Towards Noise-Resilient Document Modeling

Tao Yang
The Pennsylvania State University
University Park, PA 16802, USA
tyang@ist.psu.edu

Dongwon Lee[*]
The Pennsylvania State University
University Park, PA 16802, USA
dongwon@psu.edu

ABSTRACT

We introduce a generative probabilistic document model based on latent Dirichlet allocation (LDA), to deal with textual errors in the document collection. Our model is inspired by the fact that most large-scale text data are machine-generated and thus inevitably contain many types of noise. The new model, termed as TE-LDA, is developed from the traditional LDA by adding a switch variable into the term generation process in order to tackle the issue of noisy text data. Through extensive experiments, the efficacy of our proposed model is validated using both real and synthetic data sets.

Categories and Subject Descriptors

H.1.0 [**Information Systems**]: Models and Principles

General Terms

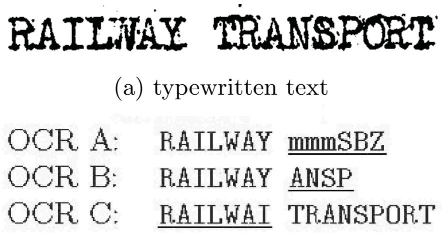
Algorithms, Theory

Keywords

Topic Models, Textual Errors

1. INTRODUCTION

Probabilistic topic models are stochastic models for text documents, which explicitly model topics in the document corpus. As generative models, they describe a procedure for generating documents using a series of probabilistic steps. Since it was introduced in 2003 [1], the latent Dirichlet allocation (LDA) model has quickly become a powerful tool for statistical analysis of text documents. LDA assumes that text documents are mixtures of hidden topics and applies Dirichlet prior distribution over the latent topic distribution of a document having multiple topics. Also, it assumes that topics are probability distribution of words and words are sampled independently from a mixture of multinomials. Therefore, LDA is a widely used Bayesian topic model which can model the semantic relations between topics and words for document corpora.



RAILWAY TRANSPORT

(a) typewritten text

OCR A: RAILWAY mmmSBZ
OCR B: RAILWAY ANSP
OCR C: RAILWAI TRANSPORT

(b) OCR output

Figure 1: Three examples of erroneous OCR output of a poor quality typewritten text (taken from [2]). Erroneous outputs are underlined.

LDA requires accurate counts of the occurrences of words in order to estimate the parameters of the model, Therefore, it assumes that the entire document corpus is clean in order to ensure correct calculation of frequencies of words. However, as text data become available in massive quantities, textual errors are appearing inevitable in large-scale document corpora. These textual errors include typos, spelling errors, transcription errors caused by text or speech recognition tools, digitization errors of Google Books and Internet Archives, etc. For example, Walker et al. [2] point out that although researchers are having increasing levels of success in digitizing hand-written manuscripts, error rates remain significantly high. Figure 1 shows an example of typewritten documents and output by three Optical Character Recognition (OCR) engines. Even on clean data, LDA will often do poorly if the very simple feature selection steps of removing stop-words is not performed first. It is shown that the performance in terms of accuracy declines significantly as word error rates increase [2], which highlights the importance of taking into account the noisy data issue in document modeling.

Motivated by the above observations, in this paper, we introduce our new model to tackle the issue of noisy data. In particular, we propose a new LDA model termed as TE-LDA to take into account textual errors in the document generation process. We compare the performance of our new model against the traditional LDA model and report promising results of our proposal in terms of perplexity. Through extensive experiments, the efficacy of our proposed models is validated using both real and synthetic data sets.

[*]Partially supported by NSF DUE-0817376 and DUE-0937891 awards.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

2. RELATED WORK

Probabilistic document modeling has recently received tremendous attention in the data mining community. A series of probabilistic models including the Naive Bayesian model and the Probabilistic Latent Semantic Indexing (PLSI) model have been introduced to simulate the document generation process. The LDA model has become most popular in the data mining and information retrieval community due to its solid theoretical statistical foundation and promising performance. A wide variety of extensions of LDA model have been proposed for different modeling purposes in different contexts. For example, the correlated LDA model learns topics simultaneously from images and caption words [3].

However, topic modeling techniques require clean document corpus. This is to prevent the model from confusing patterns which emerge in the noisy data. Recent work by Walker [2] is the first comprehensive study of document clustering and LDA on synthetic and real-word Optical Character Recognition data. The character-level textual errors introduced by OCR engines serve as baseline document corpora to understand the accuracy of document modeling in erroneous environment. The study shows that the performance of topic modeling algorithms degrades significantly as word error rates increase.

