MTTM: A Multi-type Topic Model Applied to Heterogeneous Information Network

ABSTRACT

Topic modeling has shown to be useful for document analysis, and interactions among multi-type objects play a key role at disclosing the rich semantics of networks. However, it has some drawbacks. First, they infer topics by maximizing the likelihood of the collection, which leads learn topics poor coherent when the objects are in messy form. For example, since there are always less than six authors cooperating in an academic paper, authors are scattered in a bibliography network. Therefore, it is hard to assign to a same topic group, those authors who are interested in the same area but never cooperate. Second, some topic models consider network structures integrating heterogeneous information network with topic modeling, but they ignore the inner relations within these structures. In this paper, we propose a Multi-type topic model (MTTM) to solve these problems. Our model uses Gibbs sampling method and a joint different topics framework to integrate multi-type topics learning. It individually learns topics of each type of object and considers the interactions among different type of topics. In this way, MTTM can deal with more varieties of information networks, such as computing networks, social networks and social media. Our experiments on DBLP show that MTTM learns high-quality topics through mining the relationship among different types of data. By using our approach, all types of topics are more accurate, coherent and interpretable.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering*; H.2.8 [**Information Systems Applications**]: Database Applications – *data mining*

General Terms

Algorithms, Experimentation

Keywords

Topic modeling, Multi-type topic modeling, Heterogeneous information network

1. INTRODUCTION

Information networks, containing a large number of individual agents or components interacting with each other, are ubiquitous in many applications. Multiple types objects form a heterogeneous information network [1], which is different from homogeneous network, as it has more types of objects and more varieties of interconnection among them. Therefore, it is important and challenging to examine how multi-type objects in heterogeneous information network can mutually enhance each other in topic modeling.

Many topic models, such as Latent Dirichlet Allocation (LDA) [2], have been proposed and shown to be useful for document analysis. But they have two main disadvantages: First, the efficacy of topic models is challenged by insufficiency and scatteredness of information, e.g. tweets and authors of articles.

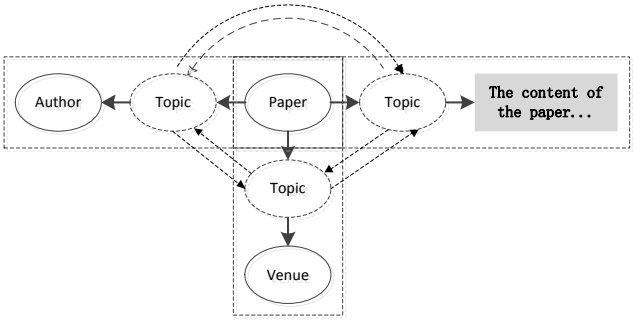


Figure 1. A Bibliographic Information Network on MTTM.

Since topics learned through LDA are formally a multinomial distribution over words, and by convention the top-10 words are used to identify the subject area or to give an interpretation of a topic [3], it is rather applied to more complete text content, such as academic abstracts and news articles. Obviously, insufficiency and scatteredness of information are also potentially relative to some specific topics. For example, in a bibliographic information network, the information about authors of academic articles is insufficiency and scatteredness. Generally, there are no more than 6 authors in an academic article. Authors still indicate some topics. Articles with author *Jiawei Han* may mostly indicate topics on data mining. The country Brazil may mostly indicate topics on World Cup in 2014. Some studies use topic models to learn topics from other type data instead of words in document. For example, *Youngchul & Junghoo* used topic model to analyze the relationship graph of popular social-network data [4]. Second, some topic models consider network structures as integrating heterogeneous information network with topic modeling, but they ignore the inner relations within these structures. Some studies include author-topic model [5], author-recipient-topic model [6] and TMBP [7] utilize additional information other than text by designing complex generative models that include additional types of objects. All of these studies are extensions to existing topic model framework, and regard text as the most relevant aspect, these methods only generate text oriented latent topics.

Conference'10, Month 1–2, 2010, City, State, Country.

For this purpose, we propose a new approach to apply topic models with heterogeneous information network to learn latent topics for multi-type objects, which is called Multi-type topic models (MTTM). As shown in Figure 1, in this model, we assume that each type of objects has its own latent topics. At the same time, each type topic may affect the other topics learning. In this way, we develop a joint framework to integrate a heterogeneous network with topic modeling. It learns topics of each type object respectively and considers the interactions among different types of topics.

