A Survey on Statistical Topic Modeling

Li Ren
Department of Computer & Information Sciences
University of Delaware

Abstract

Nowadays, as majority of literature material become digitized and is stored in the database, other than read and study the material manually, researchers need powerful tools to deal articles in thematic layer. Here comes to *topic modeling* - a group of algorithms that extra discover and annotate thematic information from documents. In the past decade, the technique on topic modeling has made a great improvement as the modern statistical topic modeling appearance. In this survey, I introduce some main works on statistical topic modeling technique, including their structure and inference techniques, and introduce some novel techniques that appeared recently.

1 Introduction

Topic models are algorithms that can discover the thematic information from a collection of documents. People have been focusing on topic model for quite a long time. The original purpose of topic modeling is to analyze and classify the semantic layer of large document collection. Nowadays the topic model has been applied to model data from varied fields, including text mining, searching technology, software technology, computer vision, bio-informatics, finance and even social sciences.

In recent decade, the topic modeling has a significant improvement according to development of probabilistic methods, especially the exploration on the application of Bayesian latent variable models. Developed from language models, the latent variable models are generative models, which assuming the documents and words are generated by a series generative process. Basically, they are joint probabilities with latent variables which are hidden from us and needed to compute and update from the learning process. They are called Bayesian latent variable models because the techniques of Bayesian inference are applied to compute the distribution of latent variables.

We used classical topic *latent Dirichlet allocation* (LDA)[2] as example. In LDA, a topic is defined as a distribution over a vocabulary. We assume that the topics are predefined before any data is generated. Then, a document is generated as follows: First randomly choose a topic distribution of this document. Then, for each word location in this document, randomly assign a topic from the distribution of the topic we chose before. Finally, for that topic we assigned corresponding to the word distribution over vocabulary, we randomly choose a word. In this model, the latent variables are the proportion of topics and topic assignment for each word. The only observed data is the set of words in document.

In statistics, the Bayesian inference is the process to compute the posterior distribution when the prior distributions, a distribution of parameters before data is observed, are given. Originally, the posterior distribution can be described as follows:

$$P(\theta \mid X, \alpha) = \frac{P(X \mid \theta)P(\theta \mid \alpha)}{\int_\theta P(X \mid \theta)P(\theta \mid \alpha)d\theta}$$

The numerator is the product of likelihood $P(X \mid \theta)$ and the prior distribution $P(\theta \mid \alpha)$, our hypothesis. The denominator is called *model evidence* or *marginal likelihood* which is the factor that all possible hypotheses are considered. Normally, in latent variable models, the numerator is easy to compute but the computation of the denominator is always intractable. Thus we have to use more complex technique to estimate the posterior distribution.

In this survey, I present a series of probabilistic topic models and their related techniques which was explored in last decade. They are not only include the classical topics model (PLSA, LDA, CTM, PAM), but also include some recent novel topic models (Biterm & NTSeg). The structure of models, the generative processes, and the related Bayesian inference technique are

introduced and explained.

2 Classic Topic Models

Latent Dirichlet Allocation is one of the most classical approaches used today. It was first presented as a graphical model for topic discovery in 2003[2]. This model was implemented based on mixture models and use Dirichlet distribution as its prior of some parameters.

In this section, we will first introduce the probabilistic approach on document indexing base on latent class model, which described each word in a document as the sample of mixture topics, then describe the classic graphic model base on Dirichlet distribution and its variants.

2.1 Aspect model

The *Probabilistic Latent Semantic Analysis(PLSA)*[5] is a solid statistical foundation on automated document indexing based on the likelihood principle. It defines a generative model for factor analysis of count data.

The basic statistical model of pLSA is *aspect* model[6], where observations sharing the same class are referred as an aspect. In this model, data is associated with unobserved class variable $z \in Z = \{z_1, \dots z_k\}$, with each appearance of word $w \in W = \{w_1, \dots, w_m\}$ in document $d \in D = \{d_1, \dots, d_n\}$. The joint probability model can be showed as follows way:

$$P(d,w) = P(d) \sum_{z \in Z} P(w \mid z) P(z \mid d)$$

This aspect model is depended on two assumptions: One is the observations of (d,w) are assumed to be independently, which is also called *bag-of-words* approach. Another assumes that there is conditional independence between document d, words w, given class variable z, which means on latent class z, words w are generated independently of the specific document identity d. To procedure the maximum likelihood estimation of the latent variable model, an Expectation Maximization(EM) algorithm, *tempered EM(TEM)* is applied. The main improvement from TEM is adding a control parameter β over the estimation of posterior calculation in order to avoid overfitting. However, pLSI is criticised not to be a proper generative model of documents because variable d is multinomial random as many possible values as there in training set and the model learn topic mixtures $p(z \mid d)$ only for documents on which it is trained.[2] Another problem is the overfitting problem caused by the increasing number of parameters with documents set.

