An Improved LDA Model for Academic Document Analysis

Yuyan Jiang, Yuan Shao
School of Management Science and Engineering/Anhui University of Technology, Ma'anshan, China
Email: {Jyy, yuanshao}@ahut.edu.cn

Ping Li and Qing Wang
School of Management Science and Engineering/Anhui University of Technology, Ma'anshan, China
Email: {pingLi, Wangq}@ahut.edu.cn

***Abstract*—Electronic documents on the Internet are always generated with many kinds of side information. Although those massive kinds of information make the analysis become very difficult, models would fit and analyze data well if they could make full use of those kinds of side information. This paper, base on the study on probabilistic topic model, proposes a new improved LDA model which is suitable for analysis of academic document. Based on the modification of standard LDA model, this new improved LDA model could analyze documents with both authors and references. To evaluate the generalization capability, this paper compares the new model with standard LDA and DMR model using the widely used Rexa dataset. Experimental results show that the new model has a high capability of document clustering and topics extraction than standard LDA and its modifications. In addition, the new model outperforms DMR model in task of authors discriminant.**

***Index Terms*—academic documents; topic model; topics extraction; authors discriminant**

I. INTRODUCTION

Characterizing the content of documents is a standard problem addressed in information retrieval, statistical natural language processing, and machine learning. Recently, probabilistic topic models such as Latent Dirichlet Allocation(LDA)[1] and its modifications are widely used in information retrieval and text mining. These algorithms provide us with new ways to organize, search and summarize large collection of text documents.

The standard LDA model treats each document as a bag-of-words and widely used in natural language process, information retrieval, emotion analysis and many other applications[2,3,4]. However, in many real world applications, user can receive text datasets with many other kinds of side information, for example, when online user post their evaluation for service or products, they also may vote or give service and products scores; Images in the MNIST dataset have many labels which indicate categories that images are belong to; Pages in the public Yahoo! Directory may also have many tags; Academic papers in the academic network search may contain authors, references and many other kinds of side information. These kinds of side information could provide important information in tasks of classification and topics extraction, while are not used by standard LDA and most of its modifications. LDA models that utilize these side information could not only improve the accuracy of labels indication, but also learn better topics and disambiguate words that may belong to different topics.

In this paper, we mainly focus on the application of LDA model in the academic network search, in which the main task is to analyze documents with both authors and references information and recommending authors, papers for users based on their queries[5]. Based on the study on LDA models and their documents generative process, this paper proposes a new improved LDA model(Author & Reference Topic Model, ART) which could analyze documents with both Authors and references information. An optimization algorithm based on the stochastic EM sampling method[6] is proposed for the optimization of ART model. Experimental results show that the ART model has a more accurate capability when predict authors for a new document. In addition, ART model could be viewed as a semi-supervised clustering model which could make full use of authors and references information to improve the performance of clustering and topic extraction.

II. DSTM AND USTM

Transforming the standard LDA model into supervised or semi-supervised model can make the model use kinds of side information efficiently and have a higher performance of documents clustering and topics extraction. what's more, supervised or semi-supervised LDA model will also can predict side information such as labels and authors of new documents accurately. Based on different ways of adding side information, there are mainly two kinds of supervised topic model, namely, downstrean supervised topic model(DSTM) and upstream supervised topic model(USTM)[7].

A. Downstream Supervised Topic Model

In a DSTM the response variable is predicted based on the latent representation of the document. Models such as

doi:10.4304/jsw.9.10.2749-2757

sLDA(supervised Latent Dirichlet Allocation)[1] and MedLDA(Maximun Margin Supervised Topic Model)[8] are examples of DSTMs. In sLDA model proposed by blei et al., response variable is generated by learning the parameters of a generalized linear model(GLM) with an appropriate link function and exponential family dispersion function, which are specified by the modeler. Based on the experiment, blei et al. illustrate the benefits of sLDA versus modern regularized regression, as well as verus an unsupervised LDA analysis followed by a separate regression. Zhu et al.[8] propose the MedLDA model which is a combination of max-margin prediction models and hierarchical Bayesian topic models. Zhu et al. show that MedLDA model can get more sparse and highly discriminative topical representtations and achieve state of the art prediction performance. Structure of DSTMs are shown in Fig. 1.

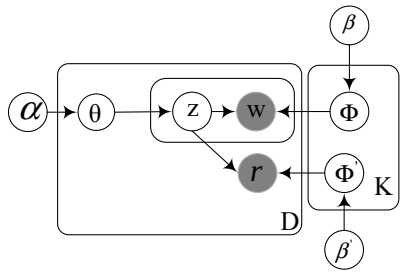


Figure 1. Structure of downstream supervised topic model.

B. Upstream Supervised Topic Model

In an USTM the response variable is being conditioned on to generate the latent representation of the documents.

