# Semantic Social Network Analysis with Text Corpora

Dong-mei Yang[1], Hui Zheng[1], Ji-kun Yan[2], and Ye Jin[2]

[1] Science and Technology on Blind Signal Processing Laboratory
ydm.123@foxmail.com, Zheng5739@163.com
[2] Southwest Electronics and Telecommunication Technology Research Institute
yanjk@126.com, djws_1982@sina.com

**Abstract.** We present the Document-Entity-Topic (DET) model for semantic social network analysis which tries to find out the interested entities through the topics we aim at, detect groups according to the entities which concern the similar topics, and rank the plentiful entities in a document to figure out the most valuable ones. DET model learns the topic distributions by the literal descriptions of entities. The model is similar to Author-Topic (AT) model, adding the key attribute that the distribution of entities in a document is not uniform but Dirichlet allocation. We experiment on the "Libya Event" data set which is collected from the Internet. DET model increases the precision on tasks of social network analysis and gives much lower perplexity than AT model.

**Keywords:** Semantic Social Network Analysis, Topic Model, Entity Modeling.

## 1 Introduction

As far as we know, topic modeling has become a most popular technology to model large collection of corpus[1-3], such as Latent Dirichlet Allocation[4]. The basic idea of topic modeling is that the latent topics can be used to describe the relationship between words and documents. In this paper we consider the problem of using latent topics to connect the words and entities in documents (such as person, location, organization). We focus on the news articles which contain lots of entities in order to convey the information about who, what, when and where. The purpose we want most is modeling the entities in terms of latent topics so that we can 1) find out the interested entities through the topics we aim at; 2)recognize groups with supposing that the entities (especially the persons) which concern the similar topics can be seen as a group; 3) rank the plentiful entities in a document to figure out the valuable ones by assuming that the more an entity contributes to a document's topic(s), the more valuable it is in the precise one. We call the three tasks Semantic Social Network Analysis for the interactions been found based on the topics of the corpus.

There are several related researches to achieve the relationship between words and entities (authors) with topic models. The Author-Topic (AT) model[5-6] learns the topics of a document conditioned on the mixture of interests with the authors. AT model assumes that the authors equally contribute to the topics of a document. The SwitchLDA and GESwitchLDA[7-8] extend LDA to capture dependencies between entities and topics, referring to entities as additional classes of words.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 493–504, 2012.

This paper presents the Document-Entity-Topic (DET) model, a directed graphical model by assuming that words were generated by the entities of the document. The model is similar to the AT model. However, it is not limited to the topic finding of authors, but tries to modeling topics of all related entities in the documents. For this application, we confront more unwanted entities of the corpus. In our experiments, there are more than five person entities in most documents, and some entities such as news reporters have little significance to the topic(s). If all entities in a document have been assumed to be equally contributed to the mixture of topics as AT model, it is not enough for us to rank the importance of entities and many noisy entities will disturb the topic modeling of corpus. So our DET model presumes that the entities have different topical contributions to their document. We use the Dirichlet allocation to describe the distribution between document and its entities; a document gives higher probabilities to several more valuable entities (not all entities) and valuable entities have more contributes to the topic modeling.

The outline of the paper is as follows: Section 2 describes the Document-Entity-Topic model, and section 3 outlines how to learn the parameters from the documents. Section 4 discusses the application of the model to the data set we collected from the internet. Section 5 contains a brief discussion and concluding comments.

## 2 Document-Entity-Topic Model

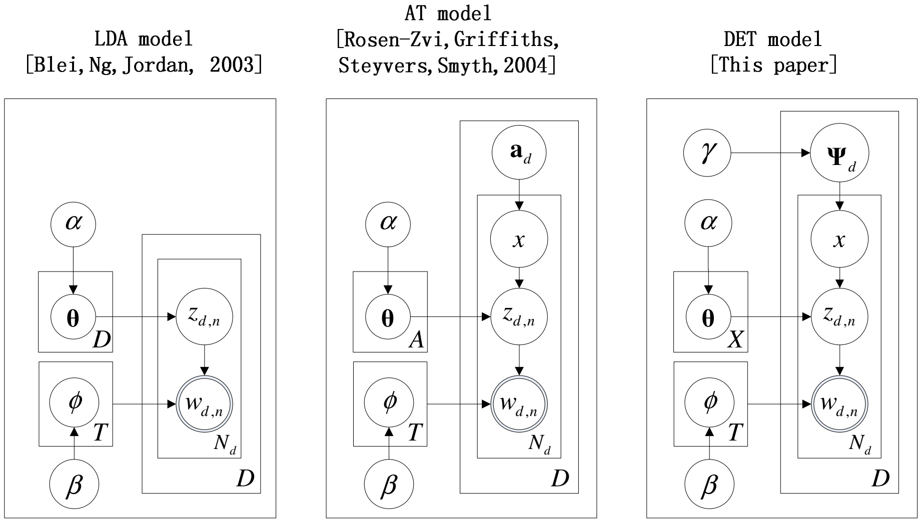
In this section we introduce the document-Entity-topic (DET) model. The DET model belongs to the family of generative models, in which each word w in a document is associated with two latent variables: an entity assignment $x$ , and a topic assignment $z$ .

