Text Categorization Based on Topic Model

Shibin Zhou, Kan Li, and Yushu Liu

School of Computer Science and Technology
Beijing Institute of Technology, Beijing 100081, P.R. China
{guoguos.zhou, likan, liuyushu}@bit.edu.cn

Abstract. In the text literature, many topic models were proposed to represent documents and words as topics or latent topics in order to process text effectively and accurately. In this paper, we propose LDACLM or Latent Dirichlet Allocation Category Language Model for text categorization and estimate parameters of models by variational inference. As a variant of Latent Dirichlet Allocation Model, LDACLM regard documents of category as Language Model and use variational parameters to estimate maximum a posteriori of terms. Experiments show LDACLM model to be effective for text categorization, outperforming standard Naive Bayes and Rocchio method for text categorization.

Keywords: Latent Dirichlet Allocation, Variational Inference, Category Language Model.

1 Introduction

In the text analysis, standard algorithms are unsatisfactory because terms often were supposed independent, which was recognized as "bag of words" model. However, the "bag of words" model offers a rather impoverished representation of the data because it ignores any relationships between the terms.

In the recent past, a new class of generative models called Topic Model has quickly become more popular in some text-related tasks. Topic Model supposes documents and corpus composed of mixture topics and then documents can be thought of "bag of topics". Thus, these models can handle the problem effectively about terms dependency. Topics can be view as a probability distribution over words, where the distribution implies semantic coherence. For example, a topic related to fruit would have high probabilities for the words "orange", "apple", and even "juicy". Wallach [10] demonstrated the "bag of topics" to surpass in performance to "bag of words" in unigram and bigram schemas.

There are many Topic Models proposed by researchers in the past such as Latent Semantic Analysis or LSA [3], the probabilistic Latent Semantic Indexing or pLSI [6], Latent Dirichlet allocation or LDA [1] and so on.

Latent Semantic Analysis (LSA) [3] is an approach that combines both term and document clustering. LSA usually takes a term-document matrix in the vector space representation as input, and uses a singular value decomposition of the input matrix to identify a linear subspace in the space of tf-idf features

G. Wang et al. (Eds.): RSKT 2008, LNAI 5009, pp. 572–579, 2008.

that captures most of the variance in the collection. Thus LSA can map text elements to a representation in the latent semantic space and can capture some aspects of basic linguistic notions such as synonymy and polysemy.

The probabilistic Latent Semantic Indexing (pLSI) model introduced by Hofmann [6], also known as the aspect model, was designed as a discrete counterpart of LSI or LSA to provide a better fit to text data and to overcome deficiencies of Latent Semantic Indexing (LSI). pLSI is a latent variable model that models each document as a mixture of topics. Although there are some problems with the generative semantics of pLSI, Hoffmann has shown some encouraging results in Information Retrieval.

One of these models, Latent Dirichlet Allocation (LDA) has quickly become one of the most popular probabilistic text modeling techniques in Information Retrieval. LDA has been shown to be effective in some text-related tasks. Processing fully generative semantics, LDA overcomes the drawbacks of previous topic models such as probabilistic Latent Semantic Indexing (pLSI) which is a MAP/ML estimated LDA model under a uniform Dirichlet distribution according to Girolami and Kaban discovery [4]. Latent Dirichlet allocation represents documents as mixtures over latent topics differentiated with pLSI, which each topic is characterized by a distribution over words. In [11], Wei and Croft shown the LDA-based document model had good performance in Information Retrieval. Moveover, Griffiths and Steyvers [5] apply LDA model to find scientific document topics.

Our goal in this paper is to address a variants of LDA and a extension of Language Model [9], which is a novel model for text categorization as we known. This generative model represents words set of each category with a mixture of topics assumed independent, as in state-of-the-art approaches like Latent Dirichlet Allocation [1], and extends these approaches to estimate maximum a posteriori of category language model parameters by assuming that variance parameters would be multinomial and dirichlet parameters of category language model.

