Probabilistic Text Modeling with Orthogonalized Topics

Enpeng Yao[1], Guoqing Zheng[2], Ou Jin[1], Shenghua Bao[3],
Kailong Chen[1], Zhong Su[3], Yong Yu[1]

[1]Shanghai Jiao Tong University, Shanghai, 200240, China
[2]Carnegie Mellon University, PA, 15213, USA
[3]IBM China Research Laboratory, Beijing, 100094, China
{yaoenpeng, kingohm, chenkl, yyu}@apex.sjtu.edu.cn
gzheng@cs.cmu.edu, {baoshhua, suzhong}@cn.ibm.com

ABSTRACT

Topic models have been widely used for text analysis. Previous topic models have enjoyed great success in mining the latent topic structure of text documents. With many efforts made on endowing the resulting document-topic distributions with different motivations, however, none of these models have paid any attention on the resulting topic-word distributions. Since topic-word distribution also plays an important role in the modeling performance, topic models which emphasize only the resulting document-topic representations but pay less attention to the topic-term distributions are limited. In this paper, we propose the Orthogonalized Topic Model (OTM) which imposes an orthogonality constraint on the topic-term distributions. We also propose a novel model fitting algorithm based on the generalized Expectation-Maximization algorithm and the Newthon-Raphson method. Quantitative evaluation of text classification demonstrates that OTM outperforms other baseline models and indicates the important role played by topic orthogonalizing.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

Keywords

Probabilistic Text Modeling; Latent Semantic Analysis; Text Classification

1. INTRODUCTION

Topic modeling has attracted increasing attention and widely used for data analysis in many domains including text documents. By assuming that each document is modeled by a distribution over a certain number of topics, each of which is a distribution over words in the vocabulary, topic models can reveal the underlying latent semantic structure which facilitates further applications such as classification, clustering and retrieval. Traditional topic models, including Probabilistic Latent Semantic Analysis (PLSA) [7], Latent Dirichlet Allocation (LDA) [1] and many of their variations, most of which put certain constraints on the *document-topic* distributions based on specific considerations, have enjoyed impressive success in tasks [2, 9], etc.

Topic models usually contain two parts of parameters, which are *documents-topic* distributions and *topic-word* distributions, and intuitively *topic-word* distributions also play an important role in the whole model. A model with better *topic-word* distributions can mine more diversified and reasonable topics, and therefore it may achieve better performance on many text mining tasks, like text classification and clustering. To the best of our knowledge, none of previous topic models pay any extra consideration on the *topic-word* distributions, such as orthogonalizing them, which is to be discussed in this paper.

Orthogonalizing techniques have been widely used in dimension reduction models [6, 8]. So we hope adding the orthogonality constraint to the topic models also can achieve performance improvement.

In this paper, we propose a new topic model, the Orthogonalized Topic Model (OTM), to focus on orthogonalizing the *topic-word* distributions. In order to address the importance of orthogonalized topics, we put a regularized factor measuring the degree of topic orthogonalities to the objective function of PLSA. The OTM model is able to take advantage of statistical foundation of PLSA without losing orthogonal property of LSA. We present an efficient algorithm to solve the proposed regularized log-likelihood maximization problem using the Expectation-Maximization(EM) algorithm [5] and the Newton-Raphson method.

To evaluate the proposed model, we apply it to two widely used text corpora on the classification task. The higher text classification accuracy based on OTM further stresses the important role played by topic orthogonalization.

2. BACKGROUND AND NOTATIONS

2.1 Latent Semantic Analysis

Latent Semantic Analysis (LSA) [4], as one of the most useful tools for learning the latent concepts from text, has widely been used in the dimension reduction task. Specifically, given a term-document matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$, where N is the number of words in the vocabulary and M is the number of documents in the corpus, by using singular value decom-

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609471.

position (SVD), LSA tries to find a low-rank construction of $\mathbf{X}$ such that

$$\mathbf{X} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}} \tag{1}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{K\times K}$ is a diagonal matrix containing the K largest singular values of $\mathbf{X}$. Orthogonal matrices $\mathbf{U} \in \mathbb{R}^{N\times K}(\mathbf{U}^{\mathbf{T}}\mathbf{U} = \mathbf{I})$ and $\mathbf{V} \in \mathbb{R}^{M\times K}(\mathbf{V}^{\mathbf{T}}\mathbf{V} = \mathbf{I})$ represent word and document embeddings onto the latent space.

