A review of topic model on micro-blogging site using Latent Dirichlet Allocation

Thuong-Hai Pham
Faculty of Information and Communication Technology
University of Malta
Msida MSD 2080, Malta
Email: thuong-hai.pham.16@um.edu.mt

***Abstract*—In this paper, the author reviews approaches to apply Latent Dirichlet Allocation (LDA) as a unsupervised topic model for micro-blogging sites. For practical evaluation, Twitter is chosen to do experiment on three approaches: standard LDA, author-topic model and Twitter-LDA. The paper also covers background behind each method and proposes a future works based on these approaches.**

1. Introduction

To deal with the explosion of data these days, especially electronic documents and text generated by world-wide-web users demands techniques that automatically organized large collection of text. One family of those techniques is called "topic model". These techniques discover underlying topic from a given corpus with or without intervention of human, in other word, supervised and unsupervised. One very successful technique and also being considered as state-of-the-art in unsupervised topic model is Latent Dirichlet Allocation (LDA) [1].

Applying topic model for micro-blogging site is a very important task to enhance our understanding of the social network. Some applications were proposed by Zhao et al. (2011) [2] with the work of comparing Twitter and traditional media by LDA, or finding topic-sensitive influencers on Twitter by Weng et al. (2010) [3]. It is important to be noted that applying LDA on micro-blogging sites is considered a more challenging problem due to the constraint of text length (140 characters in case of Twitter).

2. Latent Dirichlet Allocation

2.1. Bag-of-words assumption

2.1.1. Exchangeability. We say that $(x_1, x_2...)$ is an infinitely exchangeable sequence of random variables if, for any n, the joint probability $p(x_1, x_2, ..., x_n)$ is invariant to permutation of the indices. That is, for any permutation π,

$$p(x_1, ..., x_n) = p(x_\pi(1), ..., x_\pi(n))$$

It is important to emphasize that independent and identically distributed random variables are always infinitely exchangeable. However, infinite exchangeability is a much broader concept than being independent and identically distributed. For example, let $(x_1, x_2, \dots)$ be independent and identically distributed, and let x_0 be a non-trivial random variable independent of the rest. Then $(x_0 + x_1, x_0 + x_2, \dots)$ is infinitely exchangeable but not independent and identically distributed.

2.1.2. De Finetti theorem, 1935. A sequence of random variables $(x_1, x_2, ...)$ is infinitely exchangeable iff,for all n,

$$p(x_1, x_2, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta)$$

for some measure P on θ. If one assumes the data is infinitely exchangeable, then there must exist an underlying parameter and prior.

2.2. Latent Dirichlet Allocation

LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words. To implement this idea, LDA assumes each document is a bag-of-words. Hence, it applies infinite exchangeability on the documents and inherits the De Finetti theorem to expect an latent parameter and prior underlying in the corpus. These latent variables are illustrated in the figure 1 below.

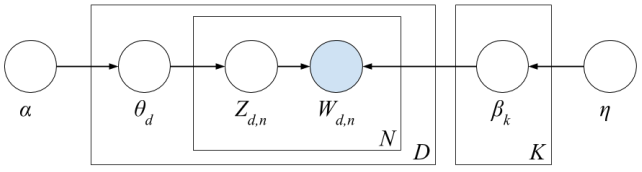

Figure 1. LDA graphical model

In figure 1:

- α is Dirichlet distribution parameter, controls the shape and sparsity of θ
- θ are per-document topic proportions.
  θ is a K-dimensional Dirichlet random variable,

takes values in the (k-1)-simplex, and has the following probabiltiy density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\dots\theta_K^{\alpha_K-1}$$

The Dirichlet is conjugate to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.

- $Z_{d,n}$ is per-word topic assignment, in which D and N are number of documents and number of words in a specific document, respectively.
- $W_{d,n}$ is observed word.
- β are topics, which is V dimensional Dirichlet.
- η is the topic hyper parameter.

The blue-shaded node denotes observed variable, the others are hidden or latent variables. Plates denote replicated structures.