Different from DSTMs, USTMs which are more close to the actual documents generative process and human vision have been widely used in many applications. For example, Ramage et al.[9] implements supervised LDA model by mapping labels of documents into several topics. when predict labels for new documents, the prediction is made by a combination of topics; Mimno et al.[10] proposed Dirichlet-multinomial regression(DMR) topic model, by adding Dirichlet-multinomial distribution into the prior of document-topic distribution. Experimental results show that in contrast to previous methods, DMR model can be able to incorporate arbitary types of observed continuous, discrete and categorical features with no additional coding, yet inference remains relatively simple. Structure of USTMs are shown in Fig. 2.

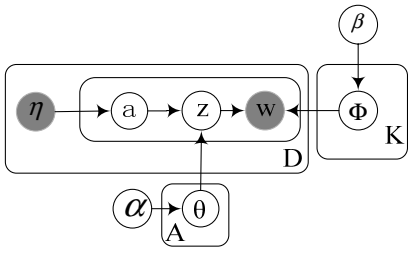


Figure 2. Structure of upstream supervised topic model.

C. Problems Definition

Although DSTMs and USTMs play an important role in the transformation of standard LDA into supervised or semi-supervised model, these two methods have ignored some key problems which make models have lower generalization power. First, documents data may have a variety of side information like authors, references, publication venue and so on. Classical DSTMs or USTMs which only process one kind of those information, can not fit dataset well. Second, different kinds of information may be generated by different methods. For example, for authors of documents a USTM method may be suitable, since authors are usually generated by some prior information instead of topics. However, for references of documents which are generated together with words, a DSTM method is needed to model their generative process.

Constructing model based on the actual documents and side information generative process not only make the model have high performance of topic extraction but also could extend application of LDA model in task of organize, index and browse large collection of documents[11]. In order to better understand the problem, we take the following description for examples. Assuming that we have a document with 9 words denoted by vector d (“bayesian”, “documents”, “topic”, “algorithms”, “gibbs”, “inference”, “approximation”, “propagation”, “parameters”). Assuming that we set the number of topics to 3. Then, for standard LDA model, it can only use the text information to extract topic, so topic 1 may contain words “documents”, “topic” which is about documents and topics. The topic 2 may contain words “bayesian”, “propagation”}, “gibbs”, “inference” which is about bayesian inference. The topic 3 may contain words “approximation”, “parameters”, “algorithms” which is about optimization algorithms. Topics generated by standard LDA model are shown in Fig. 3.

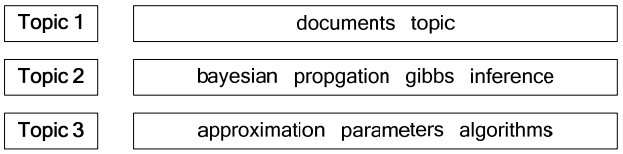


Figure 3. Topics generated by standard LDA model

In contrast, for the DMR model which can make use of authors information, assuming that we know authors of document d are a (“David Blei”, “Andrew Ng”, “Michael Jordan”), it may generate topic 1 with words “bayesian”, “topic”, “documents”, “gibbs”, “inference” which indicates Blei and topic models, topic 2 with words “approximation”, “propgation” which indicates Jordan and variational methods, topic 3 with words “algorithms”, “parameters” which indicates Andrew Ng and optimization algorithms. Topics generated by DMR model are shown in Fig. 4.

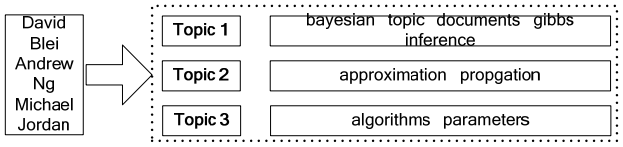


Figure 4. Topics generated by DMR model.

As shown in Fig 5, assuming that document d have both vectors a and references vector r ("variational methods", "Graphical model", "mean field method", "expectation propagatio"), then for models which can make full use of both authors and references, may generate topic 1 with words "documents"', "topic", "gibbs" which indicates Blei and topic models, topic 2 with words "approximation", "propgation", "bayesian", "inference" which indicates Jordan and bayesian variational inference, topic 3 with words "algorithms", "parameters" the same as in DMR model.

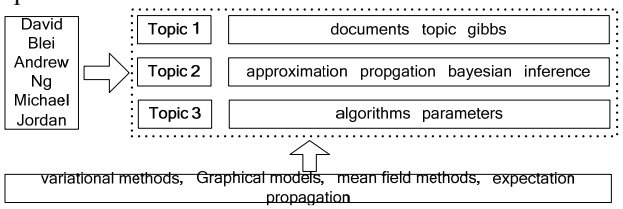


Figure 5. Topics generated by model which can analyze documents with both authors and references information.