In Section 2, we demonstrate our approaches on how to estimate parameters of models and classify documents. In section 3, we evaluate accuracy of our model on Reuters21578 and 20Newsgroups datesets. We conclude the paper with a summary, and a brief discussion of future work in section 4.

2 Latent Dirichlet Allocation Category Language Model

In this section we introduce our model that extends Latent Dirichlet Allocation and Language Model called Latent Dirichlet Allocation Category Language Model and manifest methods of inferring and estimating parameters.

2.1 Model Structure

Latent Dirichelt Allocation Category Language Model or LDACLM is a variant of LDA, which is used as classifier of text documents. Rather, LDA descripted in [1] used as dimension reducer in the discriminative framework of documents

classification. The prominent feature of LDACLM is that the model assume each word would be a independent topic that we called word topic and assume extra topics other than word topics would be model the correlation among the words. As we known, this distinguish to LDA and also tradeoff between effective and time consuming. The following process similar to LDA generates documents in the LDACLM model.

- For each category language model or words set $\mathbf{w}$, pick multinomial distribution $p(\theta_{\mathbf{w}})$ from a symmetric Dirichlet distribution $p(\theta_{\mathbf{w}}|\alpha)$ with prior scalar parameter α which is identity to all category language models.
- Pick a topic $z \in \{1, 2, \ldots, K\}$ from a multinomial distribution $p(z|\theta_{\mathbf{w}})$ with parameter vector $\theta_{\mathbf{w}}$.
- Generate a word w_t from a multinomial distribution $p(w_t|z, \beta)$ with parameter vector β, where each parameter β_z in the vector β is related to specific z respectively.

2.2 Inference

The maximum likelihood of category language model $\mathbf{w}$ with model parameter vector β and model dirichlet parameter α may formulate as:

$$p(\mathbf{w}|\alpha, \beta) \propto \int \left(\prod_{k=1}^{K} \theta_k^{\alpha-1} \right) \left(\prod_{t=1}^{V} \left\{ \sum_{k=1}^{K} (\theta_k \beta_{k,t}) \right\}^{tf_{t,\mathbf{w}}} \right) \mathrm{d}\theta$$

Where words set $\mathbf{w}$ containing words form corpus $\mathcal{D}$ who has a vocabulary of size V and $tf_{t,\mathbf{w}}$ stores the number of occurrences of a word w_t in words set $\mathbf{w}$.

Similar to LDA [1], We develop a variational approximation [8] for LDACLM by defining an approximating family distribution $q(\theta, z|\mathbf{w}, \gamma, \phi)$, and choose the variational Dirichlet parameter vector γ and variational multinomial parameter vector ϕ which are different sets for each category language model to yield a tight approximation to the true posterior. Suppose the factorized variational parameters distribution is $q(\theta, z|\mathbf{w}, \gamma, \phi) = q(\theta|\mathbf{w}, \gamma) \prod_{t=1}^{V} q(z_t|\mathbf{w}, \phi_t)$ with variational Dirichlet parameter vector γ and variational multinomial parameter vector ϕ. Especially, for each category language model, there is a different set of Multinomial and Dirichlet variational parameter vectors. Thus, minimization of the KL divergence $D(q(\theta, z|\mathbf{w}, \gamma, \phi) \| p(\theta, z|\mathbf{w}, \alpha, \beta))$ we can derive approximation of $p(\theta, z|\mathbf{w}, \alpha, \beta)$.

