DOI: http://dx.doi.org/10.26483/ijarcs.v9i3.6107

ISSN No. 0976-5697

Volume 9, No. 3, May-June 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PROBABILISTIC TOPIC MODELING AND ITS VARIANTS – A SURVEY

Padmaja CH V R
Computer Science & Engineering
Raghu Engineering College
Visakhapatnam, AP, India

S Lakshmi Narayana
Former Principal Scientist
National Institute of Oceanography
Visakhapatnam, AP, India

Divakar CH
Professor, Dept. of Information Technology
SRKR Engineering College
Bhimavaram, AP, India

Abstract **–** Topic modeling is one of the fast-growing research areas as there is a huge increase in internet users. These users are the main source of large volumes of electronic data in terms of documents, tweets, or messages and so on. Collecting, organizing, storing and retrieving the data in text format is becoming more and more typical. The topic model is one research area which focuses on classifying the textual data into groups. In this study, we are presenting a survey on the advanced algorithms that are used in topic modeling. The main purpose of this survey is to provide a brief overview of the current topic models that motivate the budding researchers to select the best suitable algorithm for their work.

Keywords: Topic modeling, pLSA, pLSI, LDA, Dynamical Topic Model.

I. INTRODUCTION

As there is a huge increase in digital data in terms of blogs, news, social networks, web pages, research papers etc., we need sophisticated techniques to extract meaningful information. Traditional search engines would search for terms and retrieve relevant documents whereas topic modeling assumes that each document is a mixture of some topics. Topic modeling is a statistical method for discovering the topics at an abstract level in a collection of documents. In other words, a document can be viewed as a mixture of hidden topics with different proportions. Due to the huge increase in textual data, topic modeling is gaining the attention of researchers and having more number of applications. Even though there are many applications of topic modeling like image processing, analysis of video etc., it was mainly invented for finding topics in textual data.

There are many surveys on topic modeling that already exists. Among them [1], [2], [3] and [4] are most significant. Classification of directed probabilistic topic models and a comprehensive view on graphical models is explained in [1]. It can aid as an initial point for the framework in the field of topic modeling. A reasonable overview is given by [2] and [3]. In paper [4] discusses the classification of probabilistic topic modeling algorithms.

In this paper, we focus on LDA and its variants in topic modeling and how they can be implemented using various tools that are available free. We also present some of the applications of topic modeling.

II. CLASSIFICATION

Topic modeling is gaining more importance in classifying the unstructured textual data. It assumes each document as a group of topics with different proportions. It uses a strong mathematical framework to study each document and discovers the probabilities of topics in documents. Based on three principles the topic models can be classified. The first one is ordering of words. In this, we have two alternatives one is bag-of-words, in which no ordering of words is required and the other is a sequence of words in which ordering of words will be maintained. In the second criteria, some external domain knowledge would be considered for classification. The third criteria are depending on labeled data.

In general, the topic models are unsupervised. But, to improve the results, they can be used as a semi-supervised and supervised approach. But, here we are focused on topic models like Latent Semantic Analysis (LSA), Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing, Latent Dirichlet Allocation and it's few variants.

III. METHODS IN TOPIC MODELING

In this section, a brief explanation is given on topic modeling methods that deals with words, documents and topics.

A. *Latent Semantic Analysis*

In this section, a brief explanation is given on topic modeling methods that deals with words, documents and

topics. Latent Semantic Analysis is a topic modeling method in Natural Language Processing (NLP). The main goal of this method is to create a vector-based representation for texts that help in gathering the more related words. Previously, this method is known as Latent Semantic Indexing(LSI) and now it is refined for text mining.

LSA uses Singular Value Decomposition (SVD) to rearrange the data. SVD uses a matrix to reconfigure and calculate the dimensions of vector space. The dimensions in vector space will be calculated and arranged from most to least important. Evangelopoulos & Visinescu (2012) [5] define three major steps for LSA represented in Figure 1.

In the pre-processing part, the documents should exclude trivial words as well as low-frequency terms and conflate terms with techniques like stemming or lemmatization.

1) A term-frequency matrix (A) must be created that includes the occurrences of each term in each document.

2) Calculate Singular Value Decomposition (SVD): Extract least-square principal components for two sets of variables i.e., for a set of terms and set of documents. These SVD products include the term eigenvectors U, the document eigenvectors V, and the diagonal matrix of singular values Σ.