2. Multi-type topic model

In this section, we introduce the problem of Multi-type topic models. Before introduce the algorithm, we define the problem of multi-type topic model and introduce the related concepts.

Definition 1. **Information Network.** Given a set of objects from T types $X = \{X_t\}_{t=1}^{T}$, where X_t is a set of objects belonging to t_{th} type, a weighted graph $G =< V, E, W >$ is called an information network on objects X, if $V=X$, E is a binary relation on V, and W: $E \rightarrow R^+$ is a weight mapping from an edge $e \in E$ to a real number $w \rightarrow R^+$. Specially, we call such an information network heterogeneous network when $T \geq 2$; and homogeneous network when $T = 1$.

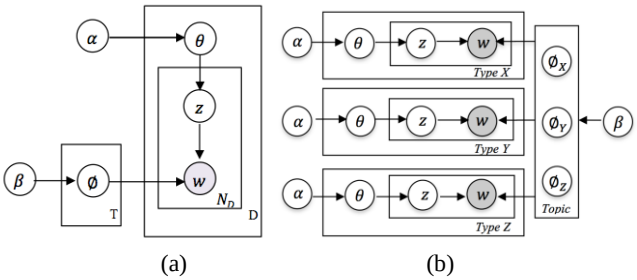
Definition 2. **Star Network Schema.** An information network $G =< V, E, W >$ on $T + 1$ types of objects $X = \{X_t\}_{t=0}^{T}$ is called with star network schema, if $\forall e =< x_i, x_j > \in E$, $x_i \in X_0$, $x_j \in X_t (t \neq 0)$. G is called star network. Type X_0 is called the center type. X_0 is also called the target type and $X_t (t \neq 0)$ are called attribute type.

Example 1 (bibliographic information network)

A bibliographic network consists of rich information about academic papers, each using a set of words, written by a group of authors, published in a certain time and published in a venue (a conference or a journal). Such a bibliographic network is composed of five types of objects: author, venues, words, time and papers. Links exist between papers and authors by the relation of "write" and "written by", between papers and terms by the relation of "contain" and "contained in", between papers and venues by the relation of "publish" and "published by", between papers and time by the relation of "publish" and "published in".

In the topic model perspective, document content are based upon the idea that the probability distribution over words in a document can be expressed as a mixture of topics, where each topic is a probability distribution over words. By the same way, the probability distribution over authors of a document can be expressed as a mixture of topics, where each topic is a probability distribution over authors; the probability distribution over venue of a document can be expressed as a mixture of topics, where each topic is a probability distribution over venues. So, the generative process corresponds to the MTTM models shown in Figure 2(b). In this model, each type object *w* is associated with a distribution over topics, θ, chosen from a symmetric $Dirichlet\ (\alpha)$ prior. The mixture weights corresponding to the chosen object are used to select a topic *z*, and an object is generated according to the distribution $\emptyset$ corresponding to that topic, drawn from a symmetric $Dirichlet\ (\beta)$ prior. Such a paper in bibliography network is generated by the following process:

1. For each type of attribution, $t \in \{1, \ldots, T\}$ choose a distribution over topic z_t.
2. For each item in the document



(a) (b)

Figure 2. Generative models (a) Latent Dirichlet Allocation (LDA); (b) The Multi-type topic model

i. Randomly choose a topic from the distribution over topics.

ii. Randomly choose an item from the corresponding distribution over the vocabulary based on other types' topic distribution.

Since some latent relationship is cross the type of topics. For example, some words are more time specific and some authors prefer some words. Every item chosen from a distribution is affected by other types of topics distribution. Thus, this model comprehensive considers all attributes of the document, make the items are chosen more rational.

3. Topics Learning Algorithm

In order to learn the topics of each type object, we consider the other type objects that related to the same center object. We observe the conditional probability given these objects topics. Enlightened by this property holding for topic model, we can have an intuition that multi-type topic model is able to catch the largest component structure of a network under the constraints that the relation between objects are recovered by 1-dimensional topic model. As a result, multi-type topic model should have better performance than simple topic model in most cases. In the DBLP dataset, according to the rules that (1) the same area authors are mostly talking about the same topic; (2) the similar topics are always talked by the authors in same area.