3. LDA MODEL

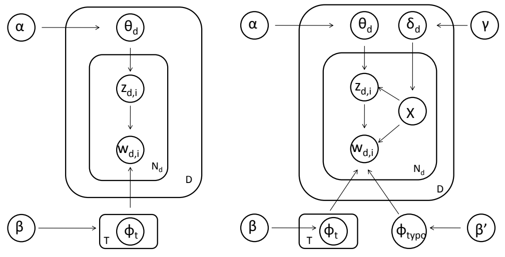
In this section, we give a brief overview of the LDA model. Blei et al. [1] introduced it as a semantically consistent topic model, which attracted a considerable interest from both the statistical machine learning and natural language processing communities. LDA models documents by assuming that a document is composed by a mixture of hidden topics and that each topic is characterized by a probability distribution over words. The model is shown as a graphical model in Figure 2(a). The notation is shown in Table 1. θ_d denotes a T-dimensional probability vector and represents the topic distribution of document d. ϕ_t denotes a W-dimensional probability vector where $\phi_{t,w}$ specifies the probability of generating word w given topic t. $Multi(.)$ denotes multinomial distribution. $Dir(.)$ denotes Dirichlet distribution. α is a T-dimensional parameter vector of the Dirichlet distribution over θ_d, and β is a W-dimensional parameter vector of the Dirichlet distribtion over ϕ_t. The process of generating documents is shown in Algorithm 1.

Algorithm 1: **LDA Model**.

```
For each of the $T$ topics $t$, sample a multinomial
  distribution $\phi_t$ from a Dirichlet distribution with prior
  $\beta$;
For each of the $D$ documents $d$, sample a multinomial
  distribution $\theta_d$ from a Dirichlet distribution with prior
  $\alpha$;
For each word $w_{d,i}$ in document $d$, sample a topic $z_{d,i}$
  from the multinomial distribution $\theta_d$;
Sample word $w_{d,i}$ from the multinomial distribution
  $\phi_{z_{d,i}}$.
```

Table 1: Notations

Symbol	Description
D	total number of documents
W	total number of word tokens
T	total number of topics
N_d	the number of words in document d
$w_{d,i}$	ith word in document d
$z_{d,i}$	latent topic at ith word in document d
$\theta_{d,i}$	probability of topic i in document d
$\phi_{t,w}$	probability of word w in topic t



(a) LDA Model. (b) TE-LDA Model.

Figure 2: Comparison between our model vs. LDA model.

4. PROPOSED MODEL

To overcome the constraints of the above traditional LDA topic model, in this section, we propose a new LDA model termed as TE-LDA (LDA with **T**extual **E**rrors) to take into account noisy data in the document generation process. In this model, we distinguish the words in the documents and separate them as tokens and typos. Given a document, each word has a probability to be an error and we want to capture this probability structure in the term generation process. In order to reflect the nature of textual errors in the generative model, we adopt a switch variable to control the influence of errors on the term generation. The proposed model is illustrated in Figure 2(b). N_d is the total number of words in document d (with $N_d = N_{term} + N_{typo}$, the sum of all the true terms and typos). α, β and β' are Dirichlet priors, θ_d is the topic-document distribution, ϕ_t is the term-topic distribution. ϕ_{typo} is the term distribution specifically for typos. We include an additional binomial distribution δ with a Beta prior of γ which controls the fraction of errors.

For each word w in a document d, a topic z is sampled first and then the word w is drawn conditional on the topic. The document d is generated by repeating the process N_d times. To decide if each word is an error or not, a switch variable X is introduced. The value of X (which is 0 or 1) is sampled based on a binomial distribution δ with a Beta prior of γ. When the sampled value of X equals 1, the word w is drawn from the topic z_t which is sampled from the topics learned from the words in document d. When the value of X equals 0, the word w is drawn directly from the term distribution for errors. Overall, the generation process for TE-LDA can be described in Algorithm 2. We omit the derivation of parameter estimation due to space limit.

5. EXPERIMENTAL VALIDATION

We trained our new model as well as the traditional LDA model on both synthetic and real text corpora to compare the generalization performance. Each model was trained on 90% of the documents in each data set and the trained model was used to calculate the estimates of the marginal log-likelihood of the remaining 10% of documents.

Algorithm 2: **TE-LDA Model**.

```
For each of the $D$ documents $d$ , sample $\theta_d \sim$ Dir($\alpha$)and $\delta_d \sim$ Beta($\gamma$);
For each of the $T$ topics $t$, sample $\phi_t \sim$ Dir($\beta$);
Sample $\phi_{typo} \sim$ Dir($\beta'$);
foreach $N_d$ words $w_{d,i}$ in document $d$ do
    Sample a flag $X \sim$ Binomial($\delta_d$);
    if $X = 1$ then
        Sample a topic $z_{d,i} \sim$ Multi($\theta_d$);
        Sample a word $w_{d,i} \sim$ Multi($\phi_{z_{d,i}}$);
    endif
    if $X = 0$ then
        Sample a word $w_{d,i} \sim$ Multi($\phi_{typo}$);
    endif
endfch
```