2.2 Latent Dirichlet Allocation

Here comes to the Latent Dirichlet Allocation(LDA), which overcomes both problems of previous model by considering the topic mixture weights k-parameter hidden random variable rather than individual parameters linked to training documents. In this section, we first introduce the core statistical model of LDA and then describe two popular inference algorithms.

2.2.1 LDA model

We treat LDA as a generative probabilistic model where our data arise from a generative process that include hidden variables. Generally, the generative model may have hidden structure inside, and we will defines a joint distribution over all observed and latent variables. In LDA, the observed variables are the words in documents, and the hidden variable are topic distributions. The process to compute the topic structure is the inference of the posterior distribution of LDA. For description of the model, we denote θ_i as the topic variable of document i, z_n as the topic for nth word in specific topic, w_n as the nth observed word in document, and α and β as the parameter of the Dirichlet prior. The joint probability of the generative model is given by:

$$p(\theta_i, z, w \mid \alpha, \beta) = \prod_{i=1}^{M} p(\theta_i \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta_i) p(w_n \mid z_n, \beta)$$

Figure 1 shows a number of dependencies from the distribution, such that topic assignment z_n for word depends on the document level topic distribution θ_i, and the observed word w_n depends on the topic z_n and the distribution prior β.

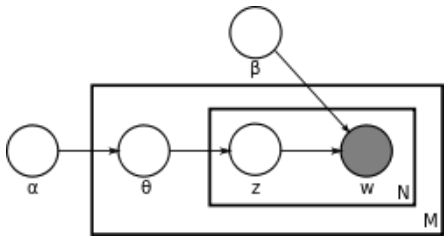

Figure 1: The plate notation representing the LDA model

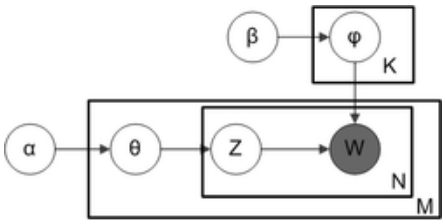

Figure 2: The Plate notation for smoothed LDA

To avoid zero probability on new documents when produce maximum likelihood estimates, smoothing need to be applied. In LDA, an extended model, shown in Figure 2, is applied. It treat β as a set of k random matrix and each one is independently drawn from a exchangeable Dirichlet distribution. The parameter $/phi$ is eventually the multinomial distribution of specific topic over K latent topics on the vocabulary.

The generative process of smoothed LDA can be presented as follows procedure:

1. Sample topic distribution θ_i from $Dir(\alpha)$ for each document i

2. Sample the word distribution ϕ_k from $Dir(\beta)$ for each topic k

3. For each document i and the word position j

   (a) Sample a topic assignment $z_{i,j}$ from $Categorical(\theta_i)$

   (b) Sample a word $w_{i,j}$ from $Categorical(\phi_{z_{i,j}})$

where $i \in \{1,...,M\}$ and $j \in \{1,...,N_i\}$.

To learn the structure of latent topic, the conditional distribution of the topic variable by given observed documents, which is the posterior of the joint distribution, should be computed. The notation of the posterior is as follows:

$$p(\beta,\theta,z \mid w,\alpha,\eta) = \frac{p(\beta,\theta,z,w,\mid \alpha,\eta)}{p(w \mid \alpha,\eta)}$$

It is always intractable to computer the denominator, the marginal probability of the observed evidence, of the posterior because it is difficult of integrate every possible hidden topic structure for every word in the corpus. Thus, approximate inferences should be applied on this problem. In following sections, two main approach of the approximate inference, the stochastic and the structure approach, will be described for the inference of LDA.

