Topic Modeling of Chinese Language Using Character-Word Relations

Qi Zhao[1], Zengchang Qin[1,2], and Tao Wan[2]

[1] Intelligent Computing and Machine Learning Lab
School of Automation Science and Electrical Engineering
Beihang University, Beijing, China
[2] Robotics Institute, Carnegie Mellon University, USA
zhaoqi1861@gmail.com, zcqin@buaa.edu.cn

Abstract. Topic models are hierarchical Bayesian models for language modeling and document analysis. It has been well-used and achieved a lot of success in modeling English documents. However, unlike English and the majority of alphabetic languages, the basic structural unit of Chinese language is character instead of word, and Chinese words are written without spaces between them. Most previous research of using topic models for Chinese documents did not take the Chinese character-word relationship into consideration and simply take the Chinese word as the basic term of documents. In this paper, we propose a novel model to consider the character-word relation into topic modeling by placing an asymmetric prior on the topic-word distribution of the standard Latent Dirichlet Allocation (LDA) model. Compared to LDA, this model can improve performance in document classification especially when test data contains considerable number of Chinese words not appeared in training data.

Keywords: Topic Models, Latent Dirichlet Allocation, CWTM, Gibbs Sampler.

1 Introduction

Topic models are a class of hierarchical probabilistic models for analyzing discrete data collections. It assumes that documents are mixtures of topics and each topic is a probability distribution over words. Topic models have attracted a lot of attentions in recent years because it tries to model document in semantic level. Unlike English and the majority of alphabetic languages, the basic structural unit of Chinese language is character instead of word [13], and Chinese words are written without spaces between them. Most previous research applying topic models to analyze Chinese documents choose Chinese word as the basic term. Chinese documents are segmented into words which are generally believed to have more specific meanings than characters. However, word-based methods completely ignore the information that a Chinese word are composed of Chinese characters. Words sharing one same character may have some semantic relations, such a relation cannot be detected in word-based models. For example, the Chinese words "*xué xí* "(study) and "*xué shēng* " (student) are literally related by sharing the same character "*xué*", and they are also semantically related in meaning and may have a high probability to occur in the same context. However, in word-based computational models, these two words are treated as two distinct words has no relations at all.

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part III, LNCS 7064, pp. 139–147, 2011.

The number of commonly used Chinese characters[1] is around 3000, while the size of word vocabulary can be way larger with new words created constantly. However, the characters constitute the new word probably already appeared in the history documents. In the standard Latent Dirichlet Allocation [3] model, all those words unseen in training data are assigned the equal probability in a topic by simply placing a symmetric dirichlet prior[8] on the topic-word distribution, regardless of their component characters having different occurrence rates in the training data. Hence, we will extend LDA by incorporating Chinese character-word relation to improve the performance of topic model when modeling Chinese documents.

This paper is structured as follows, the new proposed generative model is introduced in details in Section 2. We apply Gibbs sampling [1,5] method to inference the model in Section 3. Empirical results are given to evaluate this model in Section 4. The conclusions and future work is discussed in the end.

2 Character-Word Structure in Topic Modeling

2.1 Generative Model

To encode the character-word relation into topic model, we extend the standard LDA by placing an asymmetric prior on the topic-word distribution. This prior is obtained according to the character-word relation and topic-character distribution. The graphical model is shown in Fig. 1. Besides, we refer to this extended LDA model as *Character-Word Topic Model* (CWTM) in this paper.

