Probabilistic Topic Model on Heterogeneous Information

Author: Huaizhi Zhang
Supervisor: Joemon M. Jose

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

A dissertation presented in part fulfilment of the requirements of the Degree of Master of Science at The University of Glasgow

7th Sep 2015

Abstract

With the wide broadcast of social media and the abundance of news feeds, textual documents from disparate sources are not only getting richer, but also ubiquitously interconnected with each other in various ways. Recently, serveal topic models have been proposed for heterogeneous information and shown to be effective for event detection. However, these topic models use the same topic model on different sources, which results in the loss of original information. Interaction between different sources is also ignored within the modeling procedure. To address these problems, we studied the topic models by exploiting the individual properties of different sources and the inner-relationships between them. In this paper, a novel heterogeneous topic model is proposed to enable the interaction of topics between sources by Gibbs sampling. The underlying intuition is that multi-typed sources should be treated differently along with their extra information, which will not suffer from the loss of data in the training procedure. We proposed a novel heterogeneous topic model, which is compared with existing topic models. This topic model is evaluated on two real-world data collections, consisting of Twitters and News text streams. Traditional topic models is compared with this heterogeneous topic model on different aspects. The experimental results shows that the heterogeneous topic model has a better performance on modeling heterogeneous information.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: Huaizhi Zhang Signature: Huaizhi Zhang

Acknowledgements

I would like to thank the following people for their help and support throughout the project.

My supervisor, Professor Joemon M. Jose, who helpfully guided me until the submission of this thesis. His guidance prevented me from loosing track.

Many thanks to Post Long Chen. During this project, we discussed many aspects in topic modeling, such as news events detection. His help guided me to work out the heterogeneous topic model.

I also want to thank Professor Haitao Yu and PHD student Jie An for discussing the details of the topic model, which make me sure that this model is correct and meanful.

Contents

Chapter 1

Introduction

1.1 Background and Motivation

Topic modeling, as a form of text mining, can discover the latent patterns in corpus. Recent studies focused on modeling heterogeneous information, which consists of news articles, social media, information networks and etc. These studies suggest that by considering the relations between different sources, topic models can have a better performances than traditonal topic models, such as Probabilistic Latent Semantic Analysis (pLSA)[6] and Latent Dirichlet Allocation (LDA)[1]. Relations, such as the authors and messages in Twitter, are ubiquitously interconnected with each other as well as the latent topics between different sources. Sources, including news articles and social media, are usually talking about the same topics or events in a period of time. For example, the massive explosion in Tianjin, China that happened in 12 Auguest has drawn significant attentions in news press and social media. Numbers of articles and messages discussed this event, which means there is a latent topic between news and social media. Given the vast quantities of information, it is impossible to manage and classify the information manually. Therefore, it is reasonable and challanging to discover a model to mine topics on the heterogeneous information.

1.2 Problems

Although recent topic models have been shown to be effective for heterogeneous information, these models suffer from two main problems. The first one is that there is no correspondence between local topics of different sources. Nor is there correspondence between common topics and local topics. As a result, it is difficult to find the relations between topics of different sources. One way to solve the correspondent problem is to merge the data from different sources and do topic modeling on the whole collection. However, this approach can not maintain the individual properties of each source. These two problems make topic modeling on heterogeneous information a big challenge.

1.3 Contributions

Our work builds on recent work in topic models. More specifically, we extend topic models by sharing a common topic distribution. The experiments on topic modeling is conducted on various sources, including news articles and Twitter messages. After analyzing the feature of data and topic models, we proposed a novel heterogeneous topic model by combining LDA and Author-Topic Model. To summarize, the contributions of this paper are:

1. We proposed a novel heterogeneous topic model, in which individual properties and correspondence between sources is maintained.
2. The proposed heterogeneous topic model can discover common topics and estimate local topics for each source.

1.4 Outline

The rest of the paper is organized as follows. In Section 2, we discuss the existing topic models and their limitations. The heterogeneous topic model is introduced in Section 3. Section 4 reports the experimental results and evaluations. Finally, we present the conclusion and the future work.

Chapter 2

Literature Survey

In this chapter, a review of topic models and related techniques are presented. Firstly, a general survey about topic modeling on heterogeneous information is discussed, followed by the Gibbs Sampling technique. Then we detail two topic models, Latent Dirichlet Allocation (LDA) and Author Topic Model (ATM), which serve our Heterogeneous Topic Model (Section 3.2). Finally, we introduce a state-of-the-art topic model: Collection Model (CM)[7], which is used to compare with our heterogeneous topic model.

2.1 Overview

One way to combine heterogeneous sources is to simply merge documents from different sources into a single collection and then apply existing modeling method, such as pLSA and LDA. However, due to the impact of uneven data, the results might be biased in favor of the larger source. The combination may also lead to a loss of feature within different sources. For example, tweets are more likely based on the interests while news documents focus on significant news events in various aspects, such as politics, sports, criminal and etc. Alternatively, one can run existing topic models on each sources respectively, which could maintain the characteristic of each source. However, it is hard to capture the common topics and correspondence among various sources.

Serveal researches[11][7] have been proposed to find local topics for individual sources and common topics for the whole collection as a effective way for heterogeneous information, but these approaches is essentially equivalent to the method metioned above. By utilizing the information network, recent studies, including Laplacian PLSI[2], NetPLSA[8] and TMBP[3], were introduced to integrate surface text with network structures, which achieved a better performance. Nonetheless, all of these topic models can not solve the two main problems caused by the heterogeneous information (Section 1.2). To solve the problems, Ghosh et al (2013)[4] proposed a ProbLDA to preserve the correspondence between topics. However, this model can not work on a more complicated heterogeneous information, such as news articles together with Twitter messages. In other words, this model results in the loss of information, including authors. Aftering addressing the existing problems of topic model on heterogeneous information, we proposed a novel heterogeneous topic model.

2.2 Gibbs Sampling

In topic models, such as LDA and ATM, Gibbs sampling is commonly used to approximate the specific multivariate probability distribution since direct sampling is difficult.

Assuming that the full conditional distribution is:

$$\{\pi(\phi_1|\phi_2,\phi_3,\ldots,\phi_p);\pi(\phi_2|\phi_1,\phi_3,\ldots,\phi_p);\ldots;\pi(\phi_p|\phi_1,\phi_2,\ldots,\phi_{p-1})\}$$

where π represents the conditional function, ϕ_p represents the value in p^{th} dimension. Then the sampling procedure can be represented as:

1. Set an initial value for $\phi_1^{(0)},\phi_2^{(0)},\ldots,\phi_p^{(0)}$
2. Repeat until sampling process is convergent:

   Generate ϕ_1^{i+1} from $\pi(\phi_1|\phi_2^{(j)},\phi_3^{(j)},\ldots,\phi_p^{(j)})$
   Generate ϕ_2^{i+1} from $\pi(\phi_2|\phi_1^{(j)},\phi_3^{(j)},\ldots,\phi_p^{(j)})$
   ...
   Generate ϕ_p^{i+1} from $\pi(\phi_p|\phi_1^{(j)},\phi_2^{(j)},\ldots,\phi_p^{(j)})$

Figure 2.1 illustrates the process of Gibbs sampling in two dimensional space. From this figure, we can find that the sampling process tends to be convergent to the center of the conditional distribution.

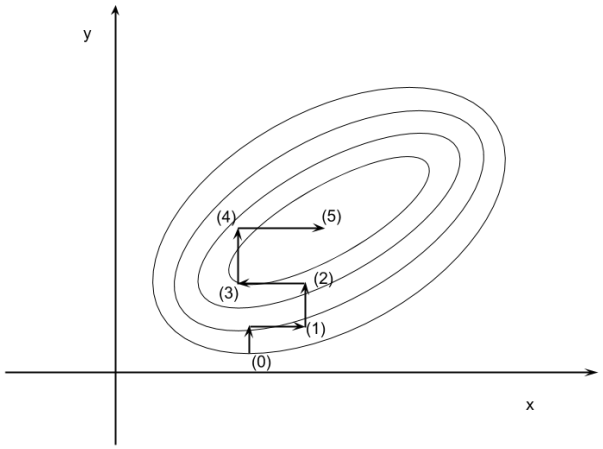


Figure 2.1: Gibbs sampling algorithm in two dimensions starting from an initial point and then completing five iterations

Expectation–Maximization (EM) algorithm[1] is another method to approximate the probability distribution in topic models. However, with the increasing numbers of datasets, EM tends to be much slower than Gibbs sampler, which might be computational expensive.

In the following sections, we talk about the topic models with Gibbs sampling.

[1] https://en.wikipedia.org/wiki/Expectation-maximization_algorithm

2.3 Topic Modeling

In this section, we first compare three traditional topic models, including pLSA, LDA and ATM. Then we address their problems and limitations. The following subsections go through a series of traditional topic models, including LDA and ATM, which is proposed to serve the heterogeneous topic model. Finally, the collection topic model is detailed.