2.2.2 Mean field variational inference

Variational methods is one of the most popular structural approximate inference. It posit a parameterized family of distributions over the hidden structure and then locate the member of family that is closest to the posterior. In this case, we can obtain the family of distributions on the latent variables:

$$P(\theta,z \mid \gamma,\phi) = q(\theta \mid \gamma)\prod_{n=1}^{N} q(z_n \mid \phi_n)$$

Then they use KL divergence to evaluate the lost of this approximate. Thus the problem become a optimization problem that determines the values of the parameters γ and ϕ. This two kind of parameters can be updated by following equations to their convergent value,

$$\phi_{ni}^{t+1} := \beta_{iw_n} exp(\Psi(\gamma_i^t))$$

$$\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$$

The variation distribution is actually a conditional distribution with variable w^*, the set of N w. Thus both the Dirichlet parameter and the multinomial parameters can be considered as the function of w^*. The variational distribution can be write as $q(\theta,z \mid \gamma(w^*),\phi(w^*))$, which is dependent on explicit variable w^*. Therefore it can be viewed as the approximation of the posterior distribution of LDA.

For the parameter estimator of α and β, the EM algorithm can be applied to find the maximized log likelihood of data:

$$l(\alpha,\beta) = \sum_{d=1}^{M} \log p(w_d \mid \alpha,\beta)$$

Since the quantity of $p(w_d \mid \alpha,\beta)$ cannot be computed tractably, we have to estimate the maximum likelihood respect to parameter α and β step by step in EM procedure. Basically the each iteration of EM algorithm for LDA can be described as follow:

- E-step: For each document d, find the values of the optimized parameter of variational distribution as the posterior distribution of latent variables.

- M-step: Maximize the resulting lower bound log likelihood respecting to the parameters α and β with approximate posterior computed in previous step. This two steps are repeated until converges to meet the most raised lower bound of the log likelihood.

2.2.3 Collapsing Gibbs sampling

Another popular approximating method is sampling method, which is provided by Griffiths and Steyvers ,where they attempt to collect samples from the posterior to approximate it with an empirical distribution. In this section, we'll describe Gibbs sampling, one of Markov chain Monte Carlo (MCMC) method to inference posterior distribution of LDA.[4]

Markov chain Monte Carlo (MCMC) methods are a serious of algorithms that make samples from distribution by construct *markov chain* that has the desired distribution. The sampling will start after the chain has run a large number of steps to mix. Gibbs sampling is one of the most simple MCMC algorithm to approximate joint distribution by sample from conditional distributions.

In their method, they integrate the mixtures ϕ and θ and sample the latent variable z. By integration, they obtain the full conditional distribution of variable z:

$$P(z_i = j \mid z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{d_i} + \alpha}{n_{-i}^{(d_i)} + T\alpha} \qquad (1)$$

where $n_{-i,j}^{(w_i)}$ is the number of instances of word w assigned to topic j, $n_{-i,j}^{(\cdot)}$ is the total number of words assigned to topic j, $n_{-i,j}^{d_i}$ is the number of words from document d_i to topic j, and n^{d_i} is the total number of other words in document d_i.

To construct and initial the state of Markov chain, they first initialize the variable z_i to the initial value (between 1 and number of topics T). Then the chain runs a large number of steps and sampling z_i from (1). After a long run of Markov chain, when distribution of the chain approaches its stationary distribution, the sampled values of z_i are started to be collected. Other parameters that are independent of the individual topics could be computed by integrating across the samples.

3 Overcome the correlation

One of main limitations of LDA is the weak ability on topic correlation modeling. In many fields, researchers would be interested in the learning the correlation between different documents with similar topics. For instance, when researchers search the particular article in his working field, the retrieved article may not only relate what he was interested in, but also a set of articles that are highly correlated in topic with the original article he want to search. In this section, a series of topic models that capture the relation between different models is introduced.

3.1 Correlated Topic Models

Correlated Topic Model(CTM)[8] is a novel topic model that extend from LDA which directly model the correlation between topics. It is a more flexible distribution for the topic proportions. This section also introduce the main-field variational inference algorithm applied in CTM computation.