From a collection of documents, LDA infers:

- Per-word topic assignment $Z_{d,n}$
- Per-document topic proportions θ_d
- Per-corpus topic distributions β_k

2.2.1. Generative process. As mentioned above, LDA is a generative probabilistic model, which generative process is performed as described below:

1) Draw $\theta_d \sim Dir(\alpha)$
2) Draw $\beta_k \sim Dir(\eta)$
3) For each of the N words in document d $W_{d,n}$:
   a) Draw a topic $Z_{d,n} \sim Multinomial(\theta_d)$
   b) Draw a word $W_{d,n}$ from $p(W_{d,n}|Z_{d,n},\beta)$, a multinomial probability conditioned on the topic $Z_{d,n}$

2.2.2. Model inference. However, in the real problem, to acquire underlying latent topics, we have to reverse the generative process by solving an inferential problem. The main goal of this inferential problem is to compute the posterior distribution of the latent variables in figure 1:

$$p(\theta, Z|W, \alpha, \beta) = \frac{p(\theta, Z, W|\alpha, \beta)}{p(W|\alpha, \beta)}$$

The function $p(\theta, Z|W, \alpha, \beta)$, in practice, is not possible to compute. Due to the conjugacy of Dirichlet distribution, we can marginalize over latent variables to rewrite the posterior $p(W|\alpha, \beta)$. This posterior is still hardly be inferred exactly. Nevertheless, there exist a wide variety of approximate inference algorithms for LDA:

- Mean field variational methods [4] (Blei et al., 2001)
- Expectation propagation [5] (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling [6] (Griffiths and Steyvers, 2004)
- Collapsed variational inference [7] (Teh et al., 2006)

After being approximated, beside LDA, the posterior can be used in many other applications such as collaborative filtering, document similarity and information retrieval...

3. Latent Dirichlet Allocation approaches for Twitter

One very basic approach is to apply LDA directly to Twitter by treating each tweet as a single document. However, due to the constraint of 140 characters per tweet, a tweet is too short for LDA to figure out the topic proportion.

3.1. The author-topic model

To overcome the problem, by assuming that each users on Twitter (Twitterer) only has a fixed interest, aggregating all tweets of a Twitter into a single documents was proposed and gained a better result to direct LDA [3], [8].

This approach is very efficient on a specific task (ex. topic-sensitive influencers mining [3]) by not modifying the inference process of original LDA. However, the assumption of the approach is also its disadvantages by the fact that tweets of a Twitterer may generated/tweeted based on a variety of topic that Twitterer pay attention on.

3.2. Twitter-LDA

In an attempt to compare Twitter and traditional media, Zhao et al. [2] proposed Twitter-LDA, a modified version of LDA to work on Twitter's short tweets without assuming that Twitterer has one fixed interest.

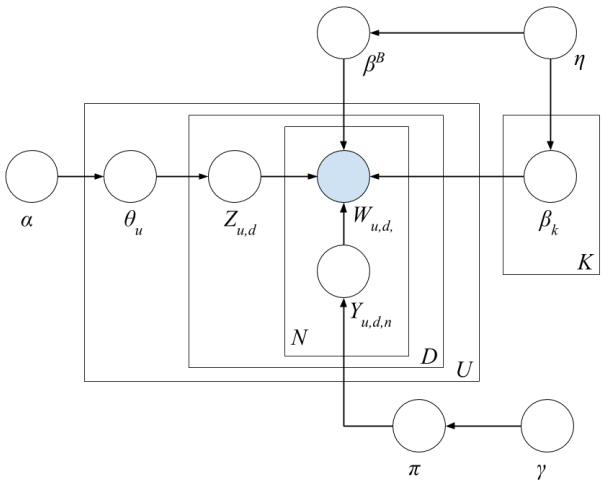

Figure 2. Twitter-LDA graphical model

In figure 2, the author introduced four more variables:

- β^B denotes the background words distribution
- π denotes a Bernoulli distribution which simulate the choice of authors between topic-related words and background words.
- γ is the parameter of distribution π.
- $Y_{u,d,v}$ denotes the selection of background or topic word.

and a slightly modification on θ that θ_u denote per-user topic proportions instead of per-document as in the original version. In addition, the D (document) plate is surrounded by a new plate U which stands for each user.

By defining the model as in figure 2, the generative process of Twitter-LDA is performed as followed:

1) Draw $\beta^B \sim Dir(\eta)$
2) Draw $\pi \sim Dir(\gamma)$
3) Draw $\beta_k \sim Dir(\eta)$
4) For each user,
   a) Draw $Z_{u,d} \sim Multi(\theta_u)$
   b) For each word in document d,
      i) Draw $Y_{u,d,n} \sim Multi(\pi)$
      ii) Draw

$$W_{u,d,n} \sim \begin{cases} Multi(\theta^B) & \text{if } Y_{u,d,n} = 0 \\ Multi(\theta^{Z_{u,d}}) & \text{if } Y_{u,d,n} = 1 \end{cases}$$

4. Evaluation

4.1. Evaluation methods

Wallach et al. [9] summarized a variety of methods to evaluate LDA. As a topic model method, LDA is commonly evaluated by intrinsic and extrinsic evaluation.