A variety of algorithms have been used to estimate the parameters of topic models. The LDA model uses Gibbs Sampling, as it provides a simple method for obtaining parameter estimates under Dirichlet priors and allows combination of estimates from several local maxima of the posterior distribution [5].

The LDA model has two sets of unknown parameters the center object distribution θ, and the *T* topic distribution ∅ - as well as the latent variables corresponding to the assignments of individual objects to topics *z*. We construct a Markov chain that converges to the posterior distribution on *z* and then use the results to infer θ and ∅ [8]. The transition between successive states of the Markov chain results from repeatedly drawing *z* from its distribution conditioned on all other variables, summing out θ and ∅ using standard Dirichlet integrals:

$$P(z_i = j | X_i = m, z_{-i}, X_{-i}, \tau_C = c) \propto \frac{C_{mj}^{XT} + \beta}{\sum_{m'} C_{m'j}^{XT} + V\beta} \frac{C_{cj}^{CT} + \alpha}{\sum_{j'} C_{cj'}^{CT} + T\alpha} \quad (1)$$

where $z_i = j$ represents the assignments of the i_{th} object related to a center object to topic j, $X_i = m$ represents the observation that the i_{th} object is the m_{th} object in the set of current type of objects, z_{-i} represents all topic assignments not including the i_{th} objects,

and $\tau_C = c$ represents the center object is *c* in type C. Furthermore, C_{mj}^{WT} is the number of times object *m* is assigned to topic *j*, not including the current instance, and C_{cj}^{CT} is the number of times topic j has related to center object *c*, not including the current instance. For any sample from this Markov chain, being an assignment of every object to a topic, we can estimate θ and $\emptyset$ using

$$\emptyset_{mj} = \frac{C_{mj}^{XT} + \beta}{\sum_{m'} C_{m'j}^{XT} + V\beta} \quad (2)$$

$$\theta_{cj} = \frac{C_{cj}^{CT} + \alpha}{\sum_{j'} C_{cj'}^{CT} + T\alpha} \quad (3)$$

where $\emptyset_{mj}$ is the probability of using object *m* in topic j, and θ_{cj} is the probability of topic *j* related to center object *c*. These values correspond to the predictive distributions over new attribute type of objects X_i and new topics *z* conditioned on X_i and *z*.

In MTTM model, in order to utilize the other attribute type information to help assign the i_{th} object in type *X to* topic *j*. We use conditional probability $\mathrm{p}(T_X = j|T_Y)$ to draw the occurred probability of topic T_X in type X given the topic T_Y in type Y related the same center object. We can estimate the probability by:

$$\mathrm{p}(T_X = j|T_Y = m) = \frac{C_{mj}^{XY}}{\sum_{j'} C_{mj'}^{XY}} \quad (4)$$

$$\mathrm{p}(T_X = j|T_Y) = \frac{\sum_m \mathrm{p}(T_X = j|T_Y = m)}{|T_Y|} \quad (5)$$

where $T_X = j$ represents the assignments of the objects in type X to topic *j*, $T_Y = m$ represents the topic in type Y is *m*, and C_{mj}^{XY} is the number of times object in type X is assigned to topic *j* given the topic *m* in type Y is related to the same center object, not including the current instance. $|T_Y|$ respects the number of topics in type Y. For utilizing all types information to learn topics for type *X*, we combine all the multi-type topics probabilistic results. These random variables are estimated from samples via:

$$\mathrm{P}(z_i = \mathrm{j}|X_i = \mathrm{m}, z_{-i}, X_{-i}) \propto \mathrm{P}(z_i = \mathrm{j}|X_i = \mathrm{m}, z_{-i}, X_{-i}, \tau_C = c) \prod_t \mathrm{p}(T_X = j|T_t) \quad (6)$$

where the t represents all types not including type X In the examples considered here, we do not estimate the hyper-parameters α and β – instead the smoothing parameters are fixed at 50/T and 0.01 respectively.