3.1.1 Logistic normal Topic Model

In Correlated topic model(CTM), the topic proportions are drawn from logistic normal distribution, a distribution on the simplex obtained by transforming random variable drawn from a normal distribution. Here is the notation to be describe in the model:

- $w_{d,n}$ denote the nth word observed in the dth document. It is an element in a V-term vocabulary.
- β_k denote the distribution over the V-term vocabulary of topic k, which is a point on the V-1 simplex.
- $z_{d,n}$ denote the topic assignment which is the same we described in LDA. The topic assignment $z_{d,n}$ is associated with the nth word and dth document.
- θ_d is simply the topic proportion for document d, which is point on the K-1 simplex. Different from LDA, in this model we typically consider the natural parameterization of the distribution η

For the model, the CTM assume the document is arisen from following generative process:

1. Drawn a topic distribution θ, considered the natural parameterization mapping $\theta_d = f(\eta) = \frac{exp\{\eta\}}{\sum_i exp\{\eta_i\}}$, where η is drew from multivariate Normal distribution $N(\mu, \Sigma)$
2. For each position in document $n \in \{1, ..., N_d\}$
   (a) Drawn a topic assignment $z_n \mid \eta$ from Mult(θ^d)
   (b) Drawn a word $w_{d,n} \mid \{z_n, \beta_1 : K\}$

The CTM is eventually build based on LDA except drawing topic proportion from logistic normal distribution instead of Dirichlet distribution. From the generative process, we know that CTM draws the natural parameters from a multivariate Gaussian distribution and the map it to simplex to get the multinomial parameter θ. The dependencies between topic portion is caused by the dependencies between the components of the transformed vector induced by the covariance of the Gaussian. Figure 3 shows the plate notation for CTM, where μ is a K dimensional positive vector and Σ is $K \times K$ covariance matrix for multivariate Gaussian

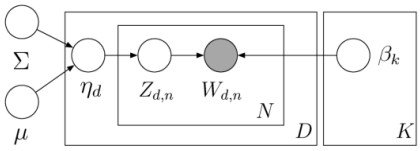

Figure 3: The Plate notation for CTM.

3.1.2 Computation of the Correlated Topic Model

In this section, learning method of CTM will be briefly introduced. Again, similar to variance inference method introduced with LDA, we still have two computational problems to solve: 1. Given a collection of topics and their distribution, the posterior of latent variables such as η and z should be estimate in order to observe documents' latent topic structure. 2. Given a collection of documents, the parameters of topic distribution and logistic normal distribution should be optimized by finding the maximum log likelihood.

For posterior inference, given document w and parameters β, μ and Σ,the posterior distribution of CTM can be described as follows:

$$P(\eta, z \mid w, \beta, \mu, \Sigma) = \frac{P(\eta \mid \mu, \Sigma) \prod_{n=1}^{N} p(z_n \mid \eta) p(w_n \mid z_n, \beta)}{\int p(\eta \mid \mu, \Sigma) \prod_{n=1}^{N} \prod_{z_n=1}^{K} p(z_n \mid \eta) p(w_n \mid z_n, \beta) d\eta}$$

Again, its intractable to computer directly because of the integrating of latent η in denominator. Thus, similar to LDA, the main idea of the inference is to optimize free parameters of variational distribution so that the distribution is close in KL divergence to the posterior of CTM. The variational distribution for the latent variables can be described as follows:

$$q(\eta, z \mid \lambda, \nu^2, \phi) = \prod_{i=1}^{K} q(\eta \mid \lambda, \nu^2) \prod_{n=1}^{K} q(z_n \mid \phi)$$

Where the distributions of the topic assignments z are depends on K-dimensional multinomial parameters ϕ_n, and the variables η are specified by K independent univariate Gaussians $\{\lambda, \nu\}$.

For parameter estimation, they designed a variational EM. In its E-step, they maximize the bound by performing variational inference for each document ,and in M-step, they maximize the likelihood estimation of the topic and multivariate Gaussian using expected sufficient statistics. Those expectation is taken with respect to the variational distributions computed during E-step.

3.1.3 Correlation Graph

The correlation between the topics in CTM can be captured by the covariance of logistic normal distribution. The covariance matrix can be used to form a topic graph where the nodes that is near each other represents the highly related topic. To implement the relation graphic model, the Gausian graphical model needed to be specific.