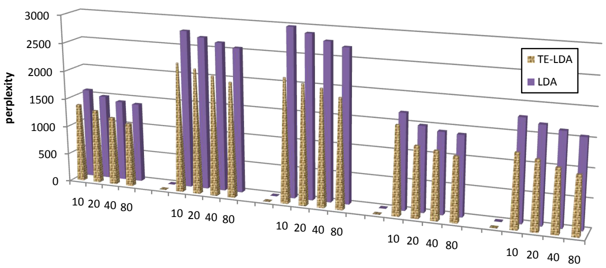


Figure 3: Perplexity of different models in Unlv data set. From left to right, the subsets are Business, Magazine, Legal, Newspaper, Magazine2.

5.1 Data Sets

In our experiment, we used three benchmark data sets TREC AP, NIPS, and Reuters-21578 in the document modeling literature. The TREC Associate Press (AP) data set[1] contains 16333 newswire articles with 23075 unique terms. The NIPS data set[2] consists of the full text of the 13 years of proceedings from 1988 to 2000 Neural Information Processing Systems (NIPS) Conferences. The data set contains 1740 research papers with 13649 unique terms. The Reuters-21578 data set[3] consists of newswire articles classified by topic and ordered by their date of issue. The data set contains of 12902 documents and 12112 unique terms. For all the above data sets, we synthetically generated noisy text data to simulate different levels of Word Error Rates (WER). We also conducted experiments on one real OCR data set Unlv[4] with five subsets, namely Business, Magazine, Legal, Newspaper, Magazine2. The average document WER generated by the OCR engine is around 30%.

5.2 Evaluation Metrics

We calculated the *perplexity* of a held-out test set to evaluate the models. A lower perplexity score corresponds to better generalization performance of the document model. Formally, for a test data of D_{test} documents the perplexity score is calculated as follows [1]. Note that the probability $p(w_d|z_k)$ is learned from the training process and $p_{test}(z_k|d)$ is estimated on the test data based on the parameters ϕ, θ and δ learned from training data.

$$perplexity(D_{test}) = \exp\{\frac{-\sum_{d=1}^{D_{test}} \log p(\mathbf{w}_d)}{\sum_{d=1}^{D_{test}} N_d}\} \qquad (1)$$

$$p(\mathbf{w}_d) = \sum_{k=1}^{K} p(w_d|z_k)p_{test}(z_k|d) \qquad (2)$$

5.3 Experimental Results

We first compare the performance of our proposed model with the baseline LDA model on the real OCR dataset. Figure 3 show the perplexity of TE-LDA as a function of the number of hidden topics in the five subsets of Unlv corpus. At different levels of WER for each subset, our TE-LDA model consistently outperforms the traditional LDA model.

We then systematically compare the performance of our model with LDA on the synthetically generated noisy corpora. In this experiment, we simulate different levels of word error rates (WER= 0.01, 0.05, 0.1). Figures 4(a)-(c) show the perplexity of TE-LDA model as a function of the number of hidden topics in the document corpus on AP data set. As we can see from Figures 4(a)-(c), at different levels of WER, our TE-LDA model consistently outperforms the traditional LDA model. Furthermore, as WER increases, the margin of improvement increases. This is due to explicit modeling of textual errors during the generation of terms in the document modeling process. In Figures 4(d)-(f), we fix the number of topics T and demonstrate how the different models perform as the word error rates increase. An interesting finding here is that the perplexity of LDA increases as the word error rates increase while the perplexity of TE-LDA models decreases as the word error rates increase. This is because LDA does not consider the textual errors in the term generation where the accuracy of calculation of word frequencies is affected significantly in the noisy text environment. In summary, our TE-LDA outperforms LDA and the margin of improvement increases as the word error rates increase. Figures 4(g)-(l) and Figures 4(m)-(r) show similar patterns on NIPS data set and Reuters data set respectively.

6. CONCLUSION AND FUTURE WORK

In this paper, we extend the traditional LDA model to account for noisy text data in latent document modeling. Our TE-LDA adopts a switching mechanism to explicitly determine whether the word is generated from the topic-document distribution through the general topic generation route or from a special word distribution through the typo processing route. We show that our proposed model achieves better generalization performance than the LDA model.

7. REFERENCES

[1] D. M.Blei et al., *Latent Dirichlet Allocation*, In Journal of Machine Learning Research, 2003.

[2] D. D.Walker et al., *Evaluating Models of Latent Document Semantics in the Presence of OCR Errors*, In EMNLP, 2010.