We start the algorithm by assigning objects to random topics. Each iteration of the algorithm involves applying Equation 6 to every object token in the bibliographic network, which leads to a time complexity that is of order of the total number of object tokens in the training data set multiplied by the number of topics.

Algorithm 1 MTTM Learning Model

Input:

- The graph $G =< V, E, W >$ V = C ∪ X ∪ Y ∪ ⋯
- Iteration number t = 1000
- Each type's maximum topics number $\{M_X, M_Y, \cdots\}$

Output:

- The topics for each type objects

Procedure:

```
assign each objects to random topics;
repeat t times
  foreach object τ_C in center type C
    foreach object τ_X related to τ_C
      foreach topic j in type X
        foreach other type topics in type t related to object τ_C
          φ_tj = (Σ_m p(T_X = j|T_t = m)) / |T_t|
        end for;
        z_i ← argmax(P(z_i = j|X_i = m, z_-i, X_-i, τ_C = c) ∏_t φ_tj)
      end for;
      assign current object to topic z_i.
    end for;
  end for;
end repeat;
return each topics with objects;
```

4. EXPERIENCE

We now study the effectiveness and accuracy of MTTM process and compare it with state-of-the-art algorithms.

4.1 Data Set

We use real data set from DBLP and build bibliographic information networks according to Example 1, we extract a data set ("four-area" data set) which contains four conferences that are most related to data mining, which are database, data mining, information retrieval and machine learning. Five representative conferences for each area are picked, and all authors have ever published papers on any of the 20 conferences, all these papers and terms appeared in these titles are included in the network. For reduce the data size of dataset, we only use the paper published from 2008 to 2011. Totally there are 14580 papers in our data set. We set the topic number of title terms as 100, the topic number of authors as 10, and we according to the number of the year divide the published year into 4 different topics: topic of 2008, 2009, 2010, 2011.

4.2 Evaluate the topics

In order to the qualitative evaluation of topics learned by MTTM, we also evaluated the proposed MTTM in terms of perplexity and accurate.

Perplexity is a standard measure for estimating the performance of a probabilistic model. The perplexity of a test set of objects τ_X with center objects τ_Z, (τ_X, τ_Z), is defined as the exponential of the negative normalized predictive likelihood under the model,

$$\mathrm{perplexity}(\tau_X|\tau_{Z_i}) = \exp - \frac{\sum_{i=1}^{|\tau_Z|} \log \mathrm{p}(\tau_X|\tau_{Z_i})}{\sum_{i=1}^{|\tau_Z|} |\tau_X \in \tau_{Z_i}|} \quad (7)$$

Better generalization performance is indicated by a lower perplexity over a set of held-out dataset. The term in the brackets is simply the probability for the set of objects τ_X given the set of center objects τ_Z.

The accuracy is calculated the classification result on paper. We use the topics learned by each method to classify the paper into different clusters that share the similar topics.

In our experiments, we compared the topic model (LDA), the TMBP model and the MTTM model. In the DBLP test set, topic model is only use the title terms of papers to learn the topics. The TMBP model utilizes the author, conference and paper structure to help to learn the topics for title terms. The MTTM model simultaneous learns the topics for author, conference and title term objects. For each test, we compare the perplexity and accuracy value in different number of iteration.

Figure 3 shows the results for the three models being compared. In Figure 3(a), since LDA method consider the least structure information, it better than the other two models, as illustrated by its lowest perplexity. Consider the same scope structure information, the MTTM is a little worse than TMBP. However, the accuracy of compared result is shown in Figure 3(b). The MTTM method expresses the best accuracy in classification test. Since the MTTM maximum use the structure information, not only the objects co-occurred information, but also the different types of topics latent relationship. Overall, the performance of multi-topic model is better in dealing with the multi-attribute and complex information.

In order to tell which kind of model's topics is more coherent and interpretable, we design an experience to compare the same author's paper's similarity through the description with content topics. In general, the same author always writes papers in several topics. The similarities among one author's paper always are high score. As illustrate in Figure 4, the performance of MTTM is the best, since the average Euclidean distance among the same author's papers is 0.195; the different authors' papers' topics is 0.42. The average Euclidean distance ratio between different author's papers and the same author's papers is the greatest with 2.15. The ratios of LDA and TMBP are 1.19 and 1.7 respectively. Further, the normal distribution of MTTM, which describes the similar value of different papers, is thinner and higher than other two methods. These express topics distribution of the same author's papers' are more compactness than other topic models. Consequently, MTTM model provides a useful method for learning objects in messy form more coherent and interpretable.