Based on the above description, this paper proposes a new LDA model which is a combination of DSTM and USTM. Different from classical LDA model, ART model can analyze documents with both authors and references. In addition, ART model adds different information in different ways which may fit actual generation process well and outperform other DSTMs and USTMs in task of documents classification and clustering.

III. ART Model and Inference

During the document generative process, the structure of the model determines ways of information generation. This paper proposes a new LDA model in which different information are generated by different ways(DSTM or USTM).

A. Document Generate Process

In this model we make the following assumptions:

(1)When analyze documents we only take into account three kinds of information, text, authors and references.

(2)Assume that authors are generated by upstream way and references are generated by upstream way.

Authors distribution of ART model are generated by the same way as DMR model. Let vector $\boldsymbol{x}_d$ denote the authors of document d where the length of $\boldsymbol{x}_d$ is equal to the length of authors list L. If document d contains the l author then the l element of $\boldsymbol{x}_d$ is set to l, otherwise 0. Each topic t has a vector $\boldsymbol{\lambda}_t$ which is generated by the gauss distribution $N(0,\sigma^2 I)$. During the topics generative process, first, each topic t draws $\boldsymbol{\lambda}_t$ from $N(0,\sigma^2 I)$, second, draws words distribution and references distribution from dirichlet distribution $D(\beta)$ and $D(\beta')$. For each document d, draws dirichlet prior $\boldsymbol{\alpha}_d$ from function $\alpha_{dt} = exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t)$, then draw $\boldsymbol{\theta}_d$ from dirichlet prior $D(\boldsymbol{\alpha}_d)$. At last, for each word w_i and reference r_j, draw topic z_i and w_i from multinomial $M(\boldsymbol{\theta}_d)$ and $M(\phi_{z_i})$, draw topic z_j and reference r_j from multinomial $M(\boldsymbol{\theta}_d)$ and $M(\phi_{z_j'})$. Table I shows the generative story of ART model. The graphical illustration of ART model is shown in Fig. 6.

TABLE I
THE GENERATIVE STORY OF ART MODEL

- For each topic t.
  - Draw $\boldsymbol{\lambda}_t \sim N(0,\sigma^2 I)$.;
  - Draw $\phi_t \sim D(\beta)$.;
  - Draw $\phi_t' \sim D(\beta')$.;
- For each documents d
  - For each topic t let $\alpha_{dt} = exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t)$.;
  - Draw $\boldsymbol{\theta}_d \sim D(\boldsymbol{\alpha}_d)$.;
  - For each word i
    Draw $z_i \sim M(\boldsymbol{\theta}_d)$.;
    Draw $w_i \sim M(\phi_{z_i})$.;
  - For each reference j
    Draw $z_j' \sim M(\boldsymbol{\theta}_d)$;
    Draw $r_j \sim M(\phi_{z_j'})$;

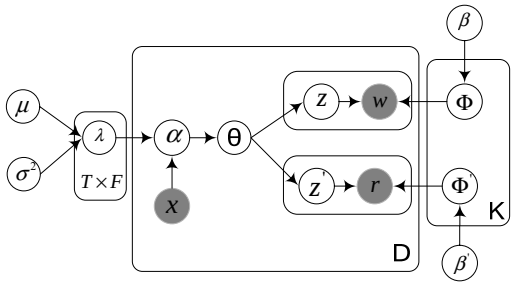


Figure 6. The graphical illustration of ART model.

B. Inference

For probabilistic models, the most frequently-used optimization method is maximum likelihood estimation(MLE). From the generative story of ART model we know that the complete likelihood is $P(z,z',\boldsymbol{\lambda})$. Using the independence assumptions of ART model, we know that $P(z \mid \boldsymbol{\lambda})$ and $P(z' \mid \boldsymbol{\lambda})$ are independent each other conditioned on $\boldsymbol{\lambda}$, so we get

$$P(z,z',\boldsymbol{\lambda}) = P(z,z' \mid \boldsymbol{\lambda})P(\boldsymbol{\lambda}) = P(z \mid \boldsymbol{\lambda})P(z' \mid \boldsymbol{\lambda})P(\boldsymbol{\lambda}) \quad (1)$$

Based on the method used in DMR model, we integrate over the multinomials $\boldsymbol{\theta}$ and get

$$P(z \mid \boldsymbol{\lambda}) = \prod_d \left(\frac{\Gamma(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t))}{\Gamma(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nw_d)} \prod_t \left(\frac{\Gamma(\exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nw_{t|d} + nr_{t|d})}{\Gamma(\exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t))} \right) \right) \tag{2}$$

$$P(z' \mid \boldsymbol{\lambda}) = \prod_d \left(\frac{\Gamma(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t))}{\Gamma(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nr_d)} \prod_t \left(\frac{\Gamma(\exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nw_{t|d} + nr_{t|d})}{\Gamma(\exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t))} \right) \right) \tag{3}$$

$$P(\boldsymbol{\lambda}) = \prod_{t,k} \left(\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\lambda_{tk}^2}{2\sigma^2} \right) \right) \tag{4}$$

where nw denotes the frequent counting of words, nr denotes the frequent counting of references. From Equation (2),(3),(4) we can get equation (1). In order to maximize the complete log likelihood, we should know the derivative of the log of Equation 1 with respect to the parameter λ_{tk} for given topic t and feature k.