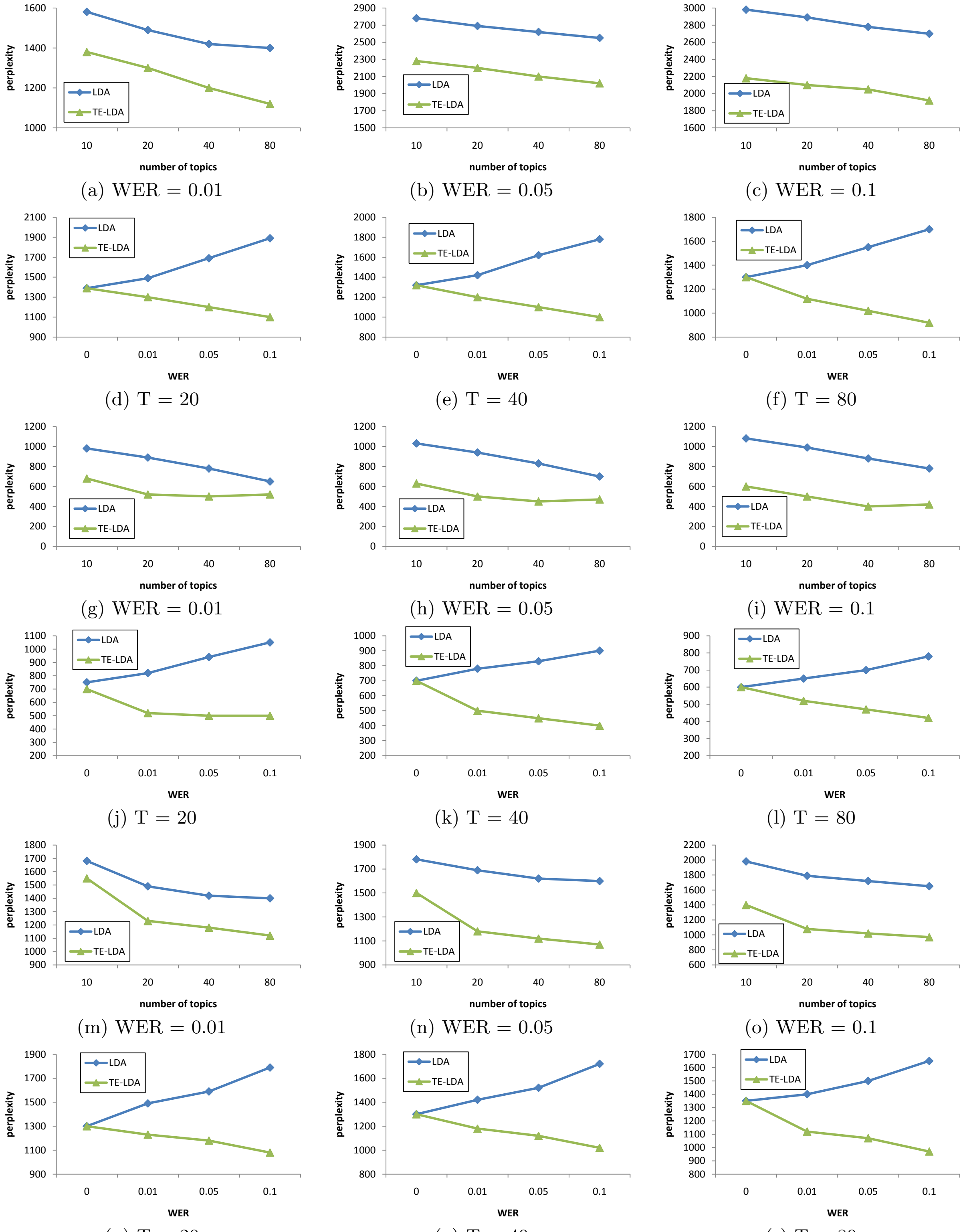
[3] X. Chen et al., *Probabilistic Models for Topic Learning from Images and Captions in Online Biomedical Literatures*, In CIKM, 2009.

[1] http://www.daviddlewis.com/resources/testcollections/trecap/

[2] http://www.cs.nyu.edu/ roweis/data.html

[3] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[4] http://code.google.com/p/isri-ocr-evaluation-tools/updates/list



(a) WER = 0.01 (b) WER = 0.05 (c) WER = 0.1

(d) T = 20 (e) T = 40 (f) T = 80

(g) WER = 0.01 (h) WER = 0.05 (i) WER = 0.1

(j) T = 20 (k) T = 40 (l) T = 80

(m) WER = 0.01 (n) WER = 0.05 (o) WER = 0.1

(p) T = 20 (q) T = 40 (r) T = 80

Figure 4: (a)-(c),(g)-(i) and (m)-(o): Perplexity of different models as a function of the number of topics in `TREC AP`, `NIPS` and `Reuters` data sets respectively. (d)-(f), (j)-(l) and (p)-(r): Perplexity of different models as a function of WER in `TREC AP`, `NIPS` and `Reuters` data sets respectively.