5. CONCLUSION

We have presented Multi-type topic models (MTTM) applied on a special heterogeneous network with star network schema. Unlike most previous work, MTTM allows to integrate multi-type objects and individually learn topics for each type object. Even if objects are insufficient and scatteredness, the topics are learned in a more accurate and coherent way. Moreover, sometimes, single type objects may contain multi-type information, such as textual content including not only words but also semantics, emotions, structure and so on. So MTTM, which extends topic models, could be applied in more applications.

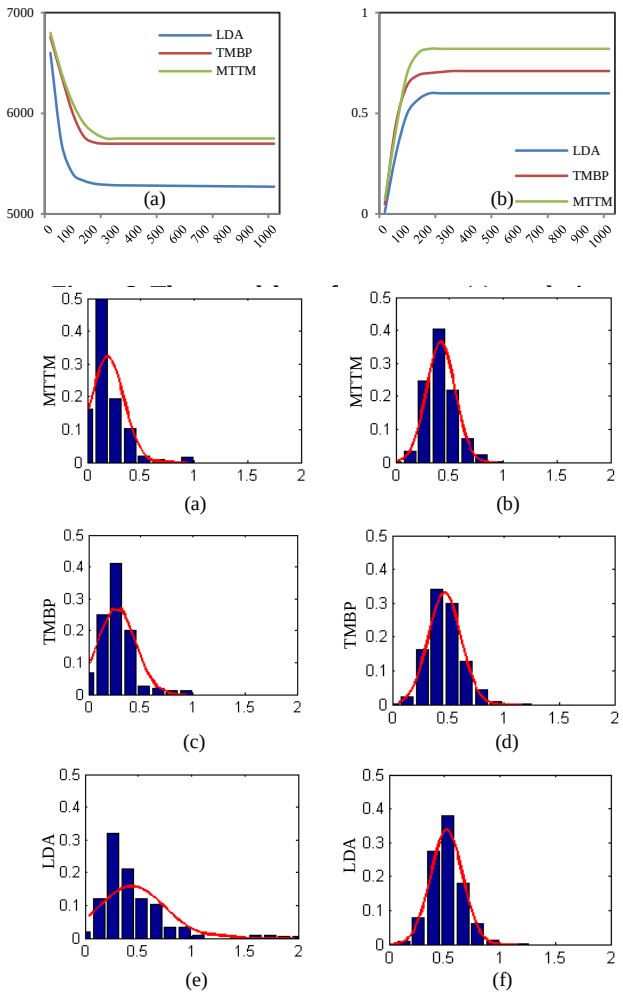


Figure 4. The topic similarity of different papers from same author (a)(c)(e) and different authors (b)(d)(f).The topics are learned by MTTM, TMBP and LDA. The similarity is calculate by Euclidean distance.

6. REFERENCES

[1] Sun, Y., Y. Yu and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009. Paris, France: ACM.

[2] Blei, D.M., A.Y. Ng and M.I. Jordan, Latent dirichlet allocation. J. Mach. Learn. Res., 2003. 3: p. 993-1022.

[3] Mehrotra, R., et al. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013. Dublin, Ireland: ACM.

[4] Cha, Y. and J. Cho. Social-network analysis using topic models. in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. 2012. Portland, Oregon, USA: ACM.

[5] Rosen-Zvi, M., et al., Learning author-topic models from text corpora. ACM Trans. Inf. Syst., 2010. 28(1): p. 1-38

[6] Mccallum, A., et al., The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and academic email. Technical report, Univ. Massachusetts, Amherst. 2005.

[7] Deng, H., et al. Probabilistic topic models with biased propagation on heterogeneous information networks. in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011. San Diego, California, USA: ACM.

[8] T.L. Griffiths, et ak., Finding scientific topics. Proceedings of the National Academy of Sciences, 2004, 101 (suppl. 1),50-57.