So, we can take decreasing steps in the KL divergence and converge to optimizing parameter by an iterative fixed-point method, bounding the marginal likelihood of a document using Jensen's inequality [8].

$$\log p(\mathbf{w}|\alpha, \beta) \geq E_q \{\log p(\theta, z, \mathbf{w}|\alpha, \beta)\} - E_q \{\log q(\theta, z|\mathbf{w}, \gamma, \phi)\} \tag{1}$$

Letting $\mathcal{L}(\gamma, \phi|\mathbf{w}, \alpha, \beta)$ denote the right-hand side of Eq.(1) and expand it, we have

$$\mathcal{L} = \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) + \sum_{k=1}^{K}\left(\alpha_k - 1\right)\left(\Psi\left(\gamma_k\right) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right)$$
$$+ \sum_{t=1}^{V}\sum_{k=1}^{K}\phi_{t,k}\left(\Psi\left(\gamma_k\right) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right) + \sum_{t=1}^{V}\sum_{k=1}^{K}tf_{t,\mathbf{w}}\phi_{t,k}\log\beta_{t,k}$$
$$- \log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right) + \sum_{k=1}^{K}\log\Gamma\left(\gamma_k\right) - \sum_{k=1}^{K}\left(\gamma_k - 1\right)\left(\Psi\left(\gamma_k\right) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right)$$
$$+ \sum_{t=1}^{V}\sum_{k=1}^{K}\phi_{t,k}\log\phi_{t,k} \tag{2}$$

Where Γ is gamma function, Ψ is digamma function.

Firstly, we maximize Eq.(2) with respect to $\phi_{t,k}$, the probability that the word t was generated by latent topic z. This is a constrained maximization with constraint $\sum_{k=1}^{K}\phi_{t,k} = 1$. With $\beta_{t,k}$ reference to $p\left(w_t|z_t = k, \beta\right)$, we form the Lagrangian by isolating the terms which contain $\phi_{t,k}$ and adding the appropriate Lagrange multipliers, so we have

$$\mathcal{L}^{\mathbf{w}}_{[\phi_{t,k}]} = \phi_{t,k}\left(\Psi\left(\gamma_k\right) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right)$$
$$+ tf_{t,\mathbf{w}}\phi_{t,k}\log\beta_{t,k} + \phi_{t,k}\log\phi_{t,k} + \lambda_t\left(\sum_{k=1}^{K}\phi_{t,k} - 1\right)$$

Taking derivatives with respect to $\phi_{t,k}$ and setting the derivative to zero yields the maximized , we have

$$\phi_{t,k} \propto \left(\beta_{t,k}\right)^{tf_{t,\mathbf{w}}}\exp\left(\Psi\left(\gamma_k\right) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right) \tag{3}$$

Secondly, we maximize Eq.(2) with respect to γ_k, the k^{th} component of the posterior Dirichlet parameter. Take the derivative with respect to γ_k and setting to zero yields a maximum:

$$\gamma_k = \alpha_k + \sum_{t=1}^{V}\phi_{t,k} \tag{4}$$

2.3 Estimating

Given a corpus of $\mathcal{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$ that $\mathbf{w}$ is a category language model, we use a variational EM algorithm (EM with a variational E Step) [1] to find the parameters and which maximize a lower bound on the log marginal likelihood:

$$\ell(\alpha, \beta) = \sum_{\mathbf{w} \in \mathcal{D}} \log p(\mathbf{w}|\alpha, \beta)$$

As we have described above, we can bound the log likelihood using

$$\log p(\mathbf{w}|\alpha, \beta) = \mathcal{L}(\gamma, \phi|\mathbf{w}, \alpha, \beta) + D(q(\theta, z|\mathbf{w}, \gamma, \phi) \| p(\theta, z|\mathbf{w}, \alpha, \beta)) \quad (5)$$

Which exhibits $\mathcal{L}(\gamma, \phi|\mathbf{w}, \alpha, \beta)$ as a lower bound because the KL term is positive. We now obtain a variational EM algorithm that repeats the following two steps until Eq.(5) converges:

- (E step) For each category language model, optimize values for the variational parameter vectors γ and ϕ, the update rules are Eq.(3) and Eq.(4).
- (M step) Maximize the resulting lower bound on the log likelihood with respect to the model parameter α and parameter vector β. We can do this by finding the maximum likelihood estimates with expected sufficient statistics computed in the E-step.