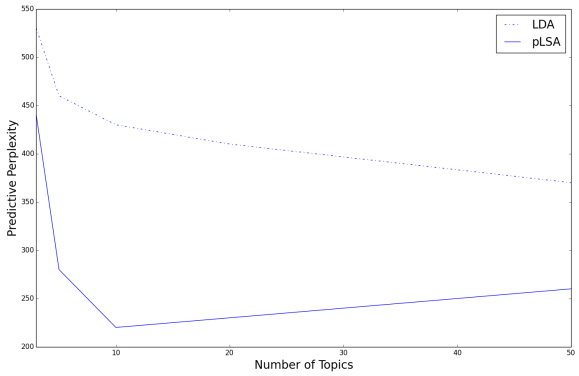


Figure 2.2: Comparison of predictive perplexity between LDA and pLSA

According to Blei et al [1], the perplexity of LDA is lower than pLSA (Figure 2.2). In other words, LDA have a better performance than pLSA to predict a document.

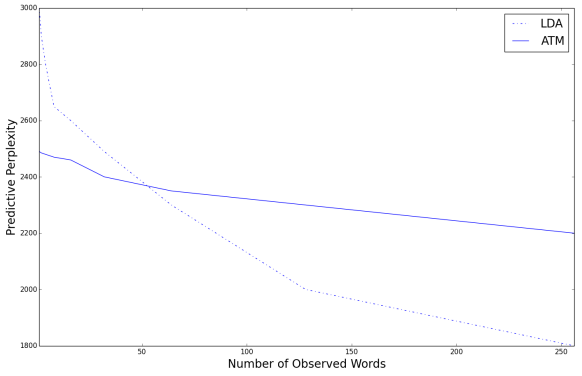


Figure 2.3: Comparison of predictive perplexity between LDA and ATM

Figure 2.3 show the results from Rosen et al's research [9], which compare the predictive perplexity between LDA and ATM. The results shows that the performance rely on the number of observed words. For the relative small numbers of observed words, LDA has higher perplexity than ATM. Therefore, it is difficult for LDA to predict the document with a few observed words. On the contrary, ATM has a better performance than LDA by utilizing the extra author information.

Motivated by the researches' results, we proposed to review LDA and ATM in detail, which is used to model news articles and Twitter messages respectively.

2.3.1 Latent Dirichlet Allocation

In this section, we introduced the LDA model. LDA is a generative topic model, in which each word is sampled from a predefined word-topic distribution and a document contains a fixed number of words. To model the distribution, the dirichlet parameters are attached to word-topic distribution and topic-document distribution respectively. Then, the sequential procedure of first sampling a topic followed by sampling words, leads to the following generative process:

For each document $d = 1, \ldots, D$, choose $\theta \sim \text{Dirichlet}(\alpha)$

For each word $w = 1, \ldots, N_d$

i. Choose $z_{di} \sim \text{Multinomial}(\theta)$

ii. Choose $w_{di} \sim \text{Multinomial}(z_{di}, \beta_{z_{di}})$

Figure 2.4 illustrates the graphic model of LDA, which is corresponding to the generative process. In this graphic model, θ represents the topic distribution over documents, β represents the word distribution over topics, z is the topic drawn from θ, w is drawn from β based on the chosen topic z. The motivation behind this model is that each word in the document is generated based on the latent topic distribution, where this distribution is characterized by a distribution over words.

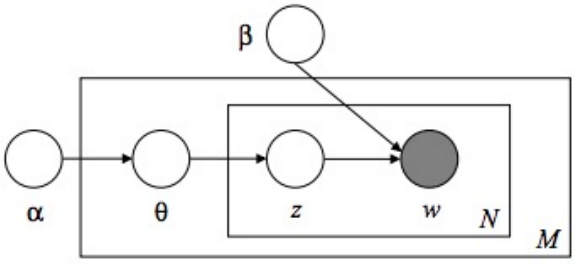


Figure 2.4: Graphic Model of LDA

According to the graphic model, we can obtain the following joint distribution of the collection, which can be viewed as the observed value of the collection:

$$p(w, z, \theta|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2.1}$$

After intergrating over θ and summing over z, the marginal distribution, which is the likelihood of the collection, can be obtained:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \tag{2.2}$$

However, since both $p(z_n|\theta)$ and $p(w_n|z_n, \beta)$ is unobserved, Gibbs Sampling is performed to compute the these values. Following is the basic equation for Gibbs sampler:

$$p(z_i = k|z_{-i}, w) = \frac{(n^{(k)}_{m,-i} + \alpha_k)}{(\sum_{k=1}^{Z} n^{(k)}_{m,-i} + \alpha_k)} \frac{(n^{(t)}_{k,-i} + \beta_t)}{(\sum_{t=1}^{V} n^{(t)}_{k,-i} + \beta_t)} \tag{2.3}$$

where $n^{(t)}_{k,-i}$ represents the frequence of word t that belongs to topic k (except i^{th} topic), $m^{(t)}_{k,-i}$ represents document t that belongs to topic k (except i^{th} topic), α_k and β_k represent the Dirichlet parameters for document-topic and topic-word distribution respectively.

Since the Dirichlet distribution is conjugate to the Multinomial distribution, the probability of words given on topics(ϕ) and topics given on documents(θ) can be obtained from Equation 2.3

$$\theta_{m,k} = \frac{m^{(t)}_k + \alpha}{\sum_{k=1}^{Z} n^{(k)}_m + \alpha_k} \tag{2.4}$$

$$\phi_{k,t} = \frac{n^{(t)}_k + \beta_t}{\sum_{t=1}^{V} n^{(t)}_k + \beta_t} \tag{2.5}$$

where $\phi_{k,t}$ represents the probability of word t belong to topic k, $\theta_{m,k}$ represents the probability of topic k that the document m is.

2.3.2 Author Topic Model

ATM[10] is also a generative model, where documents are represented as random mixtures over latent topics, and each topic is characterized based on author-topic distribution.

Figure 2.5 illustrates the graphic model of ATM. In this model, a document d is represented as a vector of words w_d. The unique words in documents make up of the corpus $N = \sum_{d=1}^{D} N_d$. Additionally, the authors of the document are represented as a vector of authors a_d. In order to model the author's interest to a topic, x is used to denote the probability of topics given on the author. In the end, z represents the chosen topic and w represents the observed word. For derivation, θ and ϕ is used to represents the mixtures of the authors distribution and mixtures of the topics distribution respectively. α and β denotes the Dirichlet parameters for θ and ϕ, which models the author-topic distribution and topic-word distribution respectively.

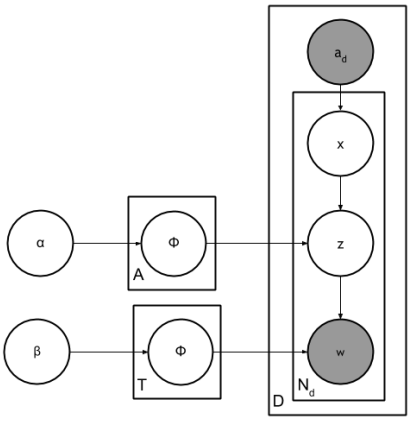


Figure 2.5: Author Topic Model

The sequential procedure of this model is given as follows:

1. For each author $a = 1, \ldots, A$ choose $\theta_a \sim \text{Dirichlet}(\alpha)$
   For each topic $t = 1, \ldots, T$ choose $\phi_t \sim \text{Dirichlet}(\beta)$
2. For each documents $d = 1, \ldots, D$:
   Given the vector of authors a_d
   For each word $i = 1, \ldots, N_d$:
   Conditioned on a_d choose an author $x_{di} \sim \text{Uniform}(a_d)$
   Conditioned on x_{di} choose a topic $z_{di} \sim \text{Discrete}(\theta_{xdi})$
   Conditioned on z_{di} choose a word $w_{di} \sim \text{Discrete}(\phi_{zdi})$

In the generative process of this topic model, each word is drawn conditional independent on ϕ and z, then the joint distribution of the corpus is obtained by:

$$p(w|\theta, \phi, \mathcal{A}) = \prod_{d=1}^{D} p(w_d|\theta, \phi, \mathcal{A}) \tag{2.6}$$

where $\mathcal{A} = \sum_{i=1}^{A} \mathcal{A}_i$ represents the authors set.