2.2 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) [7] defines a proper generative model based on a solid statistical foundation.

Given a corpus that consists of M documents $\{d_1, d_2, ..., d_M\}$ with words from a vocabulary of N words $\{w_1, w_2, ..., w_N\}$, PLSA, by assuming that the occurrence of a word w in a particular document d is assigned with one of K latent topic variables $\{z_1, z_2, ..., z_K\}$, defines the following generative process:

- Select a document d_i with probability $P(d_i)$;
- Pick a latent topic z_k with probability $P(z_k|d_i)$;
- Generate a word w_j with probability $P(w_j|z_k)$.

By summing all the latent variable z_k, the joint probability of an observed pair (d_i, w_j) can be computed as

$$P(d_i, w_j) = P(d_i)P(w_j|d_i) = P(d_i)\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{2}$$

So we can calculate the data log-likelihood as

$$\overline{\mathcal{L}} = \sum_{i=1}^{M}\sum_{j=1}^{N} n(d_i, w_j)\log\left(P(d_i)\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)\right) \tag{3}$$

where $n(d_i, w_j)$ denotes the number of times word w_j occurs in document d_i. Following the principles of maximum likelihood estimation, we can estimate $P(w_j|z_k)$ and $P(z_k|d_i)$.

According to the above modeling, PLSA successfully endows its resulting document-topic and topic-word assignments a statistical meaning which LSA lacks; however, it discards the orthogonal properties of the resulting dimensions originally possessed by LSA.

3. ORTHOGONALIZED TOPIC MODEL

In this section, we firstly introduce the measure to evaluate the degree to which topics are orthogonal to each other and then formalize our proposed model, named Orthogonalized Topic Model (OTM). We also present an efficient algorithm to solve the proposed optimization function using the Expectation Maximization (EM) algorithm [5].

3.1 Orthogonality Measure

In order to measure the degree to which topics are orthogonal to each other without harming their probabilistic property, the orthogonal constraint in LSA which ensures that each column vector of the resulting dimension are strictly orthogonal to other column vectors, such as $\mathbf{U}^{\mathbf{T}}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{\mathbf{T}}\mathbf{V} = \mathbf{I}$, can no longer be used. Instead, in this paper, we put a soft orthogonality constraint on the topic-word distributions in order to force topics as orthogonal to each other as possible. The constraint we proposed is to minimize the following orthogonalization factor:

$$\mathcal{O} = \sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\sum_{j=1}^{N} P(w_j|z_k)P(w_j|z_{k'}) \tag{4}$$

Minimizing this objective function will directly enforce the topic-word distributions of different topics to be orthogonal. One can conclude from the above factor that as $\mathcal{O}$ decreases to 0 all the topics are more likely to be orthogonal with other topics.

3.2 Objective Function of OTM

Preserving the statistical properties of the resulting topic-word distributions while ensuring that the resulting topic-word distributions are as orthogonal to others as possible, we formulate OTM as combining the statistical foundation of PLSA and the orthogonalized constraint mentioned above.

Formally, OTM tries to maximize the following objective function:

$$\begin{aligned}\mathcal{Q} =& \mathcal{L} - \lambda\mathcal{O} \\ =& \sum_{i=1}^{M}\sum_{j=1}^{N} n(d_i, w_j)\log\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \\ &- \lambda\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\sum_{j=1}^{N} P(w_j|z_k)P(w_j|z_{k'})\end{aligned} \tag{5}$$

where $\lambda \geq 0$ is the regularization parameter. Note that if $\lambda = 0$ OTM boils down to PLSA and we assume $\lambda > 0$ thereafter.