### 2.1 Dirichlet Priori on Document-Entity Distribution

The entities in news may have different weights to be described by the words, for example, "a reporter covers that, Mr. A and B contact at a national conference and have an educational barging, trying to improve the intercommunion and friendship of both." We find that the relationship between A and B is closer than that between reporter and the two people. In order to discover the different weights for different entities, we use Dirichlet allocation as the prior distribution to describe the importance of each entity in a document, which is similar to LDA model by using Dirichlet allocation to describe the relationship between topics and the document.

The reason to choose the Dirichlet is that, firstly, it can reflect the characteristic of document-entity relation, a document has primary and minor entities, and the weights can be adjusted by the hyperparameter of Dirichlet. Secondly, the conjugate prior of multinomial distribution is Dirichlet allocation, so it can simplify the computation for the posterior distribution which has the same functional form as the prior.

Thus, we propose the Document-Entity-Topic (DET) model for mining the semantic description of entities and using the topical distribution to carry out the social networking tasks. The generative process of DET model for a document can be summarized as follows: firstly, an entity is chosen randomly from the distribution over entity-document; next, a topic label is sampled for each word from the distribution over topics associated with the entities of that word; finally, the words are sampled from the distribution over words associated with each topic. The plate representation[9] for all models are shown in figure 1.


LDA model
[Blei,Ng,Jordan, 2003]

AT model
[Rosen-Zvi,Griffiths,
Steyvers,Smyth,2004]

DET model
[This paper]

**Fig. 1.** Two related models and the DET model. In all models, each word *w* is generated from a topic-specific multinomial word distribution; however topics are sampled differently in each of the models. In LDA, a topic is sampled from a document-specific topic distribution which is sampled from a Dirichlet with hyperparameter. In the AT model, a topic is sampled from an author-specific multinomial distribution, and authors are sampled uniformly from the document's author set. In DET, Dirichlet prior has been introduced to the document-entity distribution, a topic is sampled from an entity-specific multinomial distribution, and entity assignment is sampled from the Dirichlet allocation of that document.

## 2.2 Generative Process of DET Model

In DET model, the generative process of generating a word is according to the probability distributions of firstly picking an entity followed by picking a topic.

a) For each document $d = 1,\ldots,D$ choose $\boldsymbol{\psi}_d \sim Dirichlet(\gamma)$;

For each entity $x = 1,\ldots,X$ choose $\boldsymbol{\theta}_x \sim Dirichlet(\alpha)$;

For each topic $t = 1,\ldots,T$ choose $\phi_t \sim Dirichlet(\beta)$.

b) For each document $d = 1,\ldots,D$

Given the vector of entities $\boldsymbol{X}_d$, for each word $w_i$, out of the $N_d$ words

Conditioned on $\mathbf{x}_d$ choose a persona $x_i \sim Dirichlet(\boldsymbol{\psi}_d)$;

Conditioned on $x_i$ choose a topic $z_i \sim Dirichlet(\boldsymbol{\theta}_{x_i})$;

Conditioned on zi choose a word $w_i \sim Dirichlet(\phi_{z_i})$.

Under this generative process, each entity is drawn independently conditioned on $\boldsymbol{\psi}$ ; each topic is drawn independently conditioned $\boldsymbol{\Theta}$ ; and each word is drawn independently conditioned on $\boldsymbol{\Phi}$ and $z$. The probability of the corpus $\mathbf{w}$, conditioned on $\boldsymbol{\psi}$ , $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ is shown as equation (1):

$$P(\mathbf{w}|\boldsymbol{\Theta},\boldsymbol{\Phi},\boldsymbol{\Psi}) = \prod_{d=1}^{D} P(\mathbf{w}_d|\boldsymbol{\Theta},\boldsymbol{\Phi},\boldsymbol{\Psi}) \tag{1}$$