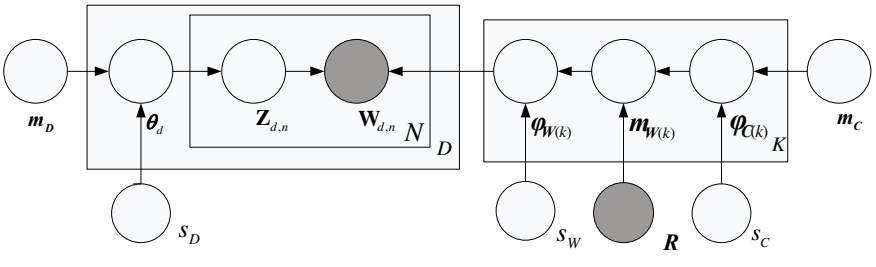

Fig. 1. Graphical model representation of Chinese character-word topic model

As illustrated in the Fig. 1, the document-topic modeling (the left part) is exactly the same as standard LDA [2,3]. $\mathbf{W}_{d,n}$ represents an observable Chinese word where d ($d = 1, \ldots, D$) is the index of document and n ($n = 1, \ldots N_d$) is the index of words in the document d. $\mathbf{Z}_{d,n}$ is a K-dimensional multinomial random variable indicate which topic is assigned to $\mathbf{W}_{d,n}$. Each document d is a mixture of topics parameterized by θ_d

[1] Previous research show that 3500 most commonly used characters can cover 99.48% of a corpus of written materials with over 2 million characters. 1000 most used character can cover 90% of a daily life corpus with over a billion characters [16].

which has a Dirichlet prior with hyperparameters m_D, s_D. The Dirichlet distribution is denote by $Dir(\mathbf{m}, s)$ in this paper, where $\mathbf{m}$ is the normalized *mean* ($\sum_i m_i = 1$) and s is called precision parameter (a scalar) that $\mathbf{m} * s$ is equivalent to α in standard Dirichelet distribution [8].

The CWTM model assumes that documents in a corpus $\mathbb{C}$ with Chinese word vocabulary size V_W and character vocabulary size V_C are generated by the following process:

1. For document $d = 1, \ldots, D$:
   (a) Draw a distribution over topics $\theta_d \sim \mathrm{Dir}(\mathbf{m}_D, s_D)$.
   (b) For each Chinese words $n = 1, \ldots, N_d$, in document d:
      (i) Draw a topic assignment $\mathbf{Z}_{d,n} \sim \mathrm{Mult}(\theta_d)$.
      (ii) Draw a word $\mathbf{W}_{d,n} \sim \mathrm{Mult}(\varphi_{\mathbf{W}(\mathbf{Z}_{d,n})})$.
2. For each topic $k = 1, \ldots, K$:
   (a) Draw a topic distribution over Chinese characters $\varphi_{\mathbf{C}(k)} \sim \mathrm{Dir}(\mathbf{m}_C, s_C)$.
   (b) Infer the distribution over Chinese words based on a deterministic function of the distribution of Chinese characters $\varphi_{C(k)}$ and character-word relationship $\mathbf{R}$: $\mathbf{m}_{\mathbf{W}(k)}=F(\varphi_{\mathbf{C}(k)}, \mathbf{R})$.
   (c) Draw a distribution over Chinese words $\varphi_{\mathbf{W}(k)} \sim \mathrm{Dir}(\mathbf{m}_{\mathbf{W}(k)}, s_W)$.

$\mathbf{R}$ is a 2-dimensional matrix containing Chinese character-word composition relationship, it is defined as the count of character (C_j) contained in word (W_i):

$$\mathbf{R}_{ij} = Count(W_i, C_j) \quad i = 1, 2, ..., V_W \quad j = 1, 2, ..., V_C \tag{1}$$

where V_W and V_C represent the size of Chinese word vocabulary and Chinese character vocabulary, respectively. We define the deterministic function F as:

$$\mathbf{m}_{\mathbf{W}(k,i)} = F(\varphi_{C(k)}, \mathbf{R})_i = N_k(\prod_{j=1}^{V_C} \varphi_{C(k,j)}{}^{R_{ij}})^{\frac{1}{\prod_{j=1}^{V_C} R_{ij}}} \quad i = 1, 2, ..., V_W \tag{2}$$

where N_k is a constant to ensure all the coordinates of $\mathbf{m}_{\mathbf{W}(k)}$ sums to 1 and the character-word relation can be used in other form by simply altering the definition of function F.