3) From these, factor loadings can be produced for terms UΣ and documents VΣ

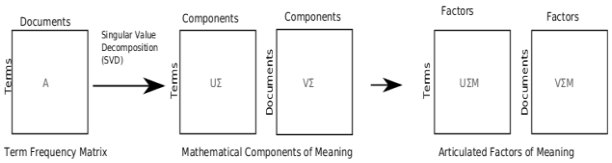

Figure 1. Steps in Latent Semantic Analysis. The figure is redrawn from Evangelopoulos & Visinescu (2012)

The major applications of LSA method are online customer support, relationship discovery, spam filtering, text summarization and so on. The limitations of LSA are high computational and memory, difficulty in determining the optimal number of dimensions to use for performing SVD. The software that is free to implement LSA includes Gensim – Topic Modeling for Humans, LSA package for R, S-Space Package, Semantic Vectors etc.,

B. Probabilistic Latent Semantic Indexing

To fix some disadvantages LSA [6], Probabilistic Latent Semantic Indexing (PLSI) was introduced by T. Hofmann in the year 1999. It includes two vital interpretations. They are allowing words with multiple meanings and revealing typical similarities by grouping together words that shared a common context [7]. According to Hofmann, PLSI is described as a generative process that probabilistically generates documents given the parameters of the model learned by Bayesian inference. The matrix factorization of PLSI is illustrated in Figure 2.

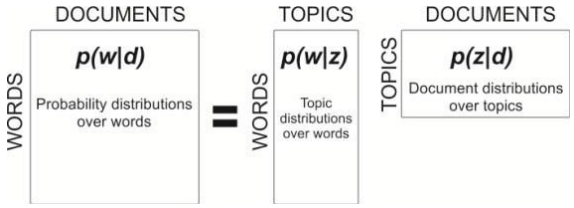

Figure 2. PLSI model

Expectation Maximization algorithm is widely used to compute word-topic and topic-document distributions. The inference of E step and M step from the generative model is shown in Figure 3. The major applications of PLSI model are computer vision and recommender systems. But, it suffers from overfitting problem as the growth of a number of parameters is proportional to the number of documents.

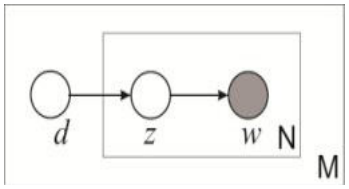

Figure 3. pLSI model

C. Latent Dirichlet Allocation

There is tremendous growth in e-documents in terms blogs, emails, twitter, research papers etc. To handle this huge collection new approaches are needed to organise them well. The probabilistic topic models LSA and PLSA can be used to organize the textual data. As an improvement of LSA and PLSA, Latent Dirichlet Allocation (LDA) [8] is introduced.

LDA is a generative probabilistic topic modeling based on statistical Bayesian topic models. It is a very widely used algorithm in text mining. It tries to identify the theme of the documents in term of topics. With this ability, LDA played a prominent role in text mining research. Many enhancements are made to LDA and applied to various types of data which include dynamic topic modeling, correlated topic modeling, author-topic analysis etc.

The complexity of LDA method is more when compared to PLSA and hence the exact inference is intractable from the generative model. To overcome this, many approximate inference algorithms are derived, which are addressed in [9] The generative process of LDA in plate notation is depicted in figure 4.

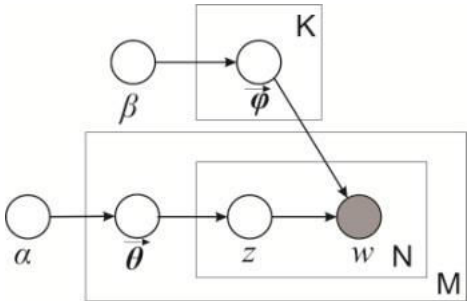

Figure 4. LDA model

The major advantage of LDA on LSA and PLSA is its ability in dimensionality reduction. LDA can be included in more complex methods, which is not possible for LSA. It can be applied to different data, leading to numerous expansions. But still, there are some limitations to LDA method. They are the number of documents, length of the documents, a very large number of topics than needed are used to fit the LDA, well-separated underlying topics in the sense of Euclidean distance, the Dirichlet parameter of the document-topic distributions should be set small.

D. Latent Dirichlet Allocation Variants

D.1. Hierarchical Latent Dirichlet Allocation

In LDA, the topics are represented as a plain structure and unable to find the relationship among the topics. To address this Blei et. al introduces Hierarchical LDA [10] which brings the relationship among topics by representing them in a tree structure. It organizes the topics in a hierarchical way, in which the more abstract topics are near the root.