Summing over the latent variables $x$ and $z$, we can obtain the probability of the words in each document $\mathbf{w}_d$ as equation (2):

$$\begin{aligned} P(\mathbf{w}_d \mid \boldsymbol{\Theta},\boldsymbol{\Phi},\boldsymbol{\Psi}) &= \prod_{i=1}^{N_d} P(\mathbf{w}_i \mid \boldsymbol{\Theta},\boldsymbol{\Phi},\boldsymbol{\Psi}) \\ &= \prod_{i=1}^{N_d}\sum_{x=1}^{X_d}\sum_{t=1}^{T} P(w_i, z_i = t, x_i = x \mid \boldsymbol{\Theta},\boldsymbol{\Phi},\boldsymbol{\Psi}) \\ &= \prod_{i=1}^{N_d}\sum_{x=1}^{X_d}\sum_{t=1}^{T} P(w_i \mid z_i = t,\boldsymbol{\Phi}) P(z_i = t \mid x_i = x,\boldsymbol{\Theta}) P(x_i = x \mid \boldsymbol{\Psi}) \\ &= \prod_{i=1}^{N_d}\sum_{x\in X_d} \psi_{x_d} \cdot \sum_{t=1}^{T} \theta_{xt} \cdot \phi_{w_i t} \end{aligned} \tag{2}$$

Factorizing in the third line of equation (2) uses the conditional independence assumptions of the model. The last line in the equations expresses the probability of the words $w$ in terms of the parameter matrices $\boldsymbol{\psi}$ , $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ . $P(x_i = x \mid \boldsymbol{\Psi})$ is the entity multinomial distribution $\boldsymbol{\psi}_d$ in $\boldsymbol{\Psi}$ which corresponds to document $d$, $P(z_i = t \mid x_i = x,\boldsymbol{\Theta})$ is the multinomial distribution $\boldsymbol{\theta}_x$ in $\boldsymbol{\Theta}$ that corresponds to entity $x$, and $P(w_i \mid z_i = t,\boldsymbol{\Phi})$ is the multinomial distribution $\phi_t$ in $\boldsymbol{\Phi}$ corresponding to topic $t$.

# 3 Learning the DET Model from Data

The DET model contains three continuous random variables $\boldsymbol{\Psi}$ , $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ . The inference scheme used in this paper is based upon a Markov chain Monte Carlo (MCMC) algorithm or more specifically, Gibbs sampling. We estimate the posterior distribution $P(\boldsymbol{\Psi},\boldsymbol{\Theta},\boldsymbol{\Phi} \mid D^{train},\gamma,\alpha,\beta)$ . The inference scheme is based upon the observation that

$$\begin{aligned} &P(\boldsymbol{\Psi},\boldsymbol{\Theta},\boldsymbol{\Phi} \mid D^{train},\gamma,\alpha,\beta) \\ &= \sum_{z,x} P(\boldsymbol{\Psi},\boldsymbol{\Theta},\boldsymbol{\Phi} \mid z,x,D^{train},\gamma,\alpha,\beta) P(z,x \mid D^{train},\gamma,\alpha,\beta) \end{aligned} \tag{3}$$

Where $z$ is the topic variable and $x$ is the entity assignment. This inference process involves two steps. Firstly, we use Gibbs sampling to obtain an empirical sample-based estimate of $P\left(z,x \mid D^{train},\gamma,\alpha,\beta\right)$. Second, we compute each specific sample corresponding to particular $x$ and $z$ using the conjugation trait between Dirichlet and multinomial distribution.

### 3.1 Gibbs Sampling

Gibbs sampling is a widely applicable Markov chain Monte Carlo algorithm which can be viewed as a special case of Metropolis Hastings algorithm. It often yields relatively simple algorithms for approximate inference in high-dimensional models such as topic models[9]. Here we wish to construct a Markov chain which converges to the posterior distribution over $x$ and $z$ in terms of $D^{train},\gamma,\alpha$ and $\beta$. Using Gibbs sampling we can generate a sample from $P\left(z,x \mid D^{train},\gamma,\alpha,\beta\right)$ by firstly sampling an entity assignment $x_i$ and a topic assignment $z_i$ for an individual word $w_i$ conditioned on initialized assignments of entities and topics for all other words in the corpus. Secondly, repeating this process for each word. A single Gibbs sampling iteration consists of sequentially performing sampling of entity and topic assignments for each individual word in the corpus.