2.2 Likelihood

The major difference between this model and LDA is the prior of the topic-word distribution $\varphi_{\mathbf{W}}$. In LDA, all the K distributions $\varphi_{\mathbf{W}(1)}, \varphi_{\mathbf{W}(2)}, ..., \varphi_{\mathbf{W}(K)}$ own a common symmetric dirichlet prior. While CWTW impose each topic-word distribution $\varphi_{\mathbf{W}k}$ with a unique asymmetric Dirichlet prior parameterized by $\mathbf{m}_{\mathbf{W}(k)}$ and s_W[11]. This means that each $\varphi_{\mathbf{W}(k)}$ has a corresponding prior $Dir(\mathbf{m}_{\mathbf{W}(k)}, s_W)$. And the mean of this Dirichlet prior $\mathbf{m}_{\mathbf{W}}$ is obtained by a deterministic function F, which takes character-word relation $\mathbf{R}$ and topic-character distribution $\varphi_{\mathbf{C}(k)}$ as inputs. Owing to the deterministic characteristic of generating $\varphi_{\mathbf{W}(k)}$ from R and $\varphi_{\mathbf{C}(k)}$, K topic-character distributions $\varphi_{\mathbf{C}(1)}, \varphi_{\mathbf{C}(2)}, ..., \varphi_{\mathbf{C}(K)}$ should be different to ensure that topic-word distributions $\varphi_{\mathbf{W}(1:K)}$ own priors with different mean parameters $\mathbf{m}_{\mathbf{W}(1:K)}$. Therefore, the character-word relation is incorporated as prior, which could make it less sensitive to errors caused by character-word relation, rather than is hardcoded into the model[9]. And the balance

between the prior originated from character-word relation and observed data could be controlled by adjusting the Dirichlet precision parameter s_W.

According to the description in Section 2.1, the joint distribution of all the variables given hyperparameters is:

$$p(\mathbf{W}_d, \mathbf{Z}_d, \theta_d, \varphi_{\mathbf{W}}, \varphi_{\mathbf{C}} | \mathbf{m}_D, s_D, \mathbf{m}_{\mathbf{C}}, s_C, s_W, R) = \prod_{n=1}^{N_d} p(\mathbf{W}_{d,n} | \varphi_{\mathbf{W}(Z_{d,n})}) p(Z_{d,n} | \theta_d) \tag{3}$$

$$\cdot p(\theta_d | \mathbf{m}_D, s_D) \cdot \prod_{k=1}^{K} p(\varphi_{\mathbf{W}(k)} | s_W, \mathbf{m}_{\mathbf{W}(k)}) p(\mathbf{m}_{\mathbf{W}(k)} | \varphi_{C(k)}, \mathbf{R}) p(\varphi_{C_{(k)}} | s_C, \mathbf{m}_C)$$

where $p(\mathbf{m}_{\mathbf{W}(k)} | \varphi_{C(k)}, \mathbf{R}) = 1$ only when $\mathbf{m}_{\mathbf{W}(k)} = F(\varphi_{C(k)}, \mathbf{R})$ and 0 for all the other values of $\mathbf{m}_{\mathbf{W}(k)}$ because F is a deterministic function.

3 Inference

Several methods have been proposed to do inference in LDA-like topic models [2,3]. In this paper, we will use Gibbs sampling [5], which is a special form of Markov chain Monte Carlo [1,4] for CWTM inference.

3.1 Document Likelihood

As we can see from the graphical model illustrated in Fig. 1, the joint distribution $p(\mathbf{W}, \mathbf{Z} | \mathbf{m}_D, s_D, \mathbf{m}_{\mathbf{W}}, s_W)$ could be factored as follows:

$$p(\mathbf{W}, \mathbf{Z} | \mathbf{m}_D, s_D, \mathbf{m}_{\mathbf{W}}, s_W) = p(\mathbf{W} | \mathbf{Z}, \mathbf{m}_{\mathbf{W}}, s_W) p(\mathbf{Z} | \mathbf{m}_D, s_D) \tag{4}$$

