Paper Review: "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*—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. This report is aimed to give a review of Blei et al. work in these techniques family, Latent Dirichlet Allocation, its suitability to be published in the Journal of Machine Learning Research 3, 2003, and to propose a technical improvement that should overcome the problem remained.**

1. Suitability of the topic

1.1. Topic appealing to the Journal readers

The paper was submitted to be published in the Journal of Machine Learning Research 2003. By eliminating the assumptions in tf-idf, Latent Semantic Index (LSI) and probability Latent Semantic Index (pLSI), Latent Dirichlet Allocation (LDA, Blei et al. [1]) makes use of De Finite theorem as the base idea and works as an unsupervised method to extract underlying topics.

Although the topic seems to be more relevant to Natural Language Processing (NLP), the proposed method, in conjunction with the previous methods in the family such as LSI, plays an important role in Information Retrieval (IR) when applying to a general collection of data, not just text documents. Hence, the topic is appropriate and useful for the readers of the Journal of Machine Learning Research as well.

1.2. Impact in the field

In the reviewer opinion, the topic of Blei's paper will provide an efficient tool for researchers in Information Retrieval, in general, and Natural Language Processing to extract information from a large collection of documents. As mentioned above, despite the fact that the paper uses text documents to illustrate the idea, the proposed method is appropriate to apply in a more general context of Information Retrieval. Thus, most of the works which are currently depending on the algorithm family (LSI, pLSI, LSA) can be re-evaluated with the new method. More important, after being approximated, the posterior in LDA can be used in many other applications such as collaborative filtering, document similarity, etc.

2. Content

The paper involves in discussing the topic by both examining deeply into the proposed method and covering a wide range of related mathematical concepts and theorems backing up the method.

2.1. Coverage of the topic

The proposed method, which is also the main topic that the authors focus on in this paper, is covered sufficiently in the 3^{rd} section (Latent Dirichlet Allocation). To prepare for the readers adequate background to understand that, the authors also mention the mathematical concepts of De Finite theorem in the same section.

2.2. Technical depth

By discussing inference and parameters estimation in detail, the authors have not only proposed a method which is solely a combination or improvement of the previous works but have proposed a new idea of a working mixture models and how to solve the latent variables inference problem. Even though the inference method used in the paper is not the authors work on their own, the authors have applied it wisely in the context of the work. Moreover, the authors have proved the effectiveness of the proposed method on various tasks with carefully chosen measurements. These tasks vary from topic model to text classification, which is proof of the statement the authors mention as its wide application in IR.

All in all, the paper suits a wide range of readers from nonspecialist (e.g. linguists without computational background) to expert in the field (to improve the inference methods and to modify the LDA variables model).

2.3. Technical novelty

The proposed method is not an improvement of any previous algorithm in the family, which kind of works are very common in the field of Machine Learning Research. In contrast, the work is purely a new idea of mixture model based on observation of exchangeability and bag-of-word assumption. By having that, the method works well on both word mixture and more complicated structures such as documents or paragraphs. In addition to the method itself, the author emphasize the novelty of the proposed method by comparing it with the other related probabilistic models in Section 4. The experiment result reported in Section 7 also clearly prove the effectiveness and innovation of the method.

2.4. Authoritative and originality

In general, the paper can be considered to be authoritative and provide a content with high quality and originality. The paper clearly distinguishes the authors' contribution (the model, observation...) and applying colleagues' works (the inference methods).

3. Presentation

3.1. Overall organization

The paper is well organized with 6 separated main parts: introduction of the previous methods, notation and terminology used in the paper, the proposed method, relationship with other probabilistic models, inference and parameter estimation (which is crucial for this kind of Bayesian model), and empirical results. Besides that, the length of the paper which is 30 pages fits the article length constraint of the Journal of Machine Learning Research.

3.2. Title and abstract

The title is short with only the name of the method. However, the abstract provides an ample amount of information to support the title, by first describing the method itself as "a generative probabilistic model for collections of discrete data" [1] which clearly states all the characteristics of the method and the domain that it can be applied to. In this part, the author also mentions the important method to infer this type of Bayesian model and briefly introduces how experiment is constructed with the counterpart method. In general, the abstract is adequately written to support the short title, which together provide a clear, accurate indication of the material discussed throughout the paper.

3.3. Symbols, terms, and concepts' definition

Not only adequately define all the concepts and terms used in the paper, the authors separate these into a single section after introduction. In which, they also formally define the relationship between these terms and concepts.

3.4. English usage

The authors have used proper and comprehensive academic English to convey the necessary information.

3.5. Bibliography

In the bibliography, the authors have cited all the related works mentioned throughout the article. These works are ranging from not only previous methods (LSI, td-idf, pLSI) but also mathematical concepts and proofs (De Finite theorem, exchangeability) which references are complete and accurate.

4. Overall comments & recommendations

Having analyzed in detail the work of Blei et al. as stated in the sections above, it is obviously that the paper uses an excellent literary style and has high quality and originality. Importantly, nonspecialist from other fields can easily access mostly the knowledge conveyed by the paper with high tutorial value. By claiming that value does not reject the impact of the paper for specialists in Information Retrieval and Natural Language Processing due to the fact that LDA provides a mixture model and very efficient tool which is easily customized for improvement and adapt other contexts.

In conclusion, this paper is excellent and should be published in the Journal of Machine Learning Research.

5. Technical improvement

The proposed method has been showed to outperform the other approaches in the family. However, we argue that the method may fail to model topic(s) of short-text documents, such as Twitter's tweets, Q&A questions or other social media generated contents, due to the constraint of text length (140 characters in case of Twitter). In addition, these are becoming huge sources of data to be analyzed and applying topic model for micro-blogging site is a very important task to enhance our understanding of the social network. Hence, the two following scenarios should be the future works 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.

The above approaches take Twitter's tweets as the example to do the experiment. 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.