$P\left(z_i=t \mid \mathbf{z}_{-i},\mathbf{x},D^{train},\gamma,\alpha,\beta\right)$ and $P\left(x_i=x \mid \mathbf{x}_{-i},\mathbf{z},D^{train},\gamma,\alpha,\beta\right)$ can also be the Gibbs sampler. In this paper we use the blocked sampler where we sample $x_i$ and $z_i$ jointly. It can improve the mixing time of the sampler and the method also has been used similarly by Rosen-Zvi et al[5]. In Appendix, we derive the Gibbs sampler of document $d$ and entity $x \in X_d$ as equation (4)

$$P\left(x_i=x,z_i=t \mid w_i=w,\mathbf{x}_{-i},\mathbf{z}_{-i},\mathbf{w}_{-i},\gamma,\alpha,\beta\right)$$
$$\propto \frac{C^{WT}_{wt,-i}+\beta}{\sum_{w'}C^{WT}_{w't,-i}+W\beta}\cdot\frac{C^{TX}_{tx,-i}+\alpha}{\sum_{t'}C^{TX}_{t'x,-i}+T\alpha}\cdot\frac{C^{XD}_{xd,-i}+\gamma}{\sum_{x'}C^{XD}_{x'd,-i}+X\gamma} \tag{4}$$

Here $C^{XD}$ represents the document-entity count matrix, where $C^{XD}_{xd,-i}$ is the number of words assigned to entity $x$ for document $d$ excluding word $w_i$. $C^{TX}$ is the topic-entity count matrix, where $C^{TX}_{tx,-i}$ is the number of words assigned to topic $t$ for entity $x$ excluding the topic assignment to word $w_i$. $C^{WT}$ is the number of words from the $w^{th}$ entry in the vocabulary assigned to topic $t$ excluding the topic assignment to word $w_i$. Finally, $z_{-i}$, $x_{-i}$ , $w_{-i}$ stand for the vector of topic assignments, vector of entity assignments and vector of word observations in corpus except for the $i^{th}$ word, respectively.

### 3.2 The Posterior on $\mathbf{\Psi}$ , $\mathbf{\Theta}$ and $\mathbf{\Phi}$

Given $\mathbf{z},\mathbf{x},D^{train},\gamma,\alpha$ and $\beta$ , we can compute the posterior distributions on $\mathbf{\Psi}$ , $\mathbf{\Theta}$ and $\mathbf{\Phi}$ directly. Using the fact that the Dirichlet is conjugate to the multinomial, we have

$$\phi_{w,t}=\frac{C^{WT}_{wt,-i}+\beta}{\sum_{w'}C^{WT}_{w't,-i}+V\beta},\theta_{t,x}=\frac{C^{TX}_{tx,-i}+\alpha}{\sum_{t'}C^{TX}_{t'x,-i}+T\alpha},\psi_{x,d}=\frac{C^{XD}_{xd,-i}+\gamma}{\sum_{x'}C^{XD}_{x'd,-i}+X\gamma} \tag{5}$$

These posteriors provide point estimates for $\mathbf{\Psi}$ , $\mathbf{\Theta}$ and $\mathbf{\Phi}$ . $\mathbf{\Psi}$ corresponds to the posterior predictive distribution for the documents and entities, it obeys the Dirichlet allocation other than uniform distribution, and can get the more valuable entities who effect the topics of the document more. $\mathbf{\Theta}$ corresponds to the posterior predictive distribution for the entities and topics, every entity has a vector of topics, it can tell us what topics the entity associates with and which entities are interested in the similar topics, so groups can be extracted from $\mathbf{\Theta}$ . $\mathbf{\Phi}$ corresponds to the posterior predictive distribution for the topics and words, we can get the word description of topics.

## 4 Experiment Result

We train our DET model on the "Libya Event" dataset which is collected from Internet (http://www.ifeng.com). It contains $D=4165$ documents, $P=3784$ unique entities (most are person names), $N=782043$ tokens, and a vocabulary of $V=15812$ unique words. We preprocess the document set with tryout edition of ICTCLAS whose rights reserved by ictclas.org. All documents are written in Chinese, and we translate the results in English.

We run the Markov chain for a fixed number of 2000 iterations. Furthermore, we find that the sensitivity to hyperparameters is not very strong, so that we use the fixed symmetric Dirichlet distributions $\gamma=0.5,\alpha=0.1$, and $\beta=0.01$ in all our experiments. In the comparing experiment of AT model, the author set **a** are entities extracted from the documents.