And the two terms on the right side of above equation can be obtained by canceling out $\varphi_{\mathbf{W}}$ and θ respectively:

$$\begin{aligned} p(\mathbf{W} | \mathbf{Z}, \mathbf{m}_{\mathbf{W}}, s_W) &= \int p(\mathbf{W} | \mathbf{Z}, \varphi_{\mathbf{W}}) p(\varphi_{\mathbf{W}} | \mathbf{m}_{\mathbf{W}}, s_W) d\mathbf{m}_{\mathbf{W}} \\ &= \prod_{k=1}^{K} \frac{\Gamma(s_W)}{\Gamma(s_W + n_k^T)} \prod_{i}^{V_W} \frac{\Gamma(n_{k,i}^{TW} + s_W * \mathbf{m}_{\mathbf{W}(k,i)})}{\Gamma(s_W * \mathbf{m}_{\mathbf{W}(k,i)})} \end{aligned} \tag{5}$$

$$\begin{aligned} p(\mathbf{Z} | \mathbf{m}_{\mathbf{D}}, s_D) &= \int p(\mathbf{Z} | \theta) p(\theta | \mathbf{m}_{\mathbf{D}}, s_D) d\theta \\ &= \prod_{d=1}^{N} \frac{\Gamma(s_D)}{\Gamma(s_D + n_d^D)} \prod_{k=1}^{K} \frac{\Gamma(n_{d,k}^{DT} + s_D * \mathbf{m}_{D(k)})}{\Gamma(s_D * \mathbf{m}_{D(k)})} \end{aligned} \tag{6}$$

where n^{TW} is a matrices stored the counts of the number of times each Chinese word is assigned to each topic, say $n_{k,i}^{TW}$ denotes the count number of Chinese word indexed by i is assigned to topic k. Similarly, $n_{d,k}^{DT}$, element of n^{DT}, denotes the count number of Chinese words which are assigned to topic k in document d. And $n_k^T = \sum_{i=1}^{V_W} n_{k,i}^{TW}$, $n_d^D = \sum_{k=1}^{K} n_{d,k}^{DT}$.

3.2 Collapsed Sampler

To perform Gibbs sampling, we need a sampler $p(\mathbf{Z}_{d,n} = k|\mathbf{Z}_{d,n-}, \mathbf{W})$ from which new samples are drawn. This sampler is similar to that of standard LDA in [5]. The derivation of standard LDA's inference algorithm based on Gibbs sampling is introduced with details in [6].

$$p(\mathbf{Z}_{d,n} = k|\mathbf{Z}_{d,n-}, \mathbf{W}) \propto \frac{[n^{TW}_{k,v-} + \mathbf{m}_{\mathbf{W}(k,v)}][n^{DT}_{d,k-} + s_d * \mathbf{m}_{D(k)}]}{\sum_{i=1}^{V_W} n^{TW}_{k,i-} + s_W} \tag{7}$$

Noting $\mathbf{W} = \{\mathbf{W}_{d,n} = v, \mathbf{W}_{d,n-}\}$, the subscript ' $-$ ' represents current token indexed by (d, n) is not taken into consideration. Besides, $n^{TW}_{k,i-}$ denotes the count number of Chinese word indexed by i is assigned to topic k.

Since this Gibbs sampling method directly estimate $\mathbf{Z}_{d,n}$ for each word in the corpus, topic-word distributions and document-topic distributions can be obtained by:

$$\varphi_{\mathbf{W}(k,i)} = \frac{n^{TW}_{k,i} + s_W * \mathbf{m}_{\mathbf{W}(k,t)}}{\sum_{i=1}^{V_W} n^{TW}_{k,i} + s_W} \tag{8}$$

$$\theta_{d,k} = \frac{n^{DT}_{d,k} + s_D * m_{D(k)}}{\sum_{k=1}^{K} n^{DT}_{d,k} + s_D} \tag{9}$$

3.3 Estimate Topics over Chinese Characters

The Gibbs sampler proposed in Section 3.2 assumes $\mathbf{m}_{\mathbf{W}}$ as known, which is converted from topics over Chinese character through a deterministic process by the function $F(R, \varphi_{\mathbf{C}})$. Therefore it is necessary to estimate topics over Chinese characters $\varphi_{\mathbf{C}}$.