3.3 Model Fitting with EM

When a model involves unobserved latent variables, the EM algorithm is the standard procedure for maximum likelihood estimation. EM alternates between two steps: (a) an Expectation(E) step where posterior probabilities of the latent variables are computed given the current estimates of the parameters, (b) a maximization(M) step where parameters are updated by maximizing the expected complete data log-likelihood given the posterior probabilities computed in the E-step. The parameters of OTM are $P(z_k|d_i), P(w_j|z_k)$. Thus, $MK + NK$ parameters are to be estimated, which is the same as PLSA.

E-step: The E-step of OTM is exactly the same as that of PLSA. The posterior probabilities for the latent variables $P(z_k|d_i, w_j)$ can be computed using applying Bayes' formula.

$$P(z_k|d_i, w_j) = \frac{P(z_k|d_i)P(w_j|z_k)}{\sum_{k'=1}^{K} P(z_{k'}|d_i)P(w_j|z_{k'})} \tag{6}$$

M-step: In the M-step of OTM, we improve the expected value of the complete data log-likelihood which is

$$\begin{aligned}\mathcal{Q} =& \sum_{i=1}^{M}\sum_{j=1}^{N} n(d_i, w_j)\log\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \\ &- \lambda\sum_{k=1}^{K}\sum_{\substack{k'=1\\k'\neq k}}^{K}\sum_{j=1}^{N} P(w_j|z_k)P(w_j|z_{k'})\end{aligned} \tag{7}$$

with the constraints $\sum_{k=1}^{K} P(z_k|d_i) = 1$ and $\sum_{j=1}^{N} P(w_j|z_k) = 1$.

The M-step re-estimation for $P(z_k|d_i)$ is exactly the same as those of PLSA because the orthogonalized term of OTM does not include $P(z_k|d_i)$.

$$P(z_k|d_i) = \frac{\sum_{j=1}^{N} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{j=1}^{N} n(d_i, w_j)} \tag{8}$$

Now we turn to the re-estimation part for $P(w_j|z_k)$. Maximization of $\mathcal{Q}$ with respect to $P(w_j|z_k)$ leads to the following set of equations:

$$\begin{aligned} & \frac{\partial \mathcal{Q}}{\partial P(w_j|z_k)} \qquad (9) \\ = & \frac{\sum_{i=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{P(w_j|z_k)} - \lambda \sum_{\substack{k'=1 \\ k' \neq k}}^{K} P(w_j|z_{k'}) - \alpha_k \\ = & \ 0 \end{aligned}$$

where $\alpha_k \geq 0$ is the Lagrange multiplier for topic k. So we get

$$P(w_j|z_k) = \frac{\sum_{i=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{\alpha_k + \lambda \sum_{\substack{k'=1 \\ k' \neq k}}^{K} P(w_j|z_{k'})} \tag{10}$$

Letting $\sum_{j=1}^{N} P(w_j|z_k) = 1$ we get

$$\sum_{j=1}^{N} \frac{\sum_{i=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{\alpha_k + \lambda \sum_{\substack{k'=1 \\ k' \neq k}}^{K} P(w_j|z_{k'})} = 1 \tag{11}$$

Due to the coupling between the summation in the denominator and the summation outside the fraction, it is not easy to give an analytic solution for the $P(w_j|z_k)$ in Eq.(10). Thus we turn to find a numerical solution. We apply Newton-Raphson method to calculate α_k from (11) as below:

$$\alpha_k^{n+1} = \alpha_k^n - \frac{f_k(\alpha_k^n)}{f_k'(\alpha_k^n)} \tag{12}$$

where

$$f_k(x) = \sum_{j=1}^{N} \frac{\sum_{i=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)}{x + \lambda \sum_{\substack{k'=1 \\ k' \neq k}}^{K} P(w_j|z_{k'})} - 1 \tag{13}$$