For this work, they applied the work of Meinshausen and Bülmann[10], which shows how to estimate the graph by *Least absolute shrinkage and selection operator*(lasso).[12]. The general idea of lasso is to regress each random variable X_s onto all other variables in $X \backsim N(\mu, \Sigma)$, by imposing a penalty parameter on the parameters to encourage sparsity. In this case, they treat standardized mean vectors λ_d of variational distribution described before as data and regress each component onto the others with penalty parameter l_1.

3.2 Pachinko Allocation Model

We know that the parameters of CTM in covariance matrix successfully capture the correlation between topics. However, the main drawback of CTM is that it can only capture the pairwise correlation other than the multiple topics. To overcome this, we describe another related topic model modeling the correlation between topics is Pachinko Allocation Model (PAM) [9], which capture the correlation by directed acyclic graph (DAG). They used leaves of DAG to represent individual words in the vocabulary and each node model a correlation among its children.

One interesting point in PAM is that the concept of topic in this model has been extended to the distribution over other topics including words. Thus the model become a hierarchical DAG where each interior node represents a topic, which have a distribution over its children (other topics or words).

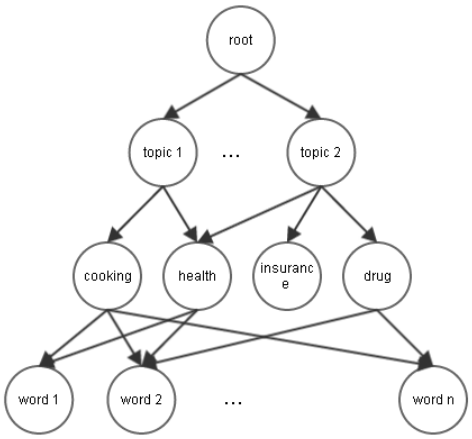

Figure 4: Example graph of PAM

Here in Figure 4 is a example of a DAG describing the hierarchy of topic distribution. It shows four topic: cooking, health, insurance and drugs. The four notes are directly connected to the words. There are many additional topics in a higher lever which is connected to the lower level topics. In this example, topic 1 is related to cooking and health while topic 2 is related to health, insurance and drugs.

3.2.1 PAM model

The model is named from a traditional game - Pachinko, a ball gambling game where metal balls bounce down around a collection of arbitrary pins until they reach the bottom pins. The generation process of the model is similar to this game. In the DAG, each interior node contains a Dirichlet distribution over its children. In the generation process, the root of DAG sample one of its children according to its multinomial distribution, then the child node continue to sample its children until reach the leafs in bottom, the words. Specifically, to generate a document d, the following steps should be followed:

1. Sample topic proportions $< \theta^d_{t_1}, \theta^d_{t_2} ... \theta^d_{t_s} >$ over children from Dirichlet distributions $< g_1(\alpha_1), g_2(\alpha_2), ..., g_s(\alpha_s) >$.
2. For each word w in the document.
   (a) Simple the topic path z_w with length L_w ($L_w - 1$ level) : $z_{w1}, z_{w2}, ..., z_{wL_w}$, where z_{w1} is the toor and z_{w2} through z_{wL_w} are topic nodes in hierarchy, where z_{wi} is always the child of $z_{w(i-1)}$ which is sampled from multinomial distribution $\theta^d_{z_{w(i-1)}}$
   (b) Sample the word from $\theta^d_{z_w L_w}$

From this process, we can write the joint probability of PAM for specific document d is as follows:

$$P(z^d, \theta^d, d \mid \alpha) = \prod_{i=1}^{s} P(\theta^d_{t_i} \mid \alpha_i) \prod_w (\prod_{i=2}^{L_w} P(z_{w_i} \mid \theta^d_{z_{w(i-1)}})) P(w \mid \theta^d_{z_w L_w}) \quad (2)$$

We marginalize z^d and θ^d from (2) for d, then product of the probability for document set D can be described as follows:

$$P(D \mid \alpha) = \prod_d \int \prod_{i=1}^{s} P(\theta^d_{t_i} \mid \alpha_i) \prod_w \sum_{z_w} (\prod_{i=2}^{L_w} P(z_{w_i} \mid \theta^d_{z_{w(i-1)}})) P(w \mid \theta^d_{z_w L_w}) d\theta^d$$

For simplification, similar to LDA, the multinomial distributions for topics in last level are simpled from whole corpus from a single Dirichlet distribution with parameter β. In this view LDA can be considered as a special PAM where DAG is a three-level hierarchy with a root at the top and a set of topics in the middle and a word vocabulary at the bottom.

