly, which is different from models like LDA. Recent work [11, 12] shows that topic coherence measures could assess topic quality very well. Thus we evaluate the quality of topics t based on PMI and LCP as mentioned in [12].

$$PMI(t) = \sum_{j=2}^{N}\sum_{i=1}^{j-1}\log\frac{P(w_i, w_j)}{P(w_j)P(w_i)} \quad (10)$$

$$LCP(t) = \sum_{j=2}^{N}\sum_{i=1}^{j-1}\log\frac{P(w_i, w_j)}{P(w_i)} \quad (11)$$

We evaluate our method and the baseline methods using the average PMI score and LCP score of the detected topics. For each topic, we use the top L words for topic coherence evaluation. In our experiment, L is set to 5 and 10. After the sampling process, we omit topics with less than 10 biterms, as these topics contain too limited information to form any meaningful topics in practice.

Our model finally discovers 53 topics in this experiment. The results on the comparison of topic coherence for each method are shown in Table 1. We can see that both LDA and HDP did not perform well because these two models were not designed for short texts and their model performances were affected by the sparsity of data. BTM performs better, but its topic coherence score varies with the predefined topic numbers. We can see that our npCTM-UB performs slightly better than BTM. Our npCTM model achieves the best results among all the models under the PMI measure and LCP measure when L=5, demonstrating the advantage of differentiating general and topical information. Generally, the experimental results demonstrate that our model performs better than the baseline methods and produces more coherent topics. The results also reveal that our model can automatically determine a proper topic number through the sampling process.

Table 2 illustrates the top 10 words of the football topic produced by our model and BTM. The non-representative topic words are marked bold. The example shows that our model derives relatively more coherent topics than BTM with different settings.

4. CONCLUSIONS

Mining topics in short texts is an important technique for information retrieval and text processing, and has been widely used in many applications. Existing methods either suffer from the data sparsity problem or need to specify a fixed topic number in advance. In this paper, we propose a non-parametric topic model

Table 1. Results on the comparison of topic coherence for each method

Method		LDA				BTM				HDP	npCTM -UB	npCTM
Topic Number		20	50	80	Average	20	50	80	Average	150	44	53
PMI	L=5	-5.74	-4.97	-5.76	-5.49	4.01	6.07	3.93	4.67	-3.77	6.54	**9.65**
	L=10	-32.22	-36.85	-43.10	-37.39	6.39	13.85	9.67	9.97	-24.48	14.46	**20.61**
LCP	L=5	-43.88	-46.24	-47.79	-45.97	-35.68	-36.42	-37.66	-36.59	-40.68	-34.84	**-33.49**
	L=10	-269.64	-288.75	-301.65	-286.68	-235.67	-236.74	-240.02	-237.48	-246.32	**-233.82**	-234.48

Table 2. Illustration of top 10 words of the football topic

BTM	npCTM
football season win team play state nfl college **open top** (k=20) football team nfl win season play sports fans **fantasy top** (k=50) football season nfl team college play state win sports **fantasy** (k=80)	football season team college nfl sports fans beat tickets games

which can distinguish general words from topical words by jointly considering the distribution of these two word types for each word as well as using word coherence information as prior knowledge. We carry out empirical studies on real-world twitter dataset. The experimental results show that our model can automatically determine a proper topic number and discover coherent topics more effectively compared with the baseline methods.

5. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation of China under Grant No. 71621002, No. 91224008 and No. 71472175.

6. REFERENCES

[1] Belém, F., Santos, R., Almeida, J. and Gonçalves, M. 2013. Topic Diversity in Tag Recommendation. *Proceedings of the 7th ACM Conference on Recommender Systems*, 141–148.

[2] Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*. 3, 993–1022.

[3] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J.L. and Blei, D.M. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, 288–296.

[4] Chen, W., Wang, J., Zhang, Y., Yan, H. and Li, X. 2015. User Based Aggregation for Biterm Topic Model. 489–494.

[5] Ding, C., Li, T. and Peng, W. 2008. On the Equivalence Between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing. *Comput. Stat. Data Anal.* 52, 8, 3913–3927.

[6] Gao, Z., Song, Y., Liu, S., Wang, H., Wei, H., Chen, Y. and Cui, W. 2011. Tracking and Connecting Topics via Incremental Hierarchical Dirichlet Processes. *2011 IEEE 11th International Conference on Data Mining*, 1056–1061.

[7] Gershman, S.J. and Blei, D.M. 2012. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56, 1, 1–12.

[8] Harvey, M., Crestani, F. and Carman, M.J. 2013. Building user profiles from topic models for personalised search. *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, 2309–2314.

[9] Hong, L. and Davison, B.D. 2010. Empirical study of topic modeling in twitter. *Proceedings of the First Workshop on Social Media Analytics*, 80–88.

[10] Hu, B. and Ester, M. 2013. Spatial Topic Modeling in Online Social Media for Location Recommendation. *Proceedings of the 7th ACM Conference on Recommender Systems*, 25–32.

[11] Lau, J.H., Baldwin, T, and Newman, D. 2013. On Collocations and Topic Models. *ACM Transactions on Speech and Language Processing*, 10, no. 3 (2013): 10:1–10:14.

[12] Lau, J.H., Newman, D. and Baldwin, T. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539.

[13] Li, D., Shuai, X., Sun, G., Tang, J., Ding, Y. and Luo, Z. 2012. Mining topic-level opinion influence in microblog. *Proceedings of the 21st ACM international conference on Information and knowledge management*, 1562–1566.

[14] Li, R., Wang, S. and Chang, K.C.-C. Towards Social Data Platform: Automatic Topic-focused Monitor for Twitter Stream. *Proc. VLDB Endow.* 6, 14 (2013), 1966–1977.

[15] Wallach, H.M., Murray, I., Salakhutdinov, R. and Mimno, D. 2009. Evaluation Methods for Topic Models. *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–1112.

[16] Weng, J., Lim, E.-P., Jiang, J. and He, Q. 2010. TwitterRank: finding topic-sensitive influential twitters. *Proceedings of the third ACM international conference on Web search and data mining*, 261–270.

[17] Yan, X., Guo, J., Lan, Y. and Cheng, X. 2013. A Biterm Topic Model for Short Texts. *Proceedings of the 22Nd International Conference on World Wide Web*, 1445–1456.

[18] Yan, X., Guo, J., Liu, S., Cheng, X. and Wang, Y. 2013. Learning Topics in Short Texts by Non-negative Matrix Factorization on Term Correlation Matrix. *Proceedings of the 2013 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics. 749–757.

[19] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X. 2011. Comparing Twitter and Traditional Media Using Topic Models. *Advances in Information Retrieval*, 338–349.