α_k^n is the value of α_k in the nth iteration. We initialize α_k with a random real number α_k^0 and get the root of the function f_k, or the numerical solution of α_k by updating the value of α_k iteratively until a sufficiently accurate value is reached. By noting that α_k, $\lambda \sum_{\substack{k'=1 \\ k' \neq k}}^{K} P(w_j|z_{k'})$ and $\sum_{i=1}^{M} n(d_i, w_j) P(z_k|d_i, w_j)$ are non-negative real numbers, it is easy to prove that the left hand side of Eq.(11) is a monotone function for the parameter α_k. This indicates that there is only one accurate numerical solution of α_k, with nothing to do with the initial number α_k^0. By using the value of a_k, $P(w_j|z_k)$ is calculated by Eq.(10). After obtaining $P(z_k|d_i)$ and $P(w_j|z_k)$ in the M-step, our algorithm continues to calculate the EM step cycles, until the value of objective function (5) converges.

4. EXPERIMENTS

To evaluate the effectiveness of the proposed OTM model, we apply it on two widely used text corpora, Yahoo! News K-series and Reuters 21578 corpora. We will report the quantitative evaluation of document classification based on the topics extracted from various topic models.

4.1 Datasets and Experimental Settings

Yahoo! News K-series is a collection consisting of 2,340 news articles belonging to 20 different categories. The numbers of articles per categories are almost the same. We get the preprocessed version with a vocabulary size of 8104 from D. L. Boley's page[1]. The Reuters-21578 dataset is a corpus of news stories made available by Reuters, Ltd. The preprocessed version is downloaded from R. F. Corrêa's web page[2] which consists of 9,821 documents and 5,180 distinct words. Table 1 shows the statistical details of the two datasets.

All experiments are conducted on a computer with 2.6GHz Pentium Dual Core CPU and 2 GB physical memory.

4.2 Classification Evaluation

We evaluate the performance of OTM on the tasks of document classification using the method similar to [9]. To address real-world problems in a semi-supervised setting, We first using OTM to generate the document-topic and word-topic distributions from the whole dataset, and randomly select a small number of documents (one of 10,20,and 40) from each category as labeled for training and use the remaining documents for test. The selected training data number from each category are the same. Then a linear kernel support vector machine (SVM) [3] is trained on the document-topic representations of the training set and make prediction on the test set. This means that we use the rank-reduced representations of each document (the document-topic distributions) as the input space of SVM, and the output space is the categories of documents.

Beside the LSA and PLSA, we also provide two more baselines:

- Using raw word features without any dimension reduction as input for SVM
- Laplacian Probabilistic Semantic Analysis (LapPLSA) [2]

LapPLSA is a topic model based on PLSA with the modification on the document-topic distribution. It assumes that the document space is a manifold, either linear or nonlinear. By adding a term related to document-topic distribution as regularization to the objective function of PLSA, this model achieves high performance on the text clustering task. We choose LapPLSA as a representation which add constraints on the resulting document-topic distribution to compare with our OTM which endows the resulting topic-words distribution. We use the source code downloaded from D. Cai's webpage[3].

For LSA, PLSA, LapPLSA, the dimension reduction results of each document are set as the input for SVM, similar to OTM. For each model, we vary the number of latent topics K from 10 to 100. We report the average of classification accuracies after 10 test runs.

Figures 1 and 2 demonstrate the classification performance of OTM and other baseline models. On both text sets, OTM outperforms LSA, PLSA, LapPLSA in terms of classification accuracies due to the orthogonality of the topics. This indicates that the OTM model, which combines the statistical foundation of PLSA and the orthogonalized constraint, improves topic representation of documents to a certain degree.