### 4.1 Perplexity Comparison between AT and DET

Models for natural languages are often evaluated by perplexity as a measure of the goodness fit of models. The lower perplexity a model has, the better it predicts the unseen words. The perplexity of a previously unseen document *d* consisting of words $\mathbf{w}_d$ can be defined as equation (6) when the entities $x_d$ are given:

$$Perplexity(\mathbf{w}_d)=\exp\left(-\frac{\log\left(p(\mathbf{w}_d\mid\mathbf{x}_d)\right)}{N_d}\right) \tag{6}$$

in which

$$p\left(\mathbf{w}_d\mid\mathbf{x}_d\right)=\prod_{i=1}^{N_d}\left(\sum_{x\in\mathbf{x}_d}\hat{\mathbf{\psi}}_x\cdot\sum_{t=1}^{T}\hat{\phi}_{w_it}\cdot\hat{\mathbf{\theta}}_{tx}\right)^{n_d^{(i)}} \tag{7}$$

Where $n_d^{(i)}$ is the number of times token *i* has been observed in document *d*. $\vec{\phi}_{w_i}$ can be determined by the training set, but $\mathbf{\theta}_x$ and $\mathbf{\psi}_d$ need to be derived by querying the model. Firstly, initializing the algorithm by randomly assigning topics and entities to words of the test documents, and then performing a number of loops through the Gibbs sampling update:

$$p\left(\tilde{z}_i = t, \tilde{x}_i = x \mid \tilde{w}_i = w, \mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w}_{-i}; \mathrm{M}\right) \propto \varphi_{w\tilde{t}} \cdot \left(n_{x,-i}^{(\tilde{t})} + \alpha\right) \cdot \left(n_{d,-i}^{(\tilde{x})} + \gamma\right) \qquad (8)$$

Where $n_{x,-i}^{(\tilde{t})}$ is the number of topic *t* been assigned to persona *x*, and $n_{d,-i}^{(\tilde{x})}$ is the number of entity *x* been assigned to document *d*. Both of them exclude the topic and entity assignment of word $w_i$. We report the perplexities with different number of topics on "Libya Event" test data set with 109 documents, about 10% of the whole data set.

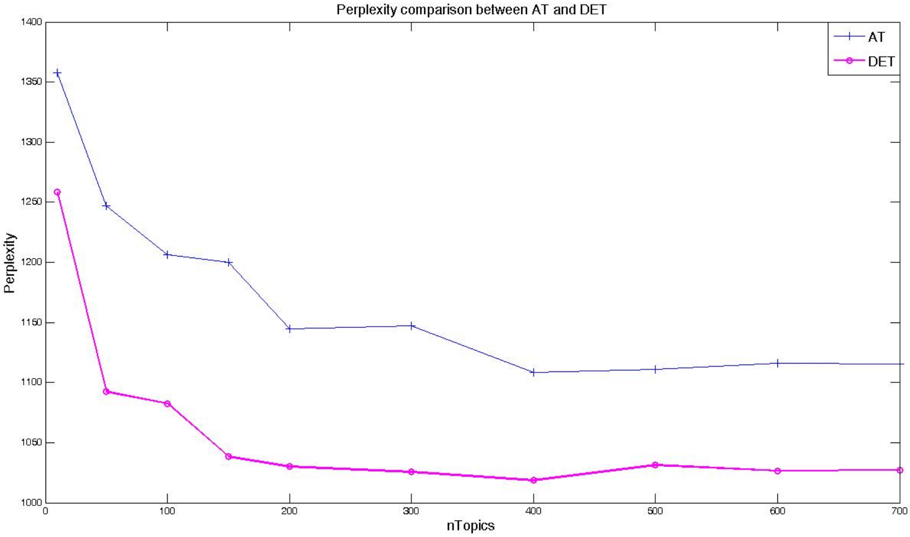


**Fig. 2.** Perplexity comparison of AT and DET on "Libya" data set. DET model has significantly better predictive power as AT over our document set. We can also find that the lowest perplexity obtained by DET is not achievable by AT with any topic number. It proves that DET can better adapt to the task of Semantic Social Network Analysis (SSNA), which discovers the topic-based relationship and group information of entities in documents.

### 4.2 Semantic Social Network Analysis with DET

**Topics and Entities.** We get the latent topics after applying the Gibbs sampling algorithm to DET model. We use the topic significance ranking method[10] to rank the topics and show two most important topics in table 1. In each topic we list the most likely words in the topic with their probability and below that the most likely entities and the topic names are named by authors.

During the experiment process, we have found that many topics own lots of same words with high probabilities, the reason we think is that all documents in "Libya Event" data set talk about one event (similar topics). We introduce in the idea of tf-idf algorithm to decide which words have high probabilities. The probability of word $w$ belonging to topic $k$ depends on both of the DET result, i.e., $\phi_{k,w}$ and the *tf_idf* value, which ranges from 0 to 1 with standardization. So the final probability of a word belonging to a topic is $\phi'_{k,w} = \delta \cdot \phi_{k,w} + (1-\delta) \cdot tf\_idf_{k,w}, 0 < \delta < 1$.