4.1.1. Intrinsic method. One very basic intrinsic evaluation method is to view the problem as document modeling [1]. By stating that, the goal of the model is t achieve high likelihood on a held-out test set, C'. In this case, the perplexity measure is used as in normal language modeling problem, in which the lower perplexity, the better performance the model achieves.

$$perp(C') = exp\left\{-\frac{\sum_{d=1}^{D} log(p(W_d))}{\sum_{d=1}^{D} N_d}\right\}$$

More advanced, measurement is estimated by the probability of unseen held-out documents given some training documents. This probability can be written as [9]

$$P(C|C') = \int d\theta d\alpha dm P(C|\theta, \alpha m) P(\theta, \alpha m|C').$$

in which, C, C' denote training documents set (corpus) and held-out documents set, respectively. Noted that m is the base measure of Dirichlet distribution, in addition to the concentration parameter α.

There is also a variation of this method, document completion, which compare predictive performance by estimating the probability of the second half of each document given the first half. In this point of view, let $c^{(1)}$ be the first half and $c^{(2)}$ be the second half, the goal of our measurement is to compute

$$P(w^{(2)}|w^{(1)}, \theta, \alpha m) = \frac{P(w^{(2)}, w^{(1)}|\theta, \alpha m)}{P(w^{(1)}|\theta, \alpha m)}$$

4.1.2. Extrinsic method. On the other hand, extrinsic approaches measure LDA performance on some secondary tasks, such as corpus comparison [2] or topic-sensitive influencers mining [3]. These approaches are similar to the way language models performance are measured.

4.2. LDA approaches evaluation result

As a part of the corpus comparing work [2], Zhao et al. also evaluated the performance between original LDA, author-topc model and their proposed Twitter-LDA. Based on preliminary experiments, the authors set number of topic K to 110 for each model, then mixed 330 topics from the three models. The topics were then scored by two human judges. Each assigned a score on each topic, ranging from 1 (meaningful) to 0 (nonsense).

The result showed that Twitter-LDA gained 25.23% higher in term of average score to author-topic model, and 32.61% higher then standard LDA. Hence, Twitter-LDA obviously outperformed the two previous methods and were used in their comparison task.

5. Conclusion

The two proposed methods of author-topic model and Twitter-LDA have overcome the problem of short tweets in the micro-blogging site Twitter. More than that, Twitter-LDA has showed to outperform the other two in experiment.

However, Twitter-LDA has to change the original LDA process and inference approximation algorithm to implement its idea. This approach is hard to be re-implemented in industrial by using existed library for other problems. We argue that aggregating into a single document may fail to model topic(s) of influencers who have more than one field of interest and/or change their interest from time to time. Hence, the two following scenarios should be the future work to solve the problem without modifying the original LDA process:

1) Applying LDA on a single user, each tweet as a document, then group these documents based on their LDA-topics.
2) Grouping tweets based on their tweeted/retweeted timestamp.

After processing the corpus by one of these above mentioned methods, we treat the grouped tweets as a single document to apply LDA.

References

[1] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European Conference on Information Retrieval*, Springer. Springer Berlin Heidelberg, 2011, pp. 338–349. [Online]. Available: http://ink.library.smu.edu.sg/sis_research http://ink.library.smu.edu.sg/sis_research/1375 http://link.springer.com/10.1007/978-3-642-20161-5_34

[3] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank," in *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, 2010, p. 261. [Online]. Available: http://ink.library.smu.edu.sg/sis_research http://portal.acm.org/citation.cfm?doid=1718487.1718520 http://ink.library.smu.edu.sg/sis{_}research

[4] D. M. Blei and M. I. Jordan, "Variational methods for the dirichlet process," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 12.

[5] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 352–359.

[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[7] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *Advances in neural information processing systems*, 2006, pp. 1353–1360.

[8] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*, ACM. New York, New York, USA: ACM Press, 2010, pp. 80–88. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1964858.1964870

[9] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. New York, New York, USA: ACM Press, 2009, pp. 1–8. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1553374.1553515