Since exact inference is intractable for large dataset, Gibbs Sampling is performed to estimate and obtain the topic distribution over authors and word distribution over topics. The basic equation for Gibbs sampler is given by:

$$\begin{aligned} p(x_{di} = a, z_{di} = t|w_{di} = w, z_{-di}, x_{di}, w_{di}, \mathcal{A}, \alpha, \beta) &\propto \frac{p(w, x, z|\alpha, \beta, \mathcal{A})}{p(w_{di}, x_{di}, z_{di}|\alpha, \beta, \mathcal{A})} \\ &= \frac{C^{TA}_{ta,-di} + \alpha}{\sum_{t'} C^{TA}_{t'a,-di} + T\alpha} \frac{C^{WT}_{wt,-di} + \beta}{\sum_{w'} C^{WT}_{w't,-di} + W\beta} \end{aligned} \tag{2.7}$$

Since the Dirichlet disribution is conjugate to the Multinomial distribution, we can obtain the posterior probability matrix of author Θ and topic Φ, where each element of Θ and Φ is given as follows:

$$\theta_{ta} = \frac{(C^{TA}_{ta)^s+\alpha}}{\sum_{t'}(C^{TA}_{t'a})^s + T\alpha} \tag{2.8}$$

$$\phi_{wt} = \frac{(C^{WT}_{wt)^s+\beta}}{\sum_{w'}(C^{WT}_{w't})^s + W\beta} \tag{2.9}$$

where $C^{WT}_{w't}$ represents the number of words w' assgin assigned to topic t, $C^{TA}_{t'a}$ is the number of author t' assigned to author a. Other notations are summarized in Table 3.1.

2.3.3 Collection Model

Collection model is an extension of the LDA model, where a Bernoulli distribution over local topics and common topics is attached to the model. Assuming that there is a set S of n text streams in this model, T_s, T_c represents the local topics and common topics respectively. The local topics $\theta^{(s)}$ are drawn from a Dirichlet distribution $\beta^{(s)}$ while the common topics $\theta^{(c)}$ are drawn from a Dirichlet distribution $\beta^{(c)}$. Each text document d in a stream s is drawn from a Bernoulli distribution $\eta_{d,s}$, which follows $\eta_{d,s} \sim \text{Beta}(\gamma_s^{(s)}, \gamma_s^{(c)})$. The parameter $\eta_{d,s}$ represents the how likely the document is chosen from local topics. In addition, $\eta_{d,c}$ represents the how likely the document is chosen from common topics. Note that $\eta_{d,s} + \eta_{d,c} = 1$. For each word w in document d, a random variable $x_{d,i}$ is first drawn from Bernoulli($\eta_{d,s}$), followed by picking a topic z_{di} from Multinomial($\theta_d^{(x_{di})}$), then word w_{di} is chosen from Multinomial($\phi_{z_{di}}^{(x_{di})}$). To summarize, the sequential process is:

1. For all common topics T_c, draw $\phi^{(c)} \sim \text{Dir}(\beta^{(c)})$
2. For a particular stream s
   (a) For all local topics T_s, draw $\phi^{(s)} \sim \text{Dir}(\beta^{(s)})$
   (b) For each document d in s
      i. Draw Bernoulli parameter $\eta_{d,s} \sim \text{Beta}(\gamma_s^{(s)}, \gamma_s^{(c)})$
      ii. Draw $\theta_d^{(s)} \sim \text{Dir}(\alpha_s)$
      iii. Draw $\theta_d^{(c)} \sim \text{Dir}(\alpha_c)$
      For each word w_{di} in document d
      A. Draw $x_{di} \sim \text{Bernoulli}(\eta_{d,s})$
      B. Draw a topic $z_{di} \sim \text{Multinomial}(\theta_d^{(x_{di})})$
      C. Draw a word $w_{di} \sim \text{Multinomial}(\phi_{z_{di}}^{(x_{di})})$

By using Gibbs sampling, we can get the topic-document distribution $\theta_{d,z}$, topic-word distribution $\phi_{z,w}$ and local-common topic distribution $\eta_{d,x}$ as follows:

$$\begin{aligned}
\theta_{d,z}^{(x)} &= \frac{m_{d,z} + \alpha_z}{\sum_{z \in T_x} m_{d,z} + \alpha_z}, x \in \{s, c\} \\
\phi_{z,w}^{(x)} &= \frac{n_{z,w} + \beta_w}{\sum_{z \in T_x} n_{z,w} + \beta_w}, x \in \{s, c\} \\
\eta_{d,x} &= \frac{c_{d,x} + \gamma_s^{(x)}}{N_d + \gamma_s^{(s)} + \gamma_s^{(c)}}
\end{aligned} \tag{2.10}$$

where $m_{d,z}$ denotes the number of words in document d assigned to topic z, $n_{z,w}$ denotes the number of occurent word w assigned to topic z, $c_{d,s}$ represents the number of words in document d assigned to local topics, α_z represents the Dirichlet parameter of topic z, β_w represents the Dirichlet parameter of word w. $\gamma_s^{(x)}$, $\gamma_s^{(s)}$ and $\gamma_s^{(c)}$ is the Beta parameter for drawn topics from Bernoulli distribution, local topics and commom topics respectively. N_d is the number of words in document d.

By adding a Bernoulli distribution over topics, this method can separate the local topics and common topics. However, this method may be effected by the large data collection. In the evaluation, we illustrate the experimental results generated from collection model.

Chapter 3

Heterogensous Topic Model

In this chapter, we first introduce the property of conditional independence in directed graphic model, which serves following sections. Then, we propose a novel Heterogeneous Topic Model. Finally, we talk about the model fitting of this model and estimation of the posteriors.

3.1 Conditional Independence in Directed Graphic Model

In this section, we explain the theory of conditional independence, which is served as the basis for heterogeneous topic model.

Figure 3.1a shows the first simple example of conditional independence in directed graphic model. In this example, both variables a and b are conditioned on the variable c (variable c is observed), then the conditional distribution of a and b given c can be easily expressed as follows

$$p(a, b|c) = \frac{p(a, b, c)}{p(a, c)p(b|c)} = p(a|c)p(b|c) \tag{3.1}$$

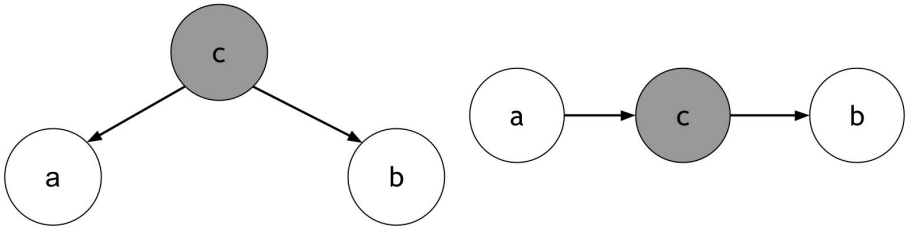
then we can prove a is conditional independent on b given c.

The second example is shown in Figure 3.1b. If $p(a, b|c) = p(a|c)p(b|c)$ can be obtained, then we can say a and b is conditional independent given observed variable c.

Using Bayes' theorem, we have

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = \frac{p(a|c)}{p(b|c)} \tag{3.2}$$

then the conditional independence property is obtained.



(a) (b)

Figure 3.1: Conditional Independence in Graphic Model. Now a and b is conditioned on c

3.2 Heterogenous Topic Model

This section mainly talks about the Heterogeneous Topic Model (HTM). The motivation of this topic model is to solve the problems that metioned in Section 1.2. The underlying initution is that topics within a period of time would receive attentions both from news and social media. By sharing a latent common topic distribution, we can make a correspondence among documents in multiple sources.

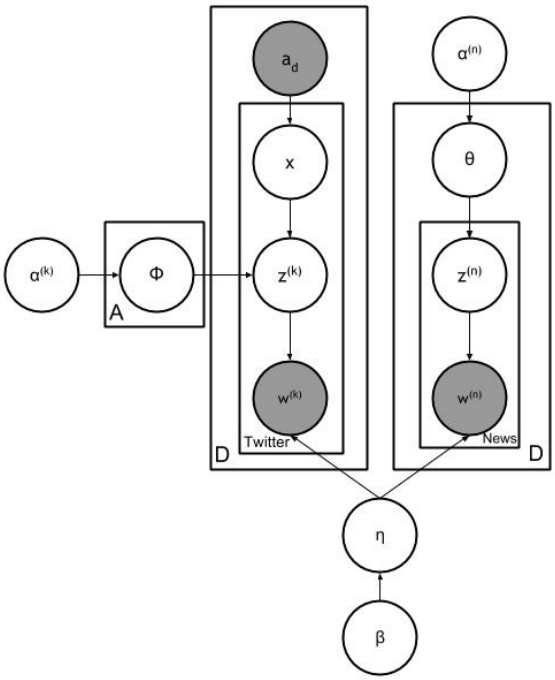


Figure 3.2: Heterogenous Topic Model. The left part model the Twitters and the right part model the News documents

Figure 3.2 illustrates the Heterogeneous Topic Model, which is a combination of ATM

and LDA. The correspondence between topics of different sources is maintained by a shared topic distribution η, which is the common topic distribution for Twitter and News collection. However, the local topic may not be able to preserve all the individual properties for both collections, because the probability of drawing a topic is effected by both collection. In HTM, we have two sets of text streams, including news articles and Twitters messages. Considering the conditional independent properties metioned above, each word $w^{(n)}$ in news articles is only associated with topic z, whlist each word $w^{(k)}$ in tweets is related to two lantent variables: an author a and a topic z.