Firstly, we maximize Eq.(2) with respect to $\beta_{t,k}$. This is a constrained maximization with constraint $\sum_{t=1}^{V} \beta_{t,k} = 1$, so we form the Lagrangian by isolating the terms which contain $\beta_{t,k}$ and adding the appropriate Lagrange multipliers, so we have

$$\mathcal{L}_{[\beta_{t,k}]} = \sum_{\mathbf{w} \in \mathcal{D}} \sum_{t=1}^{V} \sum_{k=1}^{K} tf_{t,\mathbf{w}} \phi_{t,k} \log \beta_{t,k} + \sum_{k=1}^{K} \lambda_k \left(\sum_{t=1}^{V} \beta_{t,k} - 1 \right)$$

Taking derivatives with respect to $\beta_{t,k}$ and setting the derivative to zero yields the maximized $\beta_{t,k}$, we have

$$\beta_{t,k} \propto \sum_{\mathbf{w} \in \mathcal{D}} tf_{t,\mathbf{w}} \phi_{t,k}$$

Secondly, we maximize Eq.(2) with respect to α. Then, take first derivative and second derivative with respective to α (α is a scalar dirichlet parameter). So according Newton-Raphson formula, we can find the maximal α by iteration as following:

$$\alpha^{\text{new}} = \alpha - \frac{M(\Psi(K\alpha) - K\Psi(\alpha)) + \sum_{\mathbf{w} \in \mathcal{D}} \sum_{k=1}^{K} \left\{ \Psi(\gamma_{k,\mathbf{w}}) - \Psi\left(\sum_{k=1}^{K} \gamma_{k,\mathbf{w}} \right) \right\}}{M \times K \times (\Psi'(K\alpha) - \Psi'(\alpha))}$$

where Ψ' is trigamma function.

2.4 Maximum a Posteriori of Multinomial Parameter

After model parameter α, model parameter vector β and variational parameter vector ϕ converged, we can fit the variational parameter vector γ as Eq.(4) description. Hereafter, to specific category language model $\mathbf{w}$, the maximum a posteriori of multinomial parameter in vector $\theta_{\mathbf{w}}$ can be computed approximately as

$$\theta_k^{\text{MAP}} = \frac{\gamma_k}{\sum_{k=1}^{K} \gamma_k} \qquad k = \{1, 2, \ldots K\}$$

Eventually, based on our model, we can derive maximum likelihood of document d generating by category language model $\mathbf{w}$ as following formula:

$$p(d) \propto \prod_{t \in d} \left\{ \sum_{k=1}^{K} \left(\theta_k^{\mathrm{MAP}} \beta_{t,k} \right) \right\}^{tf_{t,d}}$$

3 Experiments and Results

We have conducted experiments on two real-world datasets, Reuters21578 and 20newsgroups, to evaluate the effectiveness of our proposed model for text categorization.

The Reuters21578 dataset contains documents collected from Reuters newswire articles are assigned to 135 categories. However, some categories are empty and thus there are only non-empty 118 categories, among which the 10 most frequent categories called R10 by Debole [2] contain about 75% of the documents as Table 1 show. There are several ways to split the documents into training and testing sets: 'ModLewis' split, 'ModApte' split, and 'ModHayes' split. The 'ModApte' train/test split is widely used in text classification research. We followed the ModApte split in which the 10 most frequent categories and the numbers of documents are used for training and testing.

Table 1. Number of Training and Test documents About R10

Category name	Num Train	Num test
earn	2877	1087
acq	1650	719
money-fx	538	179
grain	433	149
crude	389	189
trade	369	118
interest	347	131
wheat	212	71
ship	197	89
corn	182	56

The 20Newsgroups(20NG) dataset is a collection of approximately 20,000 documents that were collected from 20 different newsgroups. This collection consists of 19,974 non-empty documents distributed evenly across 20 newsgroups and we selected 19,946 non-empty documents which are all the same after feature selection . We use the newsgroups to form categories, and randomly select 70% of the documents to be used for training and the remaining 30% for testing.