To approximately estimate $\varphi_{\mathbf{C}}$, we assume each Chinese character contained in the Chinese word token $\mathbf{W}_{d,n}$ in the corpus is directly sampled from the corresponding character-topic Multinomial distribution:

$$C^{(j)}_{d,n} \sim Mult(\varphi_{C(\mathbf{Z}_{d,n})}) \tag{10}$$

where j denotes the index of the character in the word token $\mathbf{W}_{d,n}$.

This means all characters in a word shared the identical topic, which is the same as the word's topic assignment. That is to say, if Chinese word "*jì suàn*"($\mathbf{W}_{d,n}$) is assigned with topic k, then its component characters "*jì*"($C^{(1)}_{d,n}$) and "*suàn*"($C^{(2)}_{d,n}$) are assigned the same topic k. This makes sense because characters in a word are inclined to express more related meanings.

Thus, the topic-character distribution could be approximately estimated as:

$$\varphi_{\mathbf{C}(k,j)} = \frac{n^{TC}_{k,j} + s_C * \mathbf{m}_{C(k,j)}}{\sum_{j=1}^{V_C} n^{TC}_{k,j} + s_C} \tag{11}$$

where $n^{TC}_{k,j}$ denotes the count number of Chinese character indexed by j is assigned to topic k.

The hyperparameters for CWTM are: $\mathbf{m}_D, s_D, \mathbf{m}_C, s_C, s_W$, which are either Dirichlet mean parameter or Dirichlet precision parameter. As with the standard LDA model, we use symmetric priors in the latter experiment section. That is to say, we set $\mathbf{m}_D$ and $\mathbf{m}_C$ as uniform distributions. For the other Dirichlet precision parameter, Griffiths have given a setting that show good quality in [5]. So we use that setting to fix $s_D = 50$, $s_W = V_W * 0.1$ and $s_C = V_C * 0.1$ for all k topics.

4 Experimental Studies

4.1 Data Descriptions

The Chinese corpus used in this paper is a news archive for classification provided by the Sogou laboratory[2]. The documents in this corpus are news articles collected from the website of the Sohu.com, which is one of the China's biggest Internet media company. These news articles are manually edited and classified into 10 classes that cover military, education, tourism and other topics. In the original corpus, each class contains 8000 documents. But few of these documents contain nothing but some meaningless non-Chinese symbols. Therefore those documents contain less than 100 Chinese characters will be ignored in our research. We then selected 1000 documents with 100 documents per class from the corpus as our experiment data which is named as NEWS1K and it is open to public online[3]. To take Chinese word as basic unit of topic model, we split Chinese documents into word tokens by using a Chinese word segmentation tool ICTCLAS-09[4] in our experiments. We also removed rare terms that appears less than 10 times across the whole corpus. For those terms that appears in over 50% of the documents, we consider them as stop words and remove them from the corpus as well. After the above preprocessing steps, the corpus NEWS1K contains 21389 unique Chinese words and 3631 unique Chinese characters. We then split NEWS1K into training and test set by the rate 1 : 1. There are 3747 Chinese words only appeared in the test set, but are unseen in training set.

4.2 Extracted Topics and Document Classification

The implementation of CWTM is based on GibbsLDA++[5]. Then we separately apply the standard LDA and CWTM to analyze NEWS1K's training data with topic number varies from 10 to 100. Fig. 2(a) shows top 20 terms of two topics independently extracted from CWTM and standard LDA when the number of topics is 50. As we can see from the results, topics extracted from these 2 models are relevant to sports.