According to the data collection, the observed variables consist of the set of authors of each tweet, the Twitters messages and the news articles. Since nearly 70% percent of tweet mentions other people (@someone, which is viewed as the Twitter Handler), we assume all of those people that mentioned in the tweet are the authors of this tweet. We also make an assumption that news articles have no authors because most of the news-press focus on the same public event. Then, the interest of the authors for news document does not need to be modeled. Conditioned on the set of authors from Twitters and distribution over topics, we can summarize the process of generating a document as follows: for each tweet, an author is first drawn uniformly at random for each word; a topic of word is chosen based on the shared topic distribution over words; the words are sampled from the topic distribution over words. For news documents, a topic of word is sampled from the shared topic distribution associated with words; next, the word is drawn from the distribution over words associated with topics. Summarizing the notations in Table 3.1, this generative process of the HTM can be expressed more formally as follows (note that $vaiable^{(k)}$ represents the variable of twitters while $variable^{(n)}$ is the variable of news):

1. Initialization

   For each topic $t = 1, \ldots, T$, choose $\eta_t \sim \text{Dirichlet}(\beta)$
   For each author $a = 1, \ldots, A$, choose $\phi_a \sim \text{Dirichlet}(\alpha^{(\text{k})})$
   For each mixture of the topic $m = 1, \ldots, M$, choose $\theta_m \sim \text{Dirichlet}(\alpha^{(\text{n})})$

2. For each common word $i = 1, \ldots, N_d$ in News and Twitter

   (a) Choose a topic $z_{di}^{(n)} \sim \text{Multinomial}(\theta)$
   Choose a word $w_{di}^{(n)} \sim \text{Multinomial}\big(z_{di}^{(n)}, \eta_{z_{di}^{(n)}}\big)$

   (b) Given the vector of authors a_d of Twitter $d_i^{(k)}$
   Choose an author $x_{di} \sim \text{Uniform}(a_d)$
   Choose a topic $z_{di}^{(k)} \sim \text{Multinomial}\big(\phi_{x_{di}}, \theta\big)$
   Choose a word $w_{di}^{(k)} \sim \text{Multinomial}\big(\eta_{z_{di}^{(k)}}\big)$

3. (a) For each document in news collection $d = 1, \ldots, D$
   For each local word $i = 1, \ldots, N_d$
   Choose a topic $z_{di}^{(n)} \sim \text{Multinomial}(\theta)$
   Choose a word $w_{di}^{(n)} \sim \text{Multinomial}\big(z_{di}^{(n)}, \eta_{z_{di}^{(n)}}\big)$

   (b) For each document in twitter collection $d = 1, \ldots, D$
   Given the vector of authors a_d

For each local word $i = 1, \ldots, N_d$
Choose an author $x_{di} \sim \text{Uniform}(a_d)$
Choose a topic $z_{di}^{(k)} \sim \text{Multinomial}\big(\phi_{x_{di}}, \theta\big)$
Choose a word $w_{di}^{(k)} \sim \text{Multinomial}\big(\eta_{z_{di}^{(k)}}\big)$

where the local word represents the word observed in one collection and not observed in another collection. For example, word "teacher" is only shown in News collection and not in Twitter messages. The common word is the word observed in both collection.

Table 3.1: Symbols Associated with the Heterogenous Topic Model

Authors of the corpus in Twitters	$\mathcal{A}$	Set
Number of authors of the document d	A_d	Scalar
Authors of the document d	a_d	A_d-dimensional vector
Number of words in the collection(Twitter and News)	N	Scalar
Number of words in document d	N_d	Scalar
Vocabulary size	W	Scalar
Number of topics	T	Scalar
Number of authors in Twitters	A	Scalar
Number of words assigned to topic and word	C^{WT}	W$\times$ T
Number of words assigned to author and topic	C^{TA}	T$\times$ A
Set of authors and words in the training data	$\mathcal{D}^{\text{train}}$	Set
Words in Twitter document d	$w_d^{(k)}$	N_d-dimensional vector
Words in Twitter document d	$w_{di}^{(k)}$	i^{th} component of $w_d^{(k)}$
Words in News document d	$w_d^{(n)}$	N_d-dimensional vector
Words in Twitter document d	$w_{di}^{(n)}$	i^{th} component of $w_d^{(n)}$
Topic assignment	z	N-dimensional vector
Topic assignment for word w_{di}	z_{di}	i^{th} component of z_d
Author assginments	x	N-dimensional vector
Author assignment for word w_{di}	x_{di}	i^{th} component of x_d
Dirichlet prior for Twitter	$\alpha^{(k)}$	T-dimensional vector
Dirichlet prior of News document	$\alpha^{(n)}$	T-dimensional vector
Dirichlet prior for topic t in News document	$\alpha_t^{(n)}$	Scalar
Dirichlet prior	β	W-dimensional vector
Dirichlet prior of word w	β_w	Scalar
Probabilities of words given on topics	η	W $\times$ T matrix
Probabilities of words given on topic t	η_t	W-dimensional vector
Probabilities of topics given on authors	ϕ	T $\times$ A matrix
Probabilities of topics given on author a	ϕ_a	T-dimensional vector
Number of times that topic t assigned to a word in document d	$n_d^{(t)}$	Scalar
Number of times that word w assigned to topic t	$n_t^{(w)}$	Scalar

The graphical model corresponding to this process is illustrated in Figure 3.2. Note that in this article, we deal with the circumstances where the number of possible topics

T is fixed. Under the generative process, each topic z in News and Twitters is drawn independently when conditioned on Θ and Φ respectively, each word w in news and twitters is drawn independently when conditioned on $\mathcal{H}_z$.

Holding the conditional independence property, we have following basic equation for Gibbs sampler:

$$\begin{aligned}
&p(z_{di,dj} = t | w_{di}^{(k)} = w_i, z_{-di}, x_{-di}, w_{-di}^{(k)}, w_{dj}^{(n)} = w_j, z_{-dj}, w_{-dj}^{(n)}, \mathcal{A}, \alpha^{(k)}, \alpha^{(n)}, \beta) \\
&\propto p(z_{di,dj} = t, w_{di}^{(k)} = w_i, w_{dj}^{(n)} = w_j | z_{-di}, x_{-di}, w_{-di}^{(k)}, z_{-dj}, w_{-dj}^{(n)}, \mathcal{A}, \alpha^{(k)}, \alpha^{(n)}, \beta) \\
&= \frac{p(z, w^{(k)}, w^{(n)} | \mathcal{A}, \alpha^{(k)}, \alpha^{(n)}, \beta)}{p(z_{-di}, z_{-dj}, w_{-di}^{(k)}, w_{-dj}^{(n)} | \mathcal{A}, \alpha^{(k)}, \alpha^{(n)}, \beta)} \\
&= \frac{p(z, w^{(k)} | \mathcal{A}, \alpha^{(k)}, \beta)}{p(z_{-di}, w_{-di}^{(k)} | \mathcal{A}, \alpha^{(k)}, \beta)} \cdot \frac{p(z, w^{(n)} | \alpha^{(n)}, \beta)}{p(z_{-dj}, w_{-dj}^{(n)} | \alpha^{(n)}, \beta)}
\end{aligned} \tag{3.3}$$

where z_{-di}, w_{-di} stand for the vector of topic assignments and word observations except for the i^{th} word of news document d. z_{-dj}, w_{-dj} stand for the vector of topic assignments and word observations except for the j^{th} word of tweet d. We separate the sampling procedures into two parts, consisting of Twitter topic model (left side) and LDA (right side).

For the left part of this heterogeneous topic model, which corresponds to the Twitters data, the joint distribution of $(z, w^{(k)}, x, \Phi, \mathcal{H} | \alpha^{(k)}, \beta, \mathcal{A})$ in tweets data is:

$$\begin{aligned}
p(z, w^{(k)}, x, \Phi, \mathcal{H} | \alpha^{(k)}, \beta, \mathcal{A}) &= p(z, w^{(k)}, x | \Phi, \mathcal{H}, \mathcal{A}) p(\Phi, \mathcal{H} | \alpha^{(k)}, \beta) \\
&= p(z|x, \Phi) p(w^{(k)} | z, \mathcal{H}) p(x | \mathcal{A}) p(\Phi | \alpha^{(k)}) p(\mathcal{H} | \beta)
\end{aligned} \tag{3.4}$$