On the "Gerneral Text Toolkit" developing by our laboratory, We have tried our proposed LDACLM with 100 topics modeling the relationship among words, NaiveBayes with Laplace smoothing, and Rocchio algorithm [7] with TF-IDF scheme to these datasets respectively. Furthermore, We apply to Information

Table 2. Experimental results on the 20NG dataset

	NaiveBayes	LDACLM	Rocchio
macro-averaging precision	0.809	0.824	0.736
macro-averaging recall	0.808	0.813	0.739
macro-averaging F1	0.808	0.818	0.738
micro-averaging accuracy	0.803	0.813	0.736

Table 3. Experimental results on the Reuters21578 R10

	NaiveBayes	LDACLM	Rocchio
macro-averaging precision	0.662	0.660	0.647
macro-averaging recall	0.616	0.714	0.661
macro-averaging F1	0.638	0.686	0.654
micro-averaging accuracy	0.804	0.840	0.787

Gain [12] feature selecting method to the documents of both 20NG and Reuters-21578 R10 datasets with threshold 0.055 to 20NG and 0.3 to Reuters. The results of macro-averged and micro-averaged to 20NG and Reuters datasets are shown in Tables 2 and 3 for LDACLM, NaiveBayes and Rocchio respectively.

Specially, All results are averaged across 5 random runs for 20NG datatset. According experimental results, LDACLM outperform NaiveBayes with Laplace smoothing and Rocchio algorithm.

4 Conclusion and Future Work

This paper proposed Latent Dirichlet Allocation Category Language Model, a novel model based on LDA model. We have presented variational inference approach, and parameters estimation method which is similar to LDA [1] in category language model. As Results on 20NG and Reuters21578 datasets shown above, LDACLM cannot significantly improve performance. In our opinion, we think that it was because the topics modeling the relationship among words is not abundant which constraint by computer memory. In the future work, we will try use topics by collection from Wordnet based on Gibbs sample, and this maybe create many topics which approximate words dependency than variational inference do.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments and suggestions. We are grateful for Zhao Cao's helpful discussion and advice. Many thanks also give to Shidong Feng, Yingfan Gao, Jian Cao, Jinghua Bai, and Xu Zhang for their suggestions regarding this paper.

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Debole, F., Sebastiani, F.: An Analysis of the Relative Difficulty of Reuters-21578 Subsets. Journal of the American Society for Information Science and Technology 56(2), 584–596 (2004)
3. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
4. Girolami, M., Kaban, A.: On an equivalence between PLSI and LDA. In: Proceedings of the 26^{th} annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, pp. 433–434 (2003)
5. Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101, 5228–5235 (2004)
6. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: Proceedings of the 22^{nd} annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, pp. 50–57 (1999)
7. Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In: Proceedings of the 14^{th} International Conference on Machine Learning, Nashville, TN, USA (1997)
8. Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L.: An introduction to variational methods for graphical models. Machine Learning 37, 183–233 (1999)
9. Ponte, J., Croft, W.: A Language Modeling Approach to Information Retrieval. In: Proceedings of the 21^{st} annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, pp. 275–281 (1998)
10. Wallach, H.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23^{rd} International Conference on Machine Learning (2006)
11. Wei, X., Croft, W.: LDA-Based Document Models for Ad-hoc Retrieval. In: Proceedings of the 29^{th} Annual International ACM SIGIR Conference on Research and Development on Information Retrieval, Seattle, pp. 178–185 (2006)
12. Yang, Y., Pedersen, J.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the 14^{th} International Conference on Machine Learning, Nashville, TN, USA, pp. 412–420 (1997)