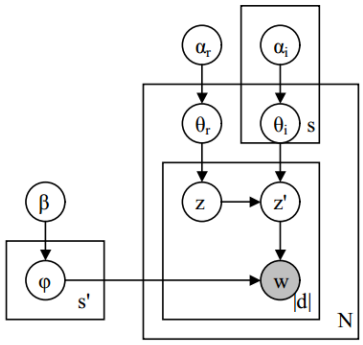

Figure 5: The Plate notation of 4-level PAM

Here is an example of identical 4-level PAM with simplification shown by Figure 5. In this case, we call the topics at the second level super-topics and call third level sub-topics with size s. The multinomial distributions ϕ of sub-topics are sampled from the whole corpus of size s' from $g(\beta)$. The joint probability of generating the whole document set is the probability of every document, which integrate the multinomial distributions for sub-topics.

$$P(D \mid \alpha, \beta) = \int \prod_{j=1}^{s'} P(\phi_{t_j} \mid \beta) \prod P(d \mid \alpha, \phi) d\phi$$

Where the $P(d \mid \alpha, \phi)$ is the marginalized joint probability with super-topics assignment z^d and the sub-topics assignment z':

$$P(d \mid \alpha, \phi) = \int P(\theta^d_r \mid \alpha_r) \prod_{i=1}^{s} P(\theta^d_{t_i} \mid \alpha_i) \prod_w \sum_{z_w, z'_w} P(z_w \mid \theta^d_{z_w}) P(w \mid \phi_{z'_w}) d\theta^d$$

3.2.2 PAM Inference and Parameter Estimation

Actually, PAM is more proper to be learned by Gibbs Sampling since EM algorithm, which is common used perform poorly for its local maxima.[9] The sample path of four-level PAM should from the root through a super-topic and a sub-topic. The root doesn't need to be sampled because the root is fixed for each word. By

marginalizing the latent distribution θ, the proportion of a super-topic and a sub-topic for word w_k in document d can be shown as follows:

$$P(z_w = t_i, z'_w = t_j \mid D, z_{-w}, \alpha, \beta) \propto \frac{n_i^d + \alpha_{ri}}{n_r^d + \sum_{i=1}^{s} \alpha_{ri}} \times \frac{n_{ij}^d + \alpha_{ij}}{n_i^d + \sum_{j=1}^{s'} \alpha_{ij}} \times \frac{n_{jk} + \beta_k}{n_j + \sum_{k=1}^{n} \beta_k}$$

Where z_{-w} is the topic assignments for all other except the super-topic z_w and sub-topic $z_{w'}$. n_r^d is the number of roots in document d. And n_i^d is the number of super-topic t_i in d. Similarly n_j is the total number of sub-topic t'_i in whole corpus and n_{jk} k is the number of word w_k in sub-topic t'_j.

For the given Dirichlet parameters α and β, they has to be learnt in each iteration of Gibbs sampling. There are several empirical study on parameter estimation in Gibbs sampling, in this case, moment matching method [3] is applied. Therefore in each iteration of Gibbs sampling, the parameters can be updated by following rule:

$$mean_{ij} = \frac{1}{(N_i + 1)} \times (\sum_d \frac{n_{ij}^d}{n_i^d} + \frac{1}{s'})$$

$$var_{ij} = \frac{1}{(N_i + 1)} \times (\sum_d (\frac{n_{ij}^d}{n_i^d} - mean_{ij})^2 + (\frac{1}{s'} - mean_{ij})^2)$$

$$m_{ij} = \frac{mean_{ij} \times (1 - mean_{ij})}{var_{ij} - 1}$$

$$\alpha_i j = \frac{mean_{ij}}{exp(\frac{\sum_j log(m_{ij})}{s' - 1})}$$

Notice that the update rule above is applied smoothing to get avoid the situation that some sub-topic will never sampled from specific super-topic, which causes α_{ij} become 0.

4 Overcome the bag-of-words

One of the most important drawback of LDA and its related models is their *bag-of-words* assumption, in which the word order does not take into consideration. The structure of the document is completely broken during leaning. Recently, two novel topic models appears during the deep research of overcome the bag-of-words assumption. In this section, some modern topic models that capture the word order are introduced briefly. One of them is *Bigram topic model* which extend LDA with bigram language model. Another is called *Unsupervised Topic Segmentation* (NTseg), which is a unsupervised topic segmentation approach considering the word order.