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda_{tk}} &= \sum_d x_{dk} \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) \times \\ &\left(\left(\Psi\left(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) \right) - \Psi\left(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nw_d \right) \right) A \right. \\ &+ \left(\Psi\left(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) \right) - \Psi\left(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nr_d \right) \right) B \\ &+ \left(\Psi\left(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nw_{t|d} \right) - \Psi\left(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) \right) \right) C \\ &+ \left. \left(\Psi\left(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nw_{t|d} \right) - \Psi\left(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) \right) \right) D \right) \\ &- \frac{\lambda_{tk}}{\sigma^2} \end{aligned} \tag{5}$$

where

$$A = \frac{\Gamma(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t))}{\Gamma(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nr_d)} \tag{6}$$

$$B = \frac{\Gamma(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t))}{\Gamma(\sum_t \exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nw_d)} \tag{7}$$

$$C = \frac{\Gamma(\exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nw_{t|d} + nr_{t|d})}{\Gamma(\exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t))} \tag{8}$$

$$D = \frac{\Gamma(\exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t) + nw_{t|d} + nr_{t|d})}{\Gamma(\exp(\mathbf{x}_d^T \boldsymbol{\lambda}_t))} \tag{9}$$

The common approximate inference algorithms of model with hidden values include EM[12], mean-field variational EM[13], expectation propagation(EP)[14] and Gibbs Sampling[15] methods based MCMC. In this paper, we use the stochastic EM sampling method to estimate the parameters of ART model. In the E-step a Gibbs Sampling method is used to sampling documents with $\boldsymbol{\lambda}$ fixed. In the M-step, given the topic assignments, we optimize the parameters $\boldsymbol{\lambda}$ using Equation 5.

IV. ART Model And Inference

A. Topics Extraction

In the experiment, we implement the ART model based on the Mallet toolkit. The E-step and M-step of our implementation is based on the LDA code and the L-BSGF module of Mallet toolkit.

In order to test the performance of topic extraction, we compare the ART model with both the standard LDA and the DMR models. During the experiment, we compute the empirical likelihood(EL) of three models and the EL is defined as follows:

$$EL(d) = \frac{1}{|S|} \sum_s \sum_i \sum_t \theta_{dts} \frac{n_{w_i|t} + \beta}{n_t + |T|\beta} \tag{10}$$

in which $|S|$ is the length of document d, $|T|$ is number of topics. We first compute the $\boldsymbol{\alpha}_d$ with fixed $\mathbf{x}_d$, then draw $\boldsymbol{\theta}_{dt}$ from dirichlet distribution, at last compute the $P(w_i \mid t)$.

We run the three models on Rexa dataset which is a corpus of research papers drawn from Rexa and contains text, publication year, publication venue, automatically disambiguated author IDs, and automatically disambiguated references. We divide the Rexa dataset into three subsets, Table II shows the detail of three subsets.

TABLE II
EXPERIMENTAL DATASETS

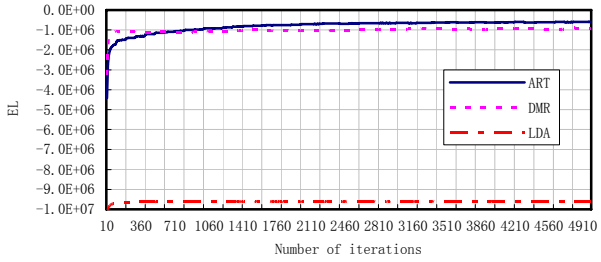
Years	Documents	Words	Authors	References
1998-2004	20284	574992	33806	190476
1999-2004	17370	514750	31845	167985
2001-2004	12090	371247	17150	122468

When use those three datasets to test models, we use 95% samples of each dataset as training samples and 5% as test samples. For hyperparameters β and β', we set them to 0.01. For the hyperparameters σ^2 of DMR and ART models, we set them to 0.1. Topics numbers of all models are set to 100 and in order to ensure models convergence set the max iterations of Gibbs Sampling to 5000. During each iteration of EM, we sample the documents 100 times at E-step and then compute the EL of model. If the change of the EL value is less than the threshold, terminate the algorithm, otherwise goto M-step. In order to make comparison more explicit, we transform each EL by equation (11), where $EL(*)$ denotes the EL that should to be transformed, $EL(LDA)$ denotes the EL of standard LDA model. Finally we get results as shown in Fig. 7.