HLDA generates a model of multiple-topic documents. It creates a mixture distribution on topics using a Nested Chinese Restaurant Process prior and the topics are joined together in a hierarchy by using the nested CRP. It picks a topic according to their distribution and generates words according to the word distribution for the topic. It's working principle in plate notation is explained in figure 5.

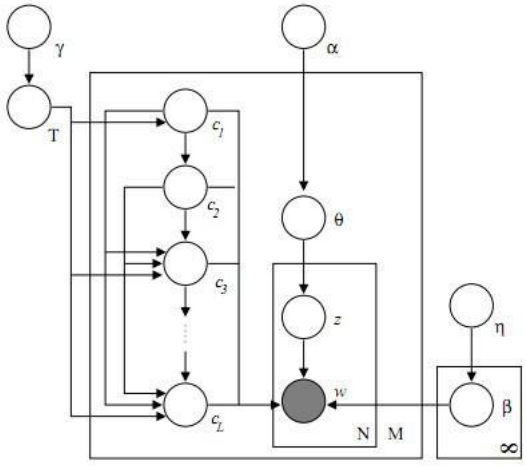

Figure 5. Distributions in Hierarchical LDA

The qualitative results for topic hierarchies in Hierarchical LDA are good but, the quantitative evaluation is not enough. The major limitation of this algorithm is the restriction that the documents can follow a single path in the tree.

D.2. Dynamical Topic Model

As an extension of LA, Dynamical Topic Model (DTM) [11] is introduced by D. Blei and J. Laferty in 2006. In LDA all documents are treated as a bag of words and assumed that they are drawn exchangeable from the same set of topics. Dynamic Topic Model relaxes these assumptions of LDA.

It allows a heterogeneous evaluation of both topic distribution for documents and word distribution for different topics illustrated through in Figure 6. This leads to distinguish the topics that are growing fast and stable over a period.

Using document metadata, it includes the notion of time in topic modeling that can describe the evaluation of word-topic distributions.

The major advantage of DTM is its competence to recognize topics over a period, which was not addressed in previous probabilistic topic models. But, a fixed number of topics and the discrete notion of time are the main limitations of this model. With the increase in time granularity, the complexity of variational inference for Dynamic Topic Model increases. This may encounter unsuitable results due to memory and computational requirements.

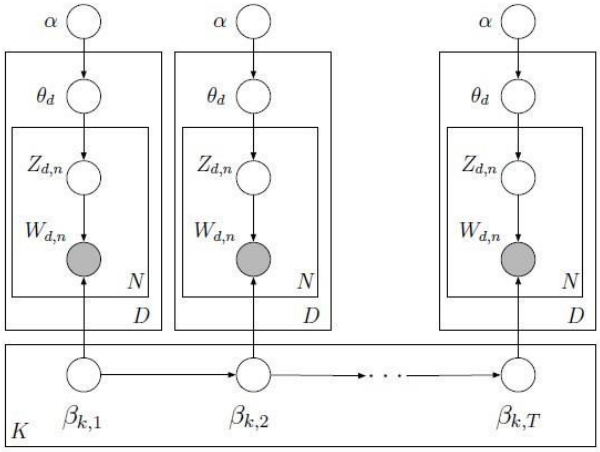

Figure 6. The plate notation of DTM

D.3. Correlated Topic Model

Correlated Topic Model (CTM) is a probabilistic topic model that addresses LDA's inability to model topic correlation was introduced by Blei and Lafferty [13] shown in Figure 7. It is a fast-variational inference model to produce a near estimated inference.

Using hierarchical model, the documents are represented, and word is modeled from mixture model in CTM. These documents share mixture components and their proportions are treated as random variables that are specific to document. In CTM, every document can be viewed as a combination of multiple topics with different proportions.

CTM uses logistic normal distribution to model the latent composition of topics associated with each document. CTM gives a more expressive document model but, it does not fit well to the multinomial and complicates approximation of posterior inference. This is due to the usage of logistic normal distribution.

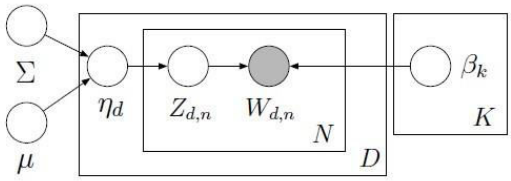

Figure 7. The Graphical model of CTM

D.3. Pachinko Allocation Model

As an alternative to Correlation Topic Model, Wei Li and McCallum [14] proposed a new model Pachinko Allocation model. This method uses Directed Acyclic Graph (DAG) to capture the correlations among topics. In this DAG, the leaves represent words and interior nodes represent the correlation among its child nodes. The generative model of PAM is represented in Figure 8.