By integrating over Φ, $\mathcal{H}$ and x, we get:

$$
\begin{aligned}
p(z, w^{(k)}|\mathcal{A}, \alpha^{(k)}, \beta) &= \int_x \int_\Phi \int_\mathcal{H} p(z|x,\Phi) p(w^{(k)}|z,\mathcal{H}) p(x|\mathcal{A}) p(\Phi|\alpha^{(k)}) p(\mathcal{H}|\beta) dx d\Phi d\mathcal{H} \\
&= \int_x \int_\Phi \int_\mathcal{H} \left[\prod_{i=1}^{N} p(z_{di}|\phi_{x_{di}})\right] \left[\prod_{d=1}^{D}\prod_{i=1}^{N_d} p(x_{di}|a_d)\right] p(\Phi|\alpha^{(k)}) p(\mathcal{H}|\beta) dx d\Phi d\mathcal{H} \\
&= \int_x \int_\Phi \int_\mathcal{H} \left[\prod_{a=1}^{A}\prod_{t=1}^{T} \eta_{ta}^{\mathcal{C}_{ta}^{TA}}\right] \left[\prod_{d=1}^{D} (\frac{1}{A_d})^{N_d}\right] \\
&\quad \left[\prod_{t=1}^{T} \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{w=1}^{W} \phi_{wt}^{\beta_w - 1}\right)\right] \left[\prod_{a=1}^{A} \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \prod_{t=1}^{T} \beta_{ta}^{\alpha_t^{(k)} - 1}\right)\right] dx d\Phi d\mathcal{H} \\
&= \int_x \int_\Phi \int_\mathcal{H} \left[\prod_{a=1}^{A}\prod_{t=1}^{T} \eta_{ta}^{\mathcal{C}_{ta}^{TA}}\right] \left[\prod_{d=1}^{D} (\frac{1}{A_d})^{N_d}\right] \\
&\quad \left[\prod_{t=1}^{T}\prod_{w=1}^{W} \eta_{wt}^{C_{wt}^{WT} + \beta_w - 1}\right] \left[\prod_{a=1}^{A}\prod_{t=1}^{T} \phi_{ta}^{C_{ta}^{TA} + \alpha_t^{(k)} - 1}\right] dx d\Phi d\mathcal{H} \\
&= \left[\prod_{a=1}^{A}\prod_{t=1}^{T} \eta_{ta}^{\mathcal{C}_{ta}^{TA}}\right] \int_{x_{:\frac{1}{A_d}}} \left[\prod_{d=1}^{D} (\frac{1}{A_d})^{N_d}\right] dx_{:\frac{1}{A_d}} \int_\mathcal{H} \left[\prod_{t=1}^{T}\prod_{w=1}^{W} \eta_{wt}^{C_{wt}^{WT} + \beta_w - 1}\right] d\mathcal{H} \\
&\quad \int_\Phi \left[\prod_{a=1}^{A}\prod_{t=1}^{T} \phi_{ta}^{C_{ta}^{TA} + \alpha_t^{(k)} - 1}\right] d\Phi \\
&= \left[\prod_{a=1}^{A}\prod_{t=1}^{T} \eta_{ta}^{\mathcal{C}_{ta}^{TA}}\right] \left[\prod_{d=1}^{D} (\frac{1}{A_d})^{N_d+1}\right] \\
&\quad \left[\prod_{a=1}^{A} \frac{\prod_{t=1}^{T} \Gamma(C_{ta}^{TA} + \alpha_t^{(k)})}{\Gamma(\sum_{t'} C_{t'a}^{TA} + T\alpha^{(k)})}\right] \left[\prod_{t=1}^{T} \frac{\prod_{w=1}^{W} \Gamma(C_{wt}^{WT} + \beta_w)}{\Gamma(\sum_{w'} C_{w't}^{WT} + W\beta)}\right]
\end{aligned}
\tag{3.5}
$$

where C_{wt}^{WT} denotes the number of times that word w in the corpus is assigned to t^{th} topic and C_{ta}^{TA} is the number of times that topic t is assigned to author a. Then we can use the same approach as Equation (3.5) to have:

$$
\begin{aligned}
p(z_{-di}, w_{-di}^{(k)}|\mathcal{A}, \alpha^{(k)}, \eta) &= \left[\prod_{a=1}^{A}\prod_{t=1}^{T} \eta_{ta}^{\mathcal{C}_{ta}^{TA}}\right] \left[\prod_{d=1}^{D} (\frac{1}{A_d})^{N_d+1}\right] \\
&\quad \left[\prod_{a=1}^{A} \frac{\prod_{t=1}^{T} \Gamma(C_{ta,-di}^{TA} + \alpha_t^{(k)})}{\Gamma(\sum_{t'} C_{t'a,-di}^{TA} + T\alpha^{(k)})}\right] \left[\prod_{t=1}^{T} \frac{\prod_{w=1}^{W} \Gamma(C_{wt,-di}^{WT} + \eta_w)}{\Gamma(\sum_{w'} C_{w't,-di}^{WT} + W\eta)}\right]
\end{aligned}
\tag{3.6}
$$

According to Equation (3.5) and (3.6), the following equation can be obtained:

$$
\frac{p(z, w^{(k)}|\mathcal{A}, \alpha^{(k)}, \beta)}{p(z_{-di}, w_{-di}^{(k)}|\mathcal{A}, \alpha^{(k)}, \beta)} = \frac{C_{wt,-di}^{WT} + \beta_w}{\sum_{w'} C_{w't,-di}^{WT} + W\beta} \cdot \frac{C_{ta,-di}^{TA} + \alpha_t^{(k)}}{\sum_{t'} C_{t'a,-di}^{TA} + T\alpha^{(k)}} \tag{3.7}
$$

Similarly, for news articles, $p(z, w^{(n)}|\alpha^{(n)}, \beta)$ can be calculated as follows:

$$
p(z, w^{(n)}|\alpha^{(n)}, \beta) = p(w^{(n)}|z, \beta) p(z|\alpha^{(n)}) \tag{3.8}
$$

where $p(w^{(n)}|\alpha^{(n)},\beta)$ and $p(z|\alpha^{(n)})$ can be obtained according to [5] as follows:

$$p(w^{(n)}|z,\beta)=\prod_{t=1}^{T}\frac{\Delta(n_t+\beta)}{\Delta(\beta)},\quad n_t=\{n_t^{(w)}\}_{t=1}^{V}. \tag{3.9}$$

$$p(w^{(n)}|z,\alpha^{(n)})=\prod_{d=1}^{D}\frac{\Delta(n_d+\alpha^{(n)})}{\Delta(\alpha^{(n)})},n_d=\{n_d^{(t)}\}_{t=1}^{T} \tag{3.10}$$

After integrating, we can get:

$$p(z,w^{(n)}|\alpha^{(n)},\beta)=\prod_{t=1}^{T}\frac{\Delta(n_t+\beta)}{\Delta(\beta)}\cdot\prod_{d=1}^{D}\frac{\Delta(n_d+\alpha^{(n)})}{\Delta(\alpha^{(n)})} \tag{3.11}$$

By using the same method, we have:

$$p(z_{-di},w_{-di}^{(n)}|\alpha^{(n)},\beta)=\prod_{t=1}^{T}\frac{\Delta(n_{t,-i}+\beta)}{\Delta(\beta)}\cdot\prod_{d=1}^{D}\frac{\Delta(n_{d,-i}+\alpha^{(n)})}{\Delta(\alpha^{(n)})} \tag{3.12}$$

then target joint distribution of words and topics can be obtained:

$$\begin{aligned}\frac{p(z,w^{(n)}|\alpha^{(n)},\beta)}{p(z_{-di},w_{-di}^{(n)}|\alpha^{(n)},\beta)}&=\frac{\Delta(n_t+\beta)}{\Delta(n_{t,-i}+\beta)}\cdot\frac{\Delta(n_d+\alpha^{(n)})}{\Delta(n_{d,-i}+\alpha^{(n)})}\\&\propto\frac{\Gamma(n_t^{(w)}+\beta_w)\Gamma(\sum_{w=1}^{V}n_{t,-i}^{(w)}+\beta_w)}{\Gamma(n_{t,-i}^{(w)}+\beta_w)\Gamma(\sum_{w=1}^{V}n_t^{(w)}+\beta_w)}\cdot\frac{\Gamma(n_d^{(t)}+\alpha_t)\Gamma(\sum_{t=1}^{T}n_{d,-i}^{(t)}+\alpha_t)}{\Gamma(n_{d,-i}^{(t)}+\alpha_t)\Gamma(\sum_{t=1}^{T}n_d^{(t)}+\alpha_t)}\\&\propto\frac{n_{t,-i}^{(w)}+\beta_w}{\sum_{w=1}^{V}n_{t,-i}^{(w)}+\beta_w}\cdot\frac{n_{d,-i}^{(t)}+\alpha_t}{[\sum_{t=1}^{T}n_d^{(t)}+\alpha_t]-1}\end{aligned} \tag{3.13}$$

where $n_{t,-i}^{(w)}$ denotes the number of times that word w except i assigned to topic t, $n_{d,-i}^{(t)}$ represents the number of times that topic t except i assgined to a word in document d.

Finally, we get the following basic equation for Gibbs Sampler:

$$\begin{aligned}&p(z_{di,dj}=t|w_{di}^{(k)}=w_i,z_{-di},x_{-di},w_{-di}^{(k)},w_{dj}^{(n)}=w_j,z_{-dj},w_{-dj}^{(n)},\mathcal{A},\alpha^{(k)},\alpha^{(n)},\beta)\\&=\frac{C_{wt,-di}^{WT}+\beta_w}{\sum_{w'}C_{w't,-di}^{WT}+W\beta}\cdot\frac{C_{ta,-di}^{TA}+\alpha^{(k)}}{\sum_{t'}C_{t'a,-di}^{TA}+T\alpha^{(k)}}\cdot\frac{n_{t,-i}^{(w)}+\beta_w}{\sum_{w=1}^{V}n_{t,-i}^{(w)}+\beta_w}\cdot\frac{n_{d,-i}^{(t)}+\alpha_t}{[\sum_{t=1}^{T}n_d^{(t)}+\alpha_t]-1}\end{aligned} \tag{3.14}$$

note that, the sampled word from one collection may not be observed in another collection. As a result, the prior probability of topic over word in another collection is treated as 1.