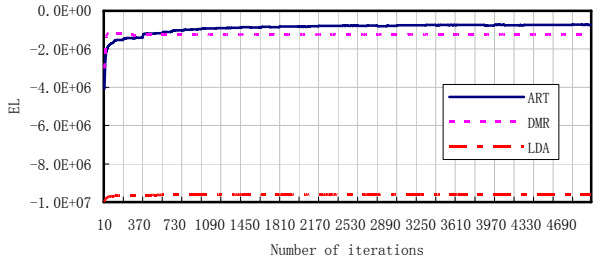
$$EL(*) = \frac{-1 \times (EL(*) - EL(LDA))}{EL(LDA)} \tag{11}$$

From Fig. 7 we can get that the standard LDA model can only make use of the text of documents, so its transfomed EL is 0 which is much lower than ART and

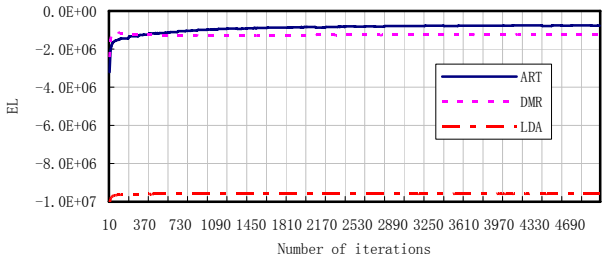
DMR model. In contrast, the DMR model which uses both text and authors information, has a higher EL value, which indicate that it has more efficient capability of topics extraction and documents clustering. The ART model outperform other two models by using not only authors but also references information. However, the convergence rate of ART model is lower than the LDA and DMR models.



a) 20284 samples



b) 17370 samples



c) 12090 samples

Figure 7: Comparison of EL value on three different datasets.

In order to illustrate the advantages of ART model directly, we list topics generated by standard LDA, DMR and ART model in Table III, IV, V and VI. As shown in Table III, standard LDA model generates topics without any prior of authors and references, which makes the model generate inexplicable topics. Comparing with standard LDA model, DMR model generates more reasonable topics. For example, topic 2 in Table IV contains words "data", "models", "clustering", "model", "algorithm" which represent the semantic meaning of "clustering model and algorithm", topic 1 in Table IV contains words "learning", "classification", "feature", "training", "data" which may indicate topic of "classification and feature learning" and there is a strong correlation between "classification" and "feature" but in standard LDA model it is not evident.

TABLE III
TOP 5 WORDS OF EACH TOPIC GENERATED BY LDA MODEL

Topics	Words
Topic#1	retrieval text information language system
Topic#2	network networks neural analysis detection
Topic#3	web digital information system knowledge
Topic#4	learning classification data abstract algorithm
Topic#5	data search web mining query large
Topic#6	model models probabilistic abstract bayesian
Topic#7	algorithm clustering data space method
Topic#8	problem abstract problems paper show
Topic#9	image images motion object objects
Topic#10	robot control planning system mobile

TABLE VI
TOP 5 WORDS OF EACH TOPIC GENERATED BY DMR MODEL

Topics	Words
Topic#1	learning classification feature training data
Topic#2	data models clustering model algorithm
Topic#3	retrieval text information language system
Topic#4	image images motion object based
Topic#5	algorithm learning problem algorithms abstract
Topic#6	model abstract information detection analysis
Topic#7	model neural network networks human
Topic#8	web user information search digital
Topic#9	knowledge abstract based reasoning logic
Topic#10	robot control system based planning

As shown in Table V and Table VI the ART topic model not only gives more reasonable topics but also finds most relevant references of topics. For example, topic 1 in Table V indicates "classification and feature learning" with references "13919877:Bagging Predictors", "18808311: Support-Vector Networks", "18215730: Inducing Features of Random Fields" which are shown in Table VI. References "17700097:Collaborative filtering via gaussian probabilistic latent semantic analysis", "12607307: Improved Algorithms for Topic Distillation in a Hyperlinked Environment", "16035499: Okapi at TREC-3" in topic 6 consists with words "retrieval", "information", "text", "web", "document" which indicate topic of "information retrieval and document analysis".