As we mentioned in Section 1, for the words not appeared in training data, characters composing these words have appeared as component of other words in the training data. CWTM takes the character-word relation into consideration. Therefore, as is illustrated

[2] The corpus can be obtained at: `http://www.sogou.com/labs/dl/c.html`

[3] Dataset is available at: `http://icmll.buaa.edu.cn/projects/topicmodel/index.html`

[4] ICTCLAS is an integrate Chinese lexical analysis system which provides a tool for Chinese word segmentation. `http://ictclas.org/`

[5] C/C++ Implementation of LDA. `http://gibbslda.sourceforge.net/`

in Fig. 2(b), some of the words relevant to sports, even not appeared in the training data, were assigned relatively higher probabilities in topics about sports by CWTM. For example, the word "*zhǔ kè chǎng*"(home and away), though never appeared in training documents, its component characters "*zhǔ*", "*kè*", and "*chǎng*" are the components of those words appeared in training data. Therefore, it obtained a higher probability in CWTM. On the other hand, not surprisingly, none of these words were ranked into top 1000 words in the corresponding topic extracted by standard LDA model. Similar situations as described above happened with different number of topics.

Term Rank	CWTM (Topic 6)	Standard LDA (Topic 45)
1	队	队
2	比赛	比赛
3	场	场
4	球	分钟
5	球队	两
6	但	球
7	下	球队
8	胜	出
9	本	但
10	轮	下
11	主场	本
12	足	联赛
13	队	轮
14	比	被
15	队员	次
16	表现	表现
17	联赛	号
18	对手	主场
19	球迷	胜
20	冠军	球迷

(a) Top 20 words of extracted topics from CWTM and the standard LDA when the number of topics is 50.

Term Rank	CWTM (Topic 6)
450	主客场
453	德国队
468	大胜
477	发球
480	国奥队
532	接球
536	对攻
580	女足
592	赛道
783	国奥
859	必胜
863	落选
933	亚运会

(b) Some words, though not appeared in training data, ranked in top 1000 of the 6th topic extracted from CWTM with topic number equals to 50.

Fig. 2. Samples of extracted topics by using two different topic models: LDA and CWTM

The above extracted topics indicates that our method to incorporating character-word relation make the topics more reasonable. Then we make experiments to see its performance in document classification compared to standard LDA[12]. When modeling documents with topic model, each document can be represented as a distribution over topics by parameter θ. If two documents are semantically similar, it means the distance between two corresponding topic distributions is small, too. By using topic distributions, we can quantitatively measure the semantic (dis)similarities between documents. In such a way, documents of each class can be mapped to a K-dimensional space where K is the number of topics. In such a space, we can use discriminative machine learning algorithms, such as Neural Network (NN), k-Nearest Neighbor (kNN) or Support Vector Machine (SVM), to classify these documents based on the *semantic distance* measure in terms of dissimilarity between topic distributions θ. In particular, we employ the SVM to the classification task on the NEWS1K dataset for its effectiveness in text classifications such as question classification [10,7]. We adopted libsvm[6] with its default RBF kernel in our experiments.

[6] LIBSVM is an integrated software for support vector classification, regression and distribution estimation. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