The probability of entity $x$ belonging to topic $k$ is not only decided by $\theta_{x,k}$, but also decided by the number of entity $x$ appearing in documents. If $x$ appears in document $d$, the number adds 1, and the appearing frequency is $df_x = |\{x \in d\}|/|D|$. So the probability of entity $x$ with topic $k$ is $\theta'_{x,k} = df_x \cdot \theta_{x,k}$.

**Table 1.** Two topics with highest probabilities from a 100-topic DET running with "Libya Event" data. In each topic we list the most likely words in lowercase with their probabilities, and below that the most likely entities in uppercase with initial.

| Topic89: Conflicts of government and opposition in Libya | | Topic31: National transition committee comes into existence | |
|---|---|---|---|
| the opposition | 0.751071 | committee | 0.313636 |
| demos | 0.098968 | transition | 0.278701 |
| fremdness | 0.018027 | nation | 0.213317 |
| relation | 0.015836 | admit | 0.046756 |
| find out | 0.013272 | chairman | 0.033091 |
| reason | 0.011508 | come into existence | 0.024036 |
| in the past | 0.008329 | spokesman | 0.020482 |
| hours | 0.007549 | intraday | 0.013421 |
| with responsibility for | 0.006162 | leaguer | 0.013068 |
| encounter | 0.005506 | promise | 0.006441 |
| Qaddafi | 0.060839 | Abdul-Jelil | 0.041501 |
| Bangh acirc | 0.043085 | Bangh acirc | 0.026637 |
| Qatar | 0.030726 | Italy | 0.025412 |
| Reuters | 0.027212 | National Transition Committee | 0.025164 |
| Italy | 0.020556 | Qatar | 0.023441 |
| Russia | 0.014663 | Bani Walid | 0.019576 |
| Abdul-Jelil | 0.014444 | Abdul-Jelil | 0.016731 |
| Egypt | 0.013855 | Beijing | 0.016373 |
| Associated Press | 0.008018 | Paris | 0.015959 |
| Muhammad | 0.007668 | London | 0.015528 |

**Entity Significance Ranking.** We suppose that if the topic distribution of an entity is much related to that of the document, the entity is significant to this document. Usually, KL divergence is used to measure the similarity between the entity and document. We show the KL divergences, probabilities and frequencies of all entities in two documents for particular information in table 2.

**Table 2.** KL divergences, probabilities and frequencies of all entities in two documents for particular information

| The National transition committee encounters plaster in Surt | | | |
|---|---|---|---|
| **entity** | **KL divergence** | **probability** | **frequency** |
| Misracirctah | 0.205826 | 0.181452 | 1 |
| Abdul-Jelil | 0.266745 | 0.181452 | 1 |
| Qaddafi | 0.485460 | 0.149194 | 10 |
| Saif-Nasser | 0.498184 | 0.125 | 1 |
| New York | 0.435166 | 0.084677 | 2 |
| Niger | 0.411440 | 0.060484 | 1 |
| Surt | 0.598851 | 0.044355 | 3 |
| Bani Walid | 0.494996 | 0.03629 | 1 |
| Tripoli | 0.635240 | 0.03629 | 1 |
| Jerusalem | 0.624283 | 0.028226 | 1 |
| UN's high conference of Libya appeals to picking up the reconstruction | | | |
| **entity** | **KL divergence** | **probability** | **frequency** |
| Abdul-Jelil | 0.018252 | 0.444015 | 1 |
| Wei Wei | 0.591291 | 0.374517 | 1 |
| Libya | 0.642270 | 0.104247 | 16 |
| UN | 0.662659 | 0.042471 | 6 |
| New York | 0.671120 | 0.019305 | 1 |
| Gu Zhengqiu | 0.669310 | 0.003861 | 1 |
| XinHua Net | 0.689017 | 0.003861 | 1 |
| UNSC | 0.652071 | 0.003861 | 1 |
| Ban ki-moon | 0.666539 | 0.003861 | 1 |

In most instances, if an entity which has a lower KL divergence with the document, the probability it belongs to that document will be higher, and the frequency is not a key factor to influence the belonging probability. In order to compare the entity ranking performances between AT and DET model on the whole data, we further adopt the weighted KL divergence which is defined as equations (9) and (10):

$$wKL\_AT = \frac{1}{D} \cdot \frac{1}{\mathbf{a}_d} \sum_{d=1}^{D} \sum_{a=1}^{\mathbf{a}_d} KL\left(\theta_{x,t} \parallel \eta_{d,t}\right) \tag{9}$$

$$wKL\_DET = \frac{1}{D} \cdot \sum_{d=1}^{D} \sum_{x=1}^{X_d} \left( \psi_{d,x} \cdot KL\left( \theta_{x,t} \parallel \eta_{d,t} \right) \right) \tag{10}$$

The smaller the weighted KL value is, the more similarity entities and documents own. In figure 3, we have shown the values with different topic numbers.