TABLE V
TOP 5 WORDS OF EACH TOPIC GENERATED BY ART MODEL

Topics	Words
Topic#1	learning classification algorithm training feature
Topic#2	logic reasoning knowledge semantics system
Topic#3	language word model system translation
Topic#4	learning search planning reinforcement problems
Topic#5	data web mining digital information
Topic#6	retrieval information text web document
Topic#7	image images camera motion shape
Topic#8	data algorithm problem model approach
Topic#9	model image models recognition tracking
Topic#10	robot control system robots mobile

B. Predicting Authors

In order to test the performance of author prediction, we run the ART model on the datasets shown in Table 2. Labeled LDA[16] and PLDA[9] model which are proposed by Ramage et al., map each document label into one or several topics and have been widely used in many applications. However, these two models do not suitable for task of author prediction. For example, there is a dataset with 10000 documents and 2000 different authors, when we use Labeled LDA or PLDA to predict authors,

we must set the number of topics to at least 2000 which will cause a higher computing cost and does not coincide with actual number of topics. Unlike Labeled LDA and PLDA models, ART and DMR models use the same dirichlet-multnomial prior added in the parameters a to avoid generating too much topics and have high performance of authors prediction. So for the experiment of author prediction, we only take into account ART and DMR models.

We predict authors of new document d by computing equation (12).

$$P(d \mid a) = \frac{\sum_t \alpha_t + n_a}{\sum_t \alpha_t + n_a + \sum_t n_t} \prod_t \frac{\alpha_t + n_{t|a} + n_t}{\alpha_t + n_{t|a}} \qquad (12)$$

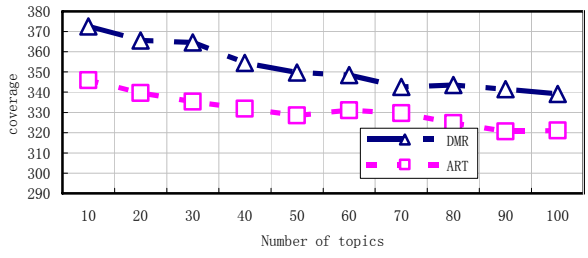
For document d, because ART and DMR models only contain topics of a subset of authors, when compute the equation (11) we just need to get the Dirichlet-multinomial distribution of these authors. Prediction of authors in LDA model belong to task of multi-label classification. At the same time, both ART and DMR models are Label Rank methods from which we can get a rank of authors. So during experiment, we use coverage to compare these two methods. Coverage is a widely used evaluation measure of multi-label classification which evaluates how far we need, on average, to go down the ranked list of labels in order to cover all the relevant labels of the example[17]. It is defined as follows:

$$coverage = \frac{1}{p}\sum_{i=1}^{p} \max_{y \in Y_i} rank_f(x_i, y) - 1 \qquad (13)$$

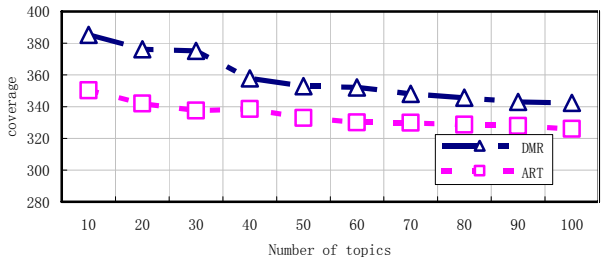
where p is the number of documents, Y_i is the set of authors given document i .Assume $f(\cdot,\cdot)$ denote the rank function of authors, then $rank_f(x_i, y)$ is a function which meet the condition that if $f(x_i, y_2) > f(x_i, y_2)$ then $rank_f(x_i, y_1) < rank_f(x_i, y_2)$.

The hyperparameters are set the same as Section IV.A. The performance of ART and DMR models under datasets of Table II is shown in Fig. 8.

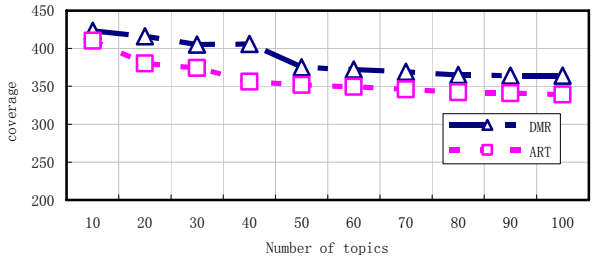
From Fig. 8 we can get that with the number of topics increase, coverage of both ART and DMR model decrease. But coverage of ART model is significantly lower then DMR model. When the number of topics is close to 100, two curves become smooth which indicate that ART and DMR models converge near 100. Based on the above results, we could find that the ART model have more accurate performance of authors prediction that DMR model.



a) 20284 samples



b) 17370 samples



c) 12090 samples

Figure 8. Comparison of coverage on three different datasets.

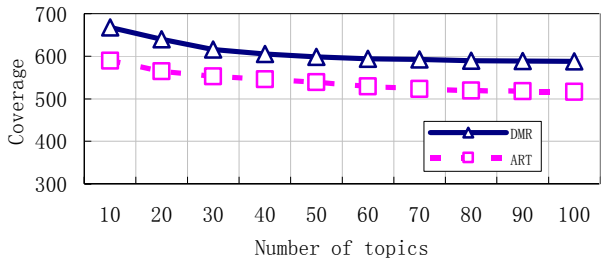
C. Citations Ranking

In the academic network search, predicting the influence of a academic paper is an difficult and important problem[18]. Based on the influence of references, academic search system could recommend the most influential documents to users, which have significant impact on good academic research[19]. By using the references information the ART model could detect the most influential references included in each topic.