3.3 Model Fitting

Gibbs sampling is a form of the Markov chain Monte Carlo for obtaining a sequence of observation when the direct sampling is hard. In our case, a Markov chain is constructed

to converge the posterior distribution over topic z conditioned on $\mathcal{D}^{\text{train}}$, $\alpha^{(n)}$, $\alpha^{(k)}$ and η. By Gibbs sampler, we can get a topic assignement z for a word w from the topic distribution $p(z|\mathcal{D}^{\text{train}}, \alpha^{(n)}, \alpha^{(k)}, \eta)$. After a series of iterations, a particular stationary topic distribution conditioned on training data and Dirichlet parameters can be obtained.

Given the basic Equation (3.14) for Gibbs Sampling, we will introduce how the algorithm works. Firstly, the vector of topic assignment z for the whole collection and the vector of author assginment x for tweets is initialized with random numbers. In each iteration, we draw the topic and author assignment of i^{th} word in Tweet $d^{(k)}$ and draw the topic assignment of i^{th} word in News document $d^{(n)}$ from the Equation (3.14). After numbers of interations, the posterior distribution $p(z|\mathcal{D}^{\text{train}}, \alpha^{(n)}, \alpha^{(k)}, \beta)$ tends to stabilize.

3.3.1 The posterior on $\mathcal{A}$, Θ and Φ

For the purpose of estimation, $(C^{WT})^{(n)}$, $(C^{WT})^{(k)}$ represents the sample of word-topic matrix in which each word is observed only in Twitter and only in News respectively, whlist (C^{WT}) is the sample of word-topic matrix where each word can be observed in both collections. Then, due to the fact that the Dirichlet distribution is conjugate to the Multinomial distribution, we can have:

$$\begin{aligned}
\beta_t^{(n)}|\mathbf{z}, \mathcal{D}^{\text{train}}, \beta &\sim \text{Dirichlet}\big(C_t^{WT} + (C_t^{WT})^{(n)} + \beta\big) \\
\beta_t^{(k)}|\mathbf{z}, \mathcal{D}^{\text{train}}, \beta &\sim \text{Dirichlet}\big(C_t^{WT} + (C_t^{WT})^{(k)} + \beta\big) \\
\beta_t|\mathbf{z}, \mathcal{D}^{\text{train}}, \beta &\sim \text{Dirichlet}(C_t^{WT} + (C^{WT})^{(n)} + (C^{WT})^{(k)} + \beta) \\
\phi_a|\mathbf{x}, \mathbf{z}, \mathcal{D}^{\text{train}}, \alpha^{(k)} &\sim \text{Dirichlet}(C_a^{TA} + \alpha^{(k)}) \\
\theta_d|\mathbf{w}, \mathbf{z}, \mathcal{D}^{\text{train}}, \alpha^{(n)} &\sim \text{Dirichlet}(n_d + \alpha^{(n)})
\end{aligned} \tag{3.15}$$

where $\beta_t^{(n)}$ represents the local topics that belong to News, $\beta_t^{(t)}$ is the local topics that belog to Twitter, which is also then the posterior mean of $\mathcal{A}$, Θ and Φ given $\mathbf{x}$, $\mathbf{z}$, $\mathbf{w}$, $\mathcal{D}^{\text{train}}$, $\alpha^{(n)}$, $\alpha^{(k)}$ and β can be obtained as follows:

$$\begin{aligned}
E[\beta_{wt}^{(n)}|\mathbf{z}^s, \mathcal{D}^{\text{train}}, \beta] &= \frac{\big(C_{wt}^{WT} + (C_{wt}^{WT})^{(n)}\big)^s + \beta_w}{\sum_{w'}\big(C_{w't}^{WT} + (C_{w't}^{WT})^{(n)}\big)^s + W\beta} \\
E[\beta_{wt}^{(k)}|\mathbf{z}^s, \mathcal{D}^{\text{train}}, \beta] &= \frac{\big(C_{wt}^{WT} + (C_{wt}^{WT})^{(k)}\big)^s + \beta_w}{\sum_{w'}\big(C_{w't}^{WT} + (C_{w't}^{WT})^{(k)}\big)^s + W\beta} \\
E[\beta_{wt}|\mathbf{z}^s, \mathcal{D}^{\text{train}}, \beta] &= \frac{\big(C_{wt}^{WT} + (C_{wt}^{WT})^{(n)} + (C_{wt}^{WT})^{(k)}\big)^s + \beta_w}{\sum_{w'}\big(C_{w't}^{WT} + (C_{w't}^{WT})^{(n)} + (C_{w't}^{WT})^{(k)}\big)^s + W\beta} \\
E[\phi_{ta}|\mathbf{z}^s, \mathbf{x}^s, \mathcal{D}^{\text{train}}, \alpha^{(k)}] &= \frac{(C_{ta}^{TA})^s + \alpha^{(k)}}{\sum_{t'}(C_{t'a}^{TA})^s + T\alpha^{(k)}} \\
E[\theta_{dt}|\mathbf{w}^s, \mathbf{z}^s, \mathcal{D}^{\text{train}}, \alpha^{(n)}] &= \frac{\big(n_d^{(t)}\big)^s + \alpha_t}{\sum_{t=1}^{T}\big(n_d^{(t)}\big)^s + \alpha_t},
\end{aligned} \tag{3.16}$$

where s refers to sample s from Gibbs sampler of the whole collection. These posterior means correspond to the topic distribution on words, author distribution on topics and document distribution on topics repsectively.

Chapter 4

Experiments and Evaluation

In this section, we detail the experiments and evaluation of this heterogeneous topic model. We first introduced the dataset that used in the experiments. Then the evaluation metrics and results about the experiments are described.

4.1 Data Collection

The data is provided by Glasgow Memory Server (GMS) system, which belongs to University of Glasgow. This data collection consists of Twitter messages and news articles. The news articles were crawled from BBC, EveningTimes, DailyRecord and Scotsman from July 2015. Both of Twitter messages and news articles are only contains the information about Scotland. Since the original Twitter dataset is quite large, we sample Tweets 10% of the whole collection for each hour, resulting in 1,384,259 Twitter messages. The orginal number of news articles is 2371 in total, which is quite smaller than Twitter messages.

4.2 Evaluation

Basically, we followed the evaluation methods, which is introduced in [7], [4], [9]. The model used in evaluation includes (1) HTM (Section 3.2); (2) LDA, which is treated as our baseline; (3) CM, which is the state-of-the-art topic model. Note that all Dirichlet parpameters are set to 0.5.

4.2.1 Perplexity

The original perplexity is defined as a measurement of how well a probability model can predict a new coming document. A low value of perplexity is an indicator of good performance for the topic model that is being evaluted. The basic equation is:

$$perplexity = exp - \frac{\sum_{d=1}^{D}\sum_{i=1}^{N_d}\log p(w_{d,i}|\mathcal{D}^{\text{train}}, \alpha^{(n)}, \alpha^{(k)}, \beta)}{\sum_{d=1}^{D} N_d} \tag{4.1}$$

where $w_{d,i}$ represents the i^{th} word in document d. Note that the perplexity is defined by summing over the documents.

However, due to the different results generated from the model, the equations differ from each other. For example, the outputs of LDA are document-topic distribution, and topic-word distribution, while ATM generates topic-word and author-topic distributions. The approximate equations of $p(w_{d,i}|\mathcal{D}^{\text{train}}, \alpha^{(n)}, \alpha^{(k)}, \beta)$ of LDA and ATM are:

$$\begin{aligned} P_{\text{LDA}}(w_{d,i}|\mathcal{D}^{\text{train}}, \alpha^{(n)}, \alpha^{(k)}, \beta) &= \sum_{t=1}^{T} \theta_{dt}\eta_{tw} \\ P_{\text{ATM}}(w_{d,i}|\mathcal{D}^{\text{train}}, \alpha^{(n)}, \alpha^{(k)}, \beta) &= \sum_{t=1}^{T} \phi_{ta}\eta_{tw} \end{aligned} \tag{4.2}$$

As a result, the different equations will lead to different results, which can not be used to compared directly. In order to make a consistence between LDA and HTM. The approximate estimate of perplexity is changed by summing over the topics:

$$perplexity = exp - \frac{\sum_{t=1}^{T} \log p(w_{d,i}|\mathcal{D}^{\text{train}}, \alpha^{(n)}, \alpha^{(k)}, \beta)}{\sum_{d=1}^{D} N_d} \tag{4.3}$$

where $p(w_{d,i}|\mathcal{D}^{\text{train}}, \alpha^{(n)}, \alpha^{(k)}, \beta) = \eta_{tw}$. After changing the equation of perplexity, we can use the same equation to evaluate how well the topic model can predict a topic.