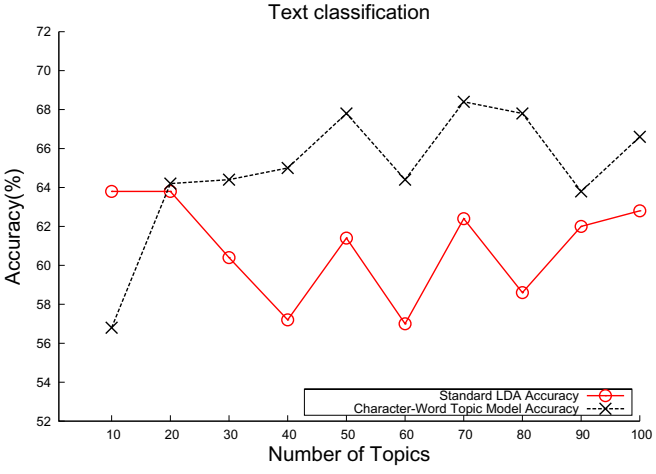

Fig. 3. Text classification accuracy based on standard LDA and CWTM

The classification accuracy of standard LDA and CWTM on the test data with different number of topics k is shown in Fig. 3. CWTM have a higher classification accuracy at all numbers of topic except when number of topic equals to 10. This may be caused by the too much errors brought to the topic-word distribution $\varphi_{\mathbf{W}}$ by the character-word relation prior when the number of topics is very small. Generally, our model have a better performance in document classification when a considerable number of words appeared only in test data while not in training data.

5 Conclusions

In this paper, we propose a method to incorporate character-word relation into the topic model by placing an asymmetric prior on the topic-word distribution of standard Latent Dirichlet Allocation (LDA) model. And experiments show, compared to LDA, CWTM can extract more reasonable topics and improve performance in document classification under certain circumstance. Besides, our proposed method can be easily applied to the most of the other topic models.

Though this method of encoding Chinese character-word relation could improve the performances of topic model, the errors brought to the model by this method prevent it from performing much better in modeling Chinese documents. This may be improved by altering the definition of the function $F(\varphi_{\mathbf{C}(k)}, \mathbf{R})$ or a more sophisticated model structure.

Acknowledgment. This work is partially funded by the NCET Program of MOE, China and the SRF for ROCS. The second author also thanks the China Scholar Council for the visiting fellowship (No. 2010307502) to CMU.

References

1. Bishop, M.C.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
2. Blei, D.M., Lafferty, J.D.: Topic Models. In: Srivastava, A., Sahami, M. (eds.) Text Mining: Classification, Clustering, and Applications. Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC (2009)
3. Blei, D.M., Ng, A., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
4. Gilks, W.R., Richardson, S., Spiegelhalter, D.J.: Markov Chain Monte Carlo in Practice. Chapman & Hall, NewYork (1996)
5. Griffiths, T.L., Steyvers, M.: Finding Scientific Topics. Proceedings of the National Academy of Science 101, 5228–5235 (2004)
6. Heinrich, G.: Parameter estimation for text analysis. Technical report, Fraunhofer IGD (2009)
7. Huang, Z., Thint, M., Qin, Z.: Question Classification using Head Words and their Hypernyms. In: Proceedings of EMNLP, pp. 927–936 (2008)
8. Minka, T.: Estimating a Dirichlet distribution (2000)
9. Petterson, J., Smola, A., Caetano, T., Buntine, W., Narayanamurthy, S.: Word Feature for Latent Dirichlet Allocation. In: Proceedings of Neural Information Processing Systems (2010)
10. Qin, Z., Thint, M., Huang, Z.: Ranking Answers by Hierarchical Topic Models. In: Chien, B.-C., Hong, T.-P., Chen, S.-M., Ali, M. (eds.) IEA/AIE 2009. LNCS, vol. 5579, pp. 103–112. Springer, Heidelberg (2009)
11. Wallach, H., Mimno, D., McCallum, A.: Rethinking LDA: Why Priors Matter. In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (2009)
12. Wu, Y., Ding, Y., Wang, X., Xu, J.: A comparative study of topic models for topic clustering of Chinese web news. In: Computer Science and Information Technology (ICCSIT), vol. 5, pp. 236–240 (2010)
13. Xu, T.Q.: Fundamental structural principles of Chinese semantic syntax in terms of Chinese Characters. Applied Linguistics 1, 3–13 (2001) (in Chinese)
14. Zhang, Y., Qin, Z.: A topic model of Observing Chinese Characters. In: Proceedings of the 2nd International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), pp. 7–10 (2010)
15. Zhao, Q., Qin, Z.: What is the Basic Semantic Unit of Chinese Language? A Computational Approach Based on Topic Models. In: Proceedings of Meeting on Mathematics of Language (MOL 2011), pp. 7–10 (to appear, 2011)
16. `http://en.wikipedia.org/wiki/Chinese_language`