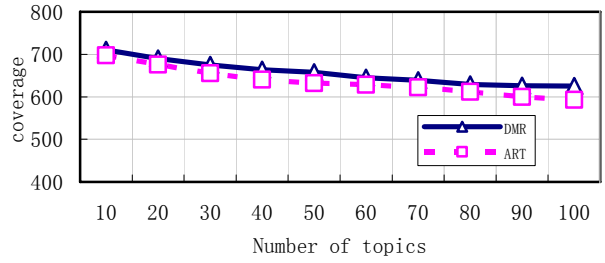
As shown in [10], the DMR could also compute the documents influence by treating each reference as a potential "author". In order to test the performance of ART and DMR models in task of measurement the influence of documents, we run the algorithm on the three datasets shown in Table 2. Since the ART model can get properties of references belong to each topic directly, what we should do is to transform those properties into the value measure the influence. After the ART model converge, we transform the properties by using equation 14, where $influence(r \mid d)$ denotes the influence of reference r on document d . We can use the influence to rank the references of document d . In order to give a exact comparison, we use coverage as measurement to evaluate these two models. The experimental results are shown in Fig. 9.

$$influence(r \mid d) = \sum_{t=1}^{T} p(r \mid t)\, p(t \mid d) \qquad (14)$$

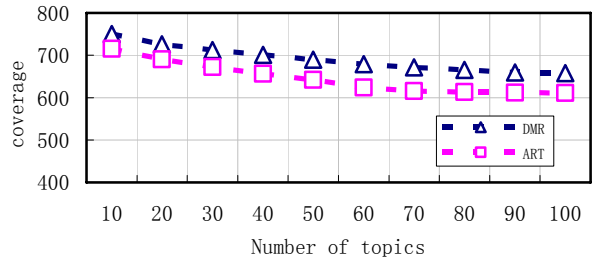
As shown in Fig. 9, at first, coverage of both ART and DMR model decrease with the increasing of topics number. However, since the ART model can make use of authors information, its coverage is much lower than the DMR model, which indicates that ART have a more accurate performance than DMR model. Especially, when the dataset is large, more obvious experimental results are achieved.



a) 20284 samples



b) 17370 samples



c) 12090 samples

Figure 9. Comparison of performance in task of citation ranking.

V. CONCLUSIONS

As the development of probabilistic topic models, they have been widely used in academic documents analysis and information retrieval[20]. Many excellent models have been proposed, include Author-Topic model(AT)[21], Author-Conference Topic model(ACT) [22] and so on. These models lay a foundation for applications of probabilistic topic model in academic documents analysis and academic search. Based on the study on the applications of probabilistic topic model in academic search, this paper proposes a new LDA model which is a combination of DSTM and USTM. Experimental results show that the new model has the following advantages:

(1) Based on the combination of DSTM and USTM, ART model can analyze documents with two kinds of side information efficiently, at the same time, can also improve the performance of document clustering and topics extraction.

(2) For different side information included in academic documents, we use different generative process which could fit the actual documents generative process well and beat other LDA models in task of documents analysis.

(3) In the process of authors generation, we use a Dirichlet-multinomial as the prior of Dirichlet prior α that makes the number of topics do not increase with the authors. So compared with the Labeled LDA and PLDA model, the ART model has a lower computational cost.

During the process of model expansion, we found that documents especially academic documents have not only authors and references information, but also other side information such as publication year, publication venue and so on. Making full use of those kinds of information could comprehensively improve the performance of labels discrimination, author prediction. Furthermore, one of our future work is to modify the standard LDA model and make it could analyze documents with a variety of side information.