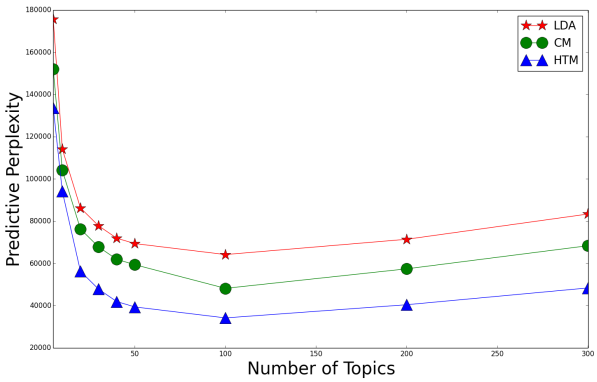


Figure 4.1: Comparison of perplexity between LDA, CM HTM

Figure 4.1 shows the results of LDA, CM and HTM. In order to make a general comparison, we generate 5, 10, 20, 30, 40, 50, 100, 200 and 300 topics separately. As is evident in the figure, the perplexity of HTM exhibits a minimum with respect to the number of topics in the setting. According to the results of T-Test, the value between LDA and HTM is 0.072 while that between CM and HTM is 0.266. The results suggest that the performance of HTM is significant better than LDA and CM.

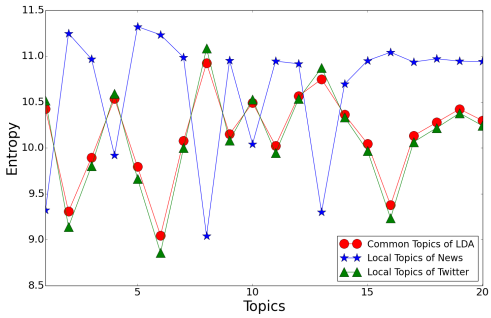
4.2.2 Entropy

Another way to evaluate the model is entropy. Entropy is defined as the expected value of information contained in the message. The smaller the value is, the more specific the topic is. In other words, there are bunches of documents talking about the same topic. We define the following equation to compute the topic entropy in HTM and approximate that for News and Twitter messages separately.

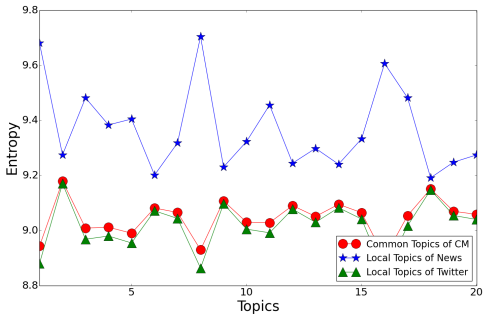
$$entropy = exp - \frac{\sum_{d=1}^{D} \frac{1}{N_d} \sum_{i=1}^{N_d} \log p(z|\mathcal{D}^{\text{train}}, \alpha^{(n)}, \alpha^{(k)}, \beta)}{N} \tag{4.4}$$

Note that the entropy of news articles and Twitter messages are approximate estimated since they effect each other in Gibbs sampling procedure.

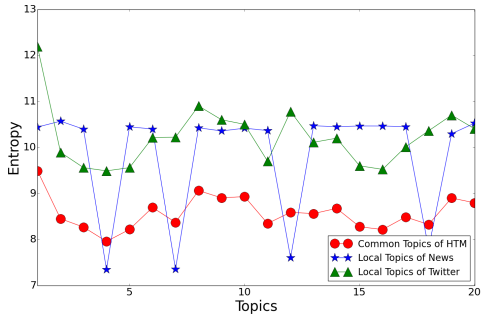
Figure 4.2 illustrates the entropy of LDA, CM and HTM on the collection. The number of topics is fixed to 20. The first three graphs shows the comparison of entropy of Twitter, News and whole collection for each model. From Figure 4.2a and 4.2b, we can find that the entropy of common topics in LDA and CM are heavily impacted by the large volumn of Twitter collection. According to Figure 4.2c, the curve of common topics is not effected by the larger dataset, from which we can believe that the HTM can handle the problems caused by the uneven data collection. Figure 4.2d compare the entropy of HTM, CM and LDA. Although the 20 topics of in each model are different, the entropy of HTM is lower than that of LDA and CM overall. In other words, the topic generated from HTM is more specific.



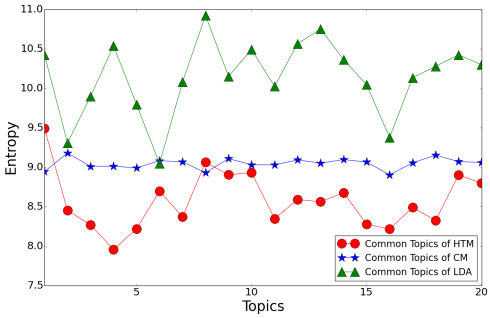
(a) LDA



(b) CM



(c) HTM



(d) A comparison between LDA and HTM

Figure 4.2: Comparison of entropy between different models and different sources

The value of T-Test of entropy between HTM and LDA is 3.99×10^{-14} while that between HTM and CM is 1.49×10^{-6}. Both of the values are lower than 0.5. In other words, there is no significant difference in entropy among these different models.

4.3 Case Study on Topics

In this section, we first illustrate the 20 topics generated by LDA, CM and HTM. Then a comparison and discussion is followed.

Topics	Words
1	cfc ferguson chelsea sir happy chelseafc falcao teacher tree alex
2	police mr year man court scotland family day glasgow incident
3	give daddy gift miley fortunately hamilton gemini joined answer cyrus
4	scotland year scottish pound glasgow work government time service council
5	rain weather friday tomorrow today whale saturday night temperature lightning
6	loch ness monster nessie x answer catfish white feltham food
7	baby island good faroe feel sun worth black summer wearing
8	duffy sand adair hughes mccrory mad ireland republican kill plot
9	game celtic club team season ranger player win league football
10	true map google sleep http idk streetviewpov var awesome getstreetview
11	im pride drag gay idea niallofficial lad aku yg thevampscon
12	hamilton lewis rosberg yang gp pole britishgp silverstone bbcsport mercedes
13	ayr en la bir ca ya da de mo ah
14	hell ah fish xd glad hurt bye friend pls boyfriend
15	world festival show day life event edinburgh year great art
16	bird seagull nest chick puffin gull egg seabird eagle nesting
17	fish amp love u ranger irvine nh hamilton mom read
18	snp scottish vote scotland party labour english referendum election minister
19	milk camel cosby funny love cow sweet wait babe skye
20	ko finally agrave ang hot shy air lang ako ng

Table 4.1: Top 10 words for each topic in LDA

Topics	Words
1	glasgow year world event people game work edinburgh life day
2	amp irvine nh tonight job read hamilton full doe great
3	lorry bin clarke driver harry inquiry telford crash hit truck
4	amp love follow lt hell bad cute hey fuck ppl gift wait
5	police mr year man court family glasgow scotland hospital found
6	scotland scottish year people pound government uk mr council snp
7	doe gurmeetramrahim ji bro lol beautiful sex mine babe wait
8	ranger power texas v angel karachi mlb news pls card
9	ayr n l k de hamilton r b y g
10	fish amp im ur shit eat whale hot w omg
11	love cfc ferguson omg chelsea mom amp happy chelseafc today
12	hamilton lewis rosberg yang gp pole britishgp silverstone bbcsport mercedes
13	god x video thing omg true love yeah finally wait
14	hell ah fish xd glad hurt bye friend pls boyfriend
15	celtic good club season game team time player win ranger
16	hell haha omg dinosaur x wing xd ah feather cute
17	duffy watch adair sand hughes kill mccrory ireland mad pls
18	baby ice daddy lol agrave shy haha wedding air gonna
19	tweet ah lol greek twitter vote angus apple black obama
20	service day road fire glasgow scotland water train park driver