4.1 Bigram Topic Model

The basic work to overcome the bag-of-words is construct the models based on the N-gram language model. Here we introduce one most outstanding N-gram models named *Bigram topic model* provided by Hanna.[13].

Bigram language model is the model predict each word based on the the measurement of previous word. The bigram language model she defined can be specified by setting Dirichlet prior $P(\Phi \mid \beta m) = \prod_j Dir(\phi_j \mid \beta m)$ with hyperparameters βm into biterm conditional distribution matrix Φ and do marginalization.

The Bigram topic model extend LDA generation which define the distribution $\phi_{j,k}$ over topic k and context j. She also describe two priors of matrix Φ: $\prod_j \prod_k Dir(\phi_{j,k} \mid \beta m)$, and $\prod_j \prod_k Dir(\phi_{j,k} \mid \beta_k m_k)$, where each topic k have a set of parameter shared only within the topic.

The generative process is a little different from LDA. Instead of drawing ϕ_k, this model draw $\phi_{j,k}$ from one of two priors. After drawing topic assignment z_n, the word is drawn from the words distribution $\phi_{j,k}$ with topic $k = z_n$ and previous word $j = w_{t-1}$.

4.2 Unsupervised Topic Segmentation

The *Unsupervised Topic Segmentation* (NTseg)[7], is another novel topic model that keep the document's structure, such as paragraphs, sentences, and word order within sentences. There are two interesting features of this model. One is considering document generated by several topically coherent segments. The other is the model is able to capture word collocations for its preservation of the ordering of words.

In this model, two levels of topic granularity are captured. One is the segment-level topic, which assigned to the segments in document. The other is word level topic which assigned to n-gram word in segments. In a document, each segment-level topic contains mixture of word-level topics and the mixture uniquely specify the segment level topic, and the word-level topic come from a set of predefined word-level topics.

The basic structure of NTseg can be briefly described as follow. The segments of a document is assumed to follow a Markov model on the topic distributions of each segment. NTseg assume that the probability that the topic for the segments s in the document will be the same as that of the segment $s-1$ will be high. For a atomic segment s, the model finds n-gram words in a word-level topic z. There is a variable for segment-level topic in document d indicating whether the topic between neighboring segment should be changed. Another random variable that NTSeg incorporate is a binary variable which indicate weather a word w_i at position i forms a bigram

with the previous word w_{i-1}.

The generative process of NTSeg is much more complex than other models we described before. Generally, for each word-level topic z, NTseg first generates the unigram word distribution ϕ_z, and for each word-level topic z and each word w it generates Bernoulli bigram status distribution ψ_{zw}, the bigram word distribution for bigrams. Then, for each document d in corpus, it draws the mixing proportion of segment-level topics τ^d, and the Bernoulli segment switch variable distribution π^d. Then, for each segment s in document d, it draws segment switch variable c_s^d from π^d, segment-level topic y_s^d from τ^d, and the mixing proportion θ_s^d from $Dir(\alpha_{y_s^d z})$ where α is a $K \times V$ mixing proportion matrix. Finally, for each of N_s^d words of segment s in document d, draws bigram status variable x_{si}^d between word at position i and $i-1$ from $Bernoulli(\psi_{z_{i-1}w_{i-1}})$, draws word-level topic assignment z_{si}^d from $Multi(\theta_s^d)$ if $x_{si}^d = 0$ otherwise $z_{si}^d = z_{s,i-1}^d$, and draws word from $Multi(\sigma_{z_{si}w_{s,i-1}})$ if $x_{si}^d = 1$ otherwise from $Multi(\phi_{z_{si}})$.

We can see that NTseg generates the word-level topic assignment the previous topic assignment if they form a bigram term, otherwise it will be generated from the mixing proportion of segment-level topics. This feature capture the structure of words within a segment and the words also generate according the bigram status variable.

5 Overcome the sparsity

Recently, to retrieval topics in short text, such as microblogs, short messages and other short text on social website, has become an important research topic and attracted many researchers. The main challenge of collecting theme information from short text is the sparsity of words in short message. Normally we can integrate short texts to learn the topic, but it doesn't help us to compare and classify texts according to their topic distribution. In this section, a novel *Biterm topic model*(BTM), which perform well on corpus with short texts, is introduced.