TABLE VI
TOP 3 REFERENCES OF EACH TOPIC GENERATED BY ART MODEL

Topics	References
Topic#1	13919877: Bagging Predictors. 18808311: Support-Vector Networks. 18215730: Inducing Features of RandomFields.
Topic#2	17713561: The Frame Problem and Knowledge-Producing Actions. 16867584: Causal Theories of Action and Change. 13877598: A Theory of Diagnosis from First Principles.
Topic#3	17939911: Three Generative, Lexicalised Models for Statistical Parsing. 16105479: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. 10031621: Maximum Entropy Markov Models for Information Extraction and Segmentation.
Topic#4	11163413: Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. 19071554: Fast Planning Through Planning Graph Analysis. 12909664: Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning.
Topic#5	9796741: Data extraction and label assignment for web databases. 9854600: Large-scale mining of usage data on web sites. 10279798: Scalable robust covariance and correlation estimates for data mining.
Topic#6	17700097: Collaborative filtering via gaussian probabilistic latent semantic analysis. 12607307: Improved Algorithms for Topic Distillation in a Hyperlinked Environment. 16035499: Okapi at TREC-3.
Topic#7	11427674: An Iterative Image Registration Technique with an Application to Stereo Vision. 17899098: Self-Calibration and Metric Reconstruction in Spite of Varying and Unknown Internal Camera Parameters. 11978148: Hierarchical Model-Based Motion Estimation.
Topic#8	17513630: Wrapper Induction for Information Extraction. 9786666: Learning to map between ontologies on the semantic web. 11468202: Wrapper induction: Efficiency and expressiveness.
Topic#9	9997923: Neural Network-Based Face Detection. 13639954: EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. 10826776: Efficient Region Tracking With Parametric Models of Geometry and Illumination.
Topic#10	13939339: Probabilistic Robot Navigation in Partially Observable Environments. 12602455: Randomized Preprocessing of Configuration Space for Fast Path Planning. 10467137: Time-optimal cooperative control of multiple robot vehicles.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their helpful comments. This work was supported by the Natural Science Foundation of Anhui Provincial Education Department under Grant No.KJ2011Z039, No.KJ2013A053

REFERENCES

[1] Blei, David M., and Jon D. McAuliffe. "Supervised topic models." arXiv preprint arXiv:1003.0783 (2010).
[2] Toutanova, Kristina, and Mark Johnson. "A Bayesian LDA-based model for semi-supervised part-of-speech tagging." Advances in Neural Information Processing Systems. 2007.
[3] Zhou, Ding, et al. "Exploring social annotations for information retrieval."Proceedings of the 17th international conference on World Wide Web. ACM, 2008.
[4] Zhao, Wayne Xin, et al. "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid." Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.
[5] Wang, Jianwen, et al. "Author-conference topic-connection model for academic network search." Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.
[6] Diebolt, J., and E. H. S. Ip. "A stochastic EM algorithm for approximating the maximum likelihood estimate." Markov chain Monte Carlo in practice (1996).
[7] Zhu, Jun, et al. "Large margin learning of upstream scene understanding models." Advances in Neural Information Processing Systems. 2010.
[8] Zhu, Jun, Amr Ahmed, and Eric P. Xing. "Medlda: maximum margin supervised topic models." Journal of Machine Learning Research 13 (2012): 2237-2278.
[9] Ramage, Daniel, Christopher D. Manning, and Susan Dumais. "Partially labeled topic models for interpretable text mining." Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011.
[10] Mimno, David, and Andrew McCallum. "Topic models conditioned on arbitrary features with dirichlet-multinomial regression." arXiv preprint arXiv:1206.3278(2012).
[11] Kim, Hyungsul, et al. "ETM: Entity Topic Models for Mining Documents Associated with Entities." Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012.
[12] Hofmann, Thomas. "Probabilistic latent semantic indexing." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.
[13] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.
[14] Minka, Thomas, and John Lafferty. "Expectation-propagation for the generative aspect model." Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 2002.
[15] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics."Proceedings of the National academy of Sciences of the United States of America 101.Suppl 1 (2004): 5228-5235.
[16] Ramage, Daniel, et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing:

Volume 1-Volume 1. Association for Computational Linguistics, 2009.
[17] Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." Pattern Recognition 40.7 (2007): 2038-2048.
[18] Gerrish, Sean, and David M. Blei. "A language-based approach to measuring scholarly impact." Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010.
[19] Wang, Lidong, Baogang Wei, and Jie Yuan. "Topic Discovery based on LDA_col Model and Topic Significance Re-ranking." Journal of Computers6.8 (2011): 1639-1647.
[20] Xie, Ming, Chanle Wu, and Yunlu Zhang. "A New Intelligent Topic Extraction Model on Web." Journal of Computers 6.3 (2011): 466-473.
[21] Rosen-Zvi, Michal, et al. "The author-topic model for authors and documents." Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004.
[22] Tang, Jie, Ruoming Jin, and Jing Zhang. "A topic modeling approach and its integration into the random walk framework for academic search." Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008.

Yuyan Jiang is a professor of the Department of Management Science and Engineering, Anhui University of technology, Ma'anshan, China.
Her current research interests include Machine learning, Data Mining.
(Jyy@ahut.edu.cn)

Shao Yuan is a undergraduate student of school of Management Science and Engineering, Anhui University of technology, Ma'anshan, China.
His current research interests include Machine learning, Data Mining.

Ping Li is a master student of school of Management Science and Engineering, Anhui University of technology, Ma'anshan, China.
His current research interests include Machine learning, Data Mining.

Qing Wang is a professor of the Department of Management Science and Engineering, Anhui University of technology, Ma'anshan, China.
His current research interests include Machine learning, Data Mining, Recommendation System.