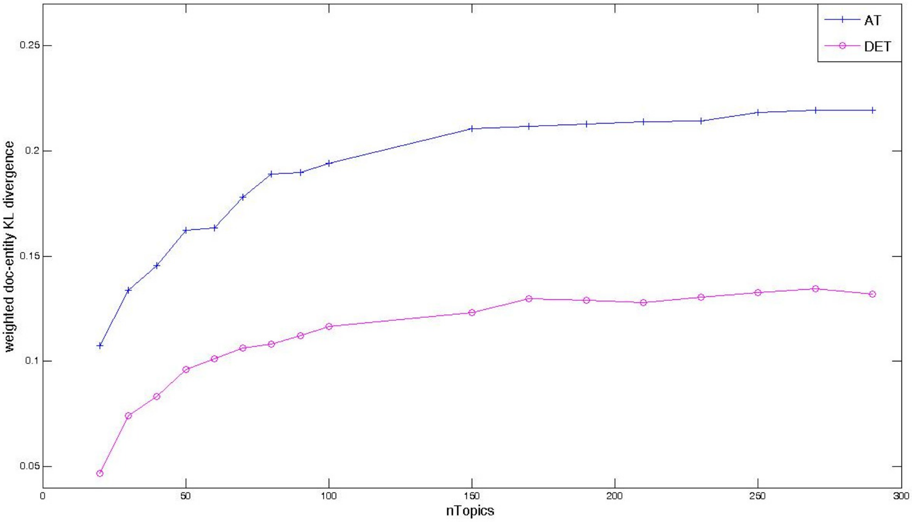


**Fig. 3.** The weighted KL divergences of AT and DET model with different topic numbers. The values of DPT model are lower than AT model. It means that the more important entities (with lower KL divergence to document which they appear in) have higher probabilities belong to the document and contributes more to the topic generativity.

# 5 Conclusions

We have presented the Document-Entity-Topic model, a probabilistic model for exploring the interactions of words, topics and entities within documents. It applies the probabilistic model to the social network analysis based on latent topics. In order to avoid the side effects of noisy entities and find out the entities which mainly affect the topics, we have introduced in the Dirichlet allocation for document-entity distribution other than uniform allocation. The model can be applied to discovering topics conditioned on entities, clustering to find semantic social groups, and ranking the significance of entities in a document.

However, while there is no entity in a document, the topics of that document can not be modeled. When such lack-of-entity documents arrive at a certain amount, the topic modeling of the corpus will be affected. Consequently, we try to improve the model for the application when there are many documents lacking of entities.

## References

1. Blei, D.: Introduction to Probabilistic Topic Models. Communications of the ACM (2011)
2. Steyvers, M., Griffiths, T.: Probabilistic Topic Models. Handbook of Latent Semantic Analysis 427 (2007)
3. Blei, D., Carin, L., Dunson, D.: Topic Models. IEEE Signal Processing Magazine 27(6), 55–65 (2010)
4. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
5. Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., Steyvers, M.: Learning Author-Topic Models from Text Corpora. ACM Transactions on Information Systems (TOIS) 28(1), 1–38 (2010)
6. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The Author-Topic Model for Authors and Documents, pp. 478–494. AUAI Press (2004)
7. Shiozaki, H., Eguchi, K., Ohkawa, T.: Entity Network Prediction Using Multitype Topic Models. Springer (2008)
8. Newman, D., Chemudugunta, C., Smyth, P.: Statistical Entity-Topic Models, pp. 680–686 (2006)
9. Bishop, C.: Pattern Recognition and Machine Learning. Springer, New York (2006)
10. AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C.: Topic Significance Ranking of LDA Generative Models. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS, vol. 5781, pp. 67–82. Springer, Heidelberg (2009)

# Appendix

We need to derive $P\left(x_i = x, z_i = t \mid w_i = w, \mathbf{x}_{-i}, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \gamma, \alpha, \beta\right)$, the conditional distribution for word $w_i$ given all other words' topic and entity assignments $z_{-i}$ and $x_{-i}$ to give out the Gibbs sampling procedure for DET model. We begin with the joint probability of the whole documents corpora. Here we can make use of conjugate priors to simplify the integrals.