[1] http://www-users.cs.umn.edu/~boley/ftp/PDDPdata/
[2] https://sites.google.com/site/renatocorrea02/
[3] http://www.zjucadcg.cn/dengcai/LapPLSA/index.html

Table 1: Datasets details

	# of docs	# of terms	# of labels	avg. terms/doc
Yahoo News K	2,340	8,104	20	105.7
Reuters-21578	9,821	5,180	20	70.3

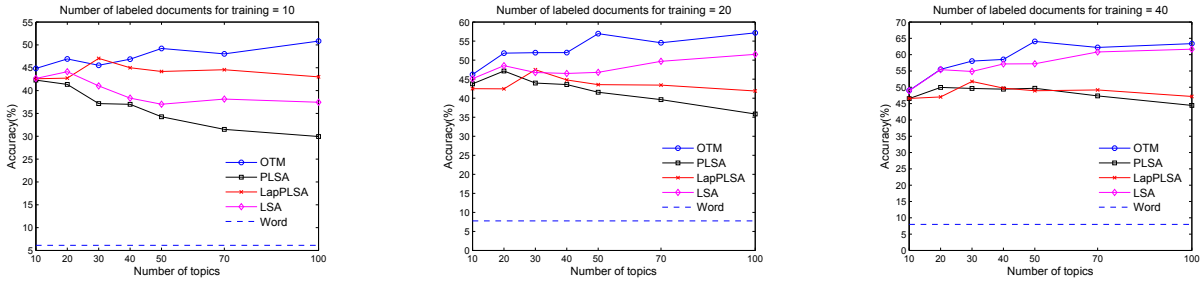

Figure 1: Classification performance on Yahoo! News K-series. The "number of labeled documents for training" is per category.

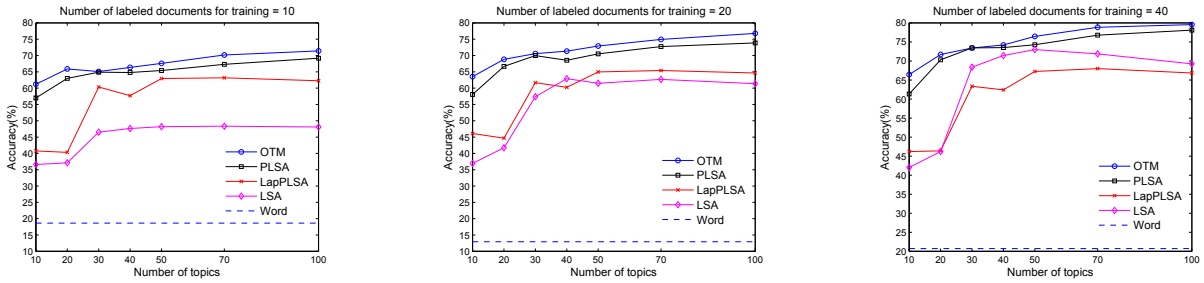

Figure 2: Classification performance on Reuters-21578. The "number of labeled documents for training" is per category.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose an Orthogonalized Topic Model (OTM) which aims to increase the resulting topic diversity by putting a constraint on topic-word distributions to orthogonalize the topics. The consideration to constrain the topic-word distribution has not been made in previous works. We formulate the model as a maximization problem and fit the model efficiently using the EM algorithm and the Newton-Raphson method. We conduct experiments on two real world text corpora. The application of topic models to text classification verifies the effectiveness of the proposed model.

In future, we plan to add the orthogonality constraint into other topic modeling algorithms such as Latent Dirichlet Allocation(LDA) to improve their performance. In another direction, we will use some other measure to express the orthogonality and obtain faster optimization process.

6. REFERENCES

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

[2] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *Proceeding of the 17th ACM conference on Information and knowledge management (CIKM'08)*, pages 911–920, 2008.

[3] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[4] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[6] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. In *Machine Learning*, pages 143–175, 2000.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of SIGIR 1999*, pages 50–57, 1999.

[8] P. O. Hoyer and P. Dayan. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[9] S. Huh and S. E. Fienberg. Discriminative topic modeling based on manifold learning. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 653–662, 2010.