Table 4.2: Top 10 words for each topic in CM

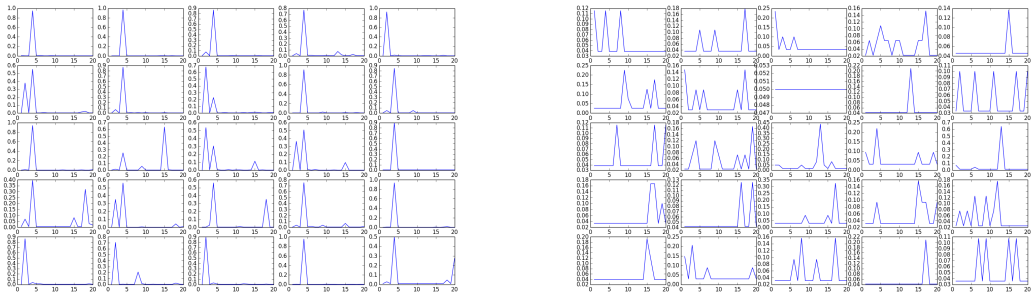
Topics	Words
1	blanket defamation testimony robreno suing iphone debit settlement valve sedative
2	accuser deposition constand quaalude card troiani lawsuit apple contactless cosby
3	gaelic gaidhlig ann bheil fo alba idhlig igrave agus gu ograve
4	ocean scone peel bae feltham catfish dry rubik carnegie monster
5	cube inverclyde nessie yard ferguson dock ship inchgreen mccoll port
6	bonnar thorniewood mortonhall elsie yau southeby dalmeny swapping phrase methandienone
7	gunman ann islamic british cumbernauld sousse extremism silence briton terrorist
8	ice graduation outsmarting hapless assignment kanye sepp piercy sandstone jack
9	unit procedure board ward transplant practice nurse gp royal bed
10	dolphin gull snh conservation puffin isle animal seabird fishery breeding
11	renewables ferguson carbon panel runway cairngorm electricity application wild renewable
12	happiness slut cow nazi stag commute stephenson happiest milk camel
13	corrigan egyptian angelina highlander foul mouthed tobi mango tweeter jolie
14	telford car bus tragedy review traffic lamara yuill control fire
15	bolton employed tartan drunk salvaged benedetti anticipation brink impaling admirable
16	hebrides union abellio industrial aslef dispute rmt strike railway passenger
17	animal cousin larger suggests geological bird fossil meat scientist organism bamboo
18	window function false com zoom roof true streetviewlatlng mylatlng myoptions
19	lesbian bisexual performer discrimination transgender trans gay lgbt drag pride
20	tuesday cider australian uncomfortable racism homophobic stonewall foster lgbtqia alternative

Table 4.3: Top 10 words for each topic in HTM

Table 4.1 shows the 20 topics generated by LDA model. Some topics give a good summarization for the collection. For example, Topic 1 talks about the Chelsea.F.C, for which Alex is playing. Topic 5 focus on the weather and temperature. However, some topics seem to be useless, such as Topic 13, Topic 14 and Topic 20. These topics heavily suffered the problem caused by vast numbers of Twitter messages. From table 4.2, although some topics in CM are meaningful, we can also find that the topics' words generated from CM have the same problems with LDA.

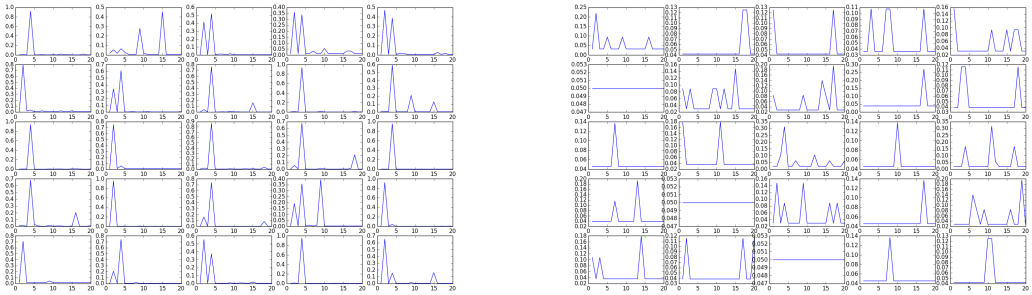
Table 4.3 illustrates the 20 topics in HTM. Compared with LDA and CM, more meaningful top words are contained in each topic, which is more useful to represent the topic. For instance, Topic 4 discusses the stuff related to the ocean. Topic 10 talks about the animinal.

The follow figures shows the document distribution over topics in LDA model. Since the collection is very large, 25 news articles and 25 tweets are randomly selected from the collection.



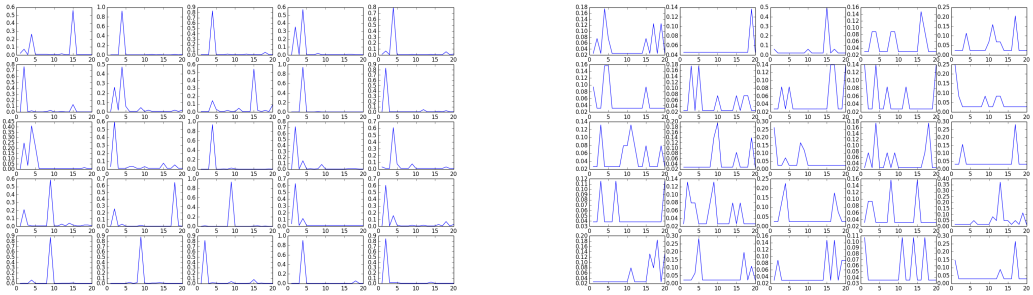
(a) (b)

Figure 4.3: The distribution of 25 news articles and 25 tweets over 20 topics in LDA model. X-axis represents topic, Y-axis represents probability of belonging to this topic.



(a) (b)

Figure 4.4: The distribution of 25 news articles and 25 tweets over 20 topics in CM model. X-axis represents topic, Y-axis represents probability of belonging to this topic.



(a) (b)

Figure 4.5: The distribution of 25 news articles and 25 tweets over 20 topics in HTM model. X-axis represents topic, Y-axis represents probability of belonging to this topic.

According to figure 4.3, 4.4 and 4.5, most of these news articles only have one spike, which are well classified to one topic, while most of the tweets are multifaceted. Using the highest probability as the probability of news articles and tweets, the average probability of news articles and tweets in LDA is 70.82% and 18.6% respectively, while that in CM is 70.31% and 19.9% respectively. In HTM, the average value of news articles 71.9%. For each tweet, the probability is computed using $p(z_d|A_d, W_d)$, where z_d, A_d and W_d represents the topic, authors and words of this document. The result shows that the average value is 23.7%, which is higher than that in LDA.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this paper, we have proposed a novel heterogeneous topic model on news articles and Twitter messages. By combining the LDA and ATM, this model not only preserves the individual properties of each source, but also make a correspondence between the local topics in each source and the common topics. Comparing the entropy, perplexity and average probabilty, our heterogeneous topic model outperforms the LDA and CM. According to the T-tests, the entropy of HTM is significant lower than than that of LDA and CM. In other words, the topics generated by our model is more specific than that LDA and CM. We modified the perplexity to evaluate how well the model can predict a topic, overcoming the comparable problems between different models. The results shows that our model have a better performance in prediction. Additionally, the average probability shows that the documents trained by HTM are better classified than that of LDA and CM.

5.2 Future Work

In the future work, the idea of heterogeneous topic model can be incorporated with news events detection. Since this heterogeneous topic model maintains the correspondence between the local topics and common topics, we can easily find on which topic both sources focus by computing the probability of topic and KL divergence between common topic and local topic. If the topic has a high topic's probability and low KL divergence, then we can say that this topic may contain a news event. Figure 5.1 illustrates the topic's probability and KL divergence.

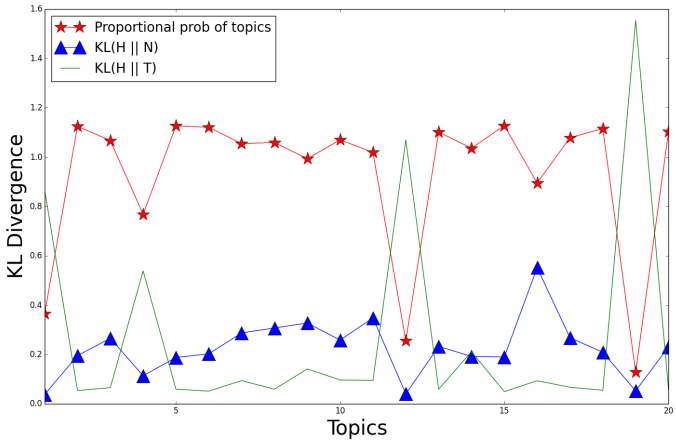


Figure 5.1: KL divergence between word-topic distributions and probabilities of topics

However, it is still unknow whether this heterogeneous topic model can have a better performance on news event detection. The evaluation proposal is doing a case study on Twitter Hashtags, which can help us define the news events.

Bibliography

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[2] Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. Modeling hidden topics on document manifold. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 911–920, New York, NY, USA, 2008. ACM.

[3] Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu, and Cindy Xide Lin. 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*, KDD '11, pages 1271–1279, New York, NY, USA, 2011. ACM.

[4] Rumi Ghosh and Sitaram Asur. Mining information from heterogeneous sources: A topic modeling approach. In *Proc. of the MDS Workshop at the 19th ACM SIGKDD (MDS-SIGKDD)*, 2013.

[5] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[6] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 289–296, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[7] Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 832–840, New York, NY, USA, 2011. ACM.

[8] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 101–110, New York, NY, USA, 2008. ACM.

[9] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Trans. Inf. Syst.*, 28(1):4:1–4:38, January 2010.

[10] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[11] ChengXiang Zhai, Atulya Velivelli, and Bei Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 743–748, New York, NY, USA, 2004. ACM.