$$
\begin{aligned}
& P\left(\mathbf{x,z,w} \mid \gamma, \alpha, \beta, X\right) \\
&= \iiint \prod_{d=1}^{D} p(\boldsymbol{\psi}_d \mid \gamma) \prod_{x=1}^{X} p(\boldsymbol{\theta}_x \mid \alpha) \prod_{t=1}^{T} p(\phi_t \mid \beta) \prod_{i=1}^{N_d} p(x_{di} \mid \boldsymbol{\psi}_d) p(z_{di} \mid \boldsymbol{\theta}_{dx_{di}}) P(w_{di} \mid \phi_{z_{di}}) d\boldsymbol{\Phi} d\boldsymbol{\Theta} d\boldsymbol{\Psi} \\
&= \int \prod_{d=1}^{D} \left( \frac{\Gamma(\sum_{d=1}^{D} \gamma_x)}{\prod_{x=1}^{X} \Gamma(\gamma_x)} \prod_{x=1}^{X} \boldsymbol{\psi}_{dx}^{\gamma_d - 1} \right) \prod_{d=1}^{D} \prod_{x=1}^{X} \boldsymbol{\psi}_{dx}^{n_{dx}} d\boldsymbol{\Psi} \\
& \quad \times \int \prod_{x=1}^{X} \left( \frac{\Gamma(\sum_{t=1}^{T} \alpha_t)}{\prod_{t=1}^{T} \Gamma(\alpha_t)} \prod_{t=1}^{T} \boldsymbol{\theta}_{xt}^{\alpha_t - 1} \right) \prod_{x=1}^{X} \prod_{t=1}^{T} \boldsymbol{\theta}_{xt}^{n_{xt}} d\boldsymbol{\Theta} \times \prod_{t=1}^{T} \left( \frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{tv}^{\beta_v - 1} \right) \prod_{t=1}^{T} \prod_{v=1}^{V} \phi_{tv}^{n_{tv}} d\boldsymbol{\Phi} \\
&\propto \prod_{d=1}^{D} \int \prod_{x=1}^{X} \boldsymbol{\psi}_{dx}^{\gamma_d + n_{dx} - 1} d\boldsymbol{\psi}_d \times \prod_{x=1}^{X} \int \prod_{t=1}^{T} \boldsymbol{\theta}_{xt}^{\alpha_t + n_{xt} - 1} d\boldsymbol{\theta}_x \times \prod_{t=1}^{T} \int \prod_{v=1}^{V} \phi_{tv}^{\beta_v + n_{tv} - 1} d\phi_t \\
&\propto \prod_{d=1}^{D} \frac{\prod_{x=1}^{X} \Gamma(\gamma_x + n_{dx})}{\Gamma\left(\sum_{x=1}^{X} (\gamma_x + n_{dx})\right)} \times \prod_{x=1}^{X} \frac{\prod_{t=1}^{T} \Gamma(\alpha_t + n_{xt})}{\Gamma\left(\sum_{t=1}^{T} (\alpha_t + n_{xt})\right)} \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta_v + n_{tv})}{\Gamma\left(\sum_{v=1}^{V} (\beta_v + n_{tv})\right)}
\end{aligned}
$$

Where $n_{dx}$ is the number of tokens assigned to persona x and document d, $n_{xt}$ is the number of tokens assigned to topic t and persona x, $n_{tv}$ is the number of tokens of word w assigned to topic t. Using the chain rule, we can obtain the conditional probability conveniently. We define $w_{-di}$ as all word tokens except the token $w_{di}$ :

$$
\begin{aligned}
& P(x_{di}, z_{di} \mid \mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \gamma, \mathbf{X}) \\
&= \frac{P(x_{di}, z_{di}, w_{di} \mid \mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}_{-di}, \alpha, \beta, \gamma, \mathbf{X})}{P(w_{di} \mid \mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}_{-di}, \alpha, \beta, \gamma, \mathbf{X})} \\
&\propto \frac{P(\mathbf{x,z,w} \mid \alpha, \beta, \gamma, \mathbf{X})}{P(\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}_{-di} \mid \alpha, \beta, \gamma, \mathbf{X})} \\
&\propto \frac{\gamma_{x_{di}} + n_{dx_{di}} - 1}{\sum_{x=1}^{X} (\gamma_x + n_{dx}) - 1} \frac{\alpha_{z_{di}} + n_{x_{di} z_{di}} - 1}{\sum_{z=1}^{T} (\alpha_z + n_{x_{di} z}) - 1} \frac{\beta_{w_{di}} + n_{z_{di} w_{di}} - 1}{\sum_{v=1}^{V} (\beta_v + n_{z_{di} v}) - 1}
\end{aligned}
$$