The *Biterm Topic Model*[14] was explored recently when people tried to overcome the sparsity problem of the LDA extended models for short text corpus. It was found that not only have better performance on short text than the state-of-the-art topic models but also outperform on normal corpus used before.

The most novel feature of Biterm topic model is that it treats a biterm, an unordered word-pair co-occurred in a short context, as a single term and treats a set of arbitrary combinations of two different words in a specific topic as biterms in topic. Here they assume that Biterm topic model is built upon a biterm set instead of documents. The generative process of Biterm topic model can be shown as follows:

1. For each topic z, draw word distribution ϕ_z from $Dir(\beta)$.
2. Draw a topic distribution θ for the whole corpus from $Dir(\alpha)$.
3. For each biterm b in the biterm set B
   (a) draw a topic assignment z from $Multi(\theta)$
   (b) draw the biterm w_i, w_j from $Multi(\phi_z)$

The joint probability of a biterm can be written as follows:

$$P(b) = \sum_z p(z)p(w_i \mid z)P(w_i \mid z)$$

We can see that instead of draw document-level topic distribution as LDA, the Biterm topic model draw topic distribution from the whole collection. Thus when text is short, Biterm topic model doesn't suffer from the sparse problem because it draws topic assignment z from the corpus-level topic distribution θ.

6 Summary

Discovering the theme information from documents is one of most important technique people explored in NLP research. Using probabilistic topic models constructed a explicit framework on solving topic modeling problems. This survey introduce one of the most classical probabilistic topic models - LDA and a series of extend models of LDA to overcome drawbacks of LDA, including the limitation of bag-of-words assumption, the weak of capture the correlation between topics, and sparsity problem in short text corpus.

Eventually, there are many other extend research direction on the extend models of LDA. For instance, we may want to assume that the topic changes according the order of documents. There are one approach on this problem respecting the ordering of the documents and gives a richer posterior topical structure than LDA.[1]. Besides, the number of the topic should not be fixed and predefined. On this problem, the non-parametric topic model-*Hierarchical dirichlet processes*[11] provides the solution that the number of topics is determined by the collection during posterior inference. Extension of the data type is another research direction. The data type adapted by LDA can be extend to multimedia, such as audio, image and video, or others such as user context, program code and social networks.

There are many new directions for research on Topic modeling. One is the evaluation of the models. This relate to how topic models are evaluated and checked. Developing a evaluation method of topic models is still a open problem. For topic checking, the main problem is how to select topic models for a specific corpus and

task. Another open problem is that the topic models should have a better interface - a more useful structure other than a collection of word with different probability. These structures should tell people more potentially useful information about the document.

References

[1] BLEI, D. M., AND LAFFERTY, J. D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 113–120.

[2] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

[3] CASELLA, G., AND BERGER, R. L. *Statistical inference*, vol. 70. Duxbury Press Belmont, CA, 1990.

[4] GRIFFITHS, T. L., AND STEYVERS, M. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America 101*, Suppl 1 (2004), 5228–5235.

[5] HOFMANN, T. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (1999), ACM, pp. 50–57.

[6] HOFMANN, T., PUZICHA, J., AND JORDAN, M. Unsupervised learning from dyadic data. *International Computer Science Institute* (1998), 1–33.

[7] JAMEEL, S., AND LAM, W. An unsupervised topic segmentation model incorporating word order. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (2013), ACM, pp. 203–212.

[8] LAFFERTY, J. D., AND BLEI, D. M. Correlated topic models. In *Advances in neural information processing systems* (2005), pp. 147–154.

[9] LI, W., AND MCCALLUM, A. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 577–584.

[10] MEINSHAUSEN, N., AND BÜHLMANN, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*, 3 (2006), 1436–1462.

[11] TEH, Y. W., JORDAN, M. I., BEAL, M. J., AND BLEI, D. M. Hierarchical dirichlet processes. *Journal of the american statistical association 101*, 476 (2006).

[12] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[13] WALLACH, H. M. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 977–984.

[14] YAN, X., GUO, J., LAN, Y., AND CHENG, X. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (2013), International World Wide Web Conferences Steering Committee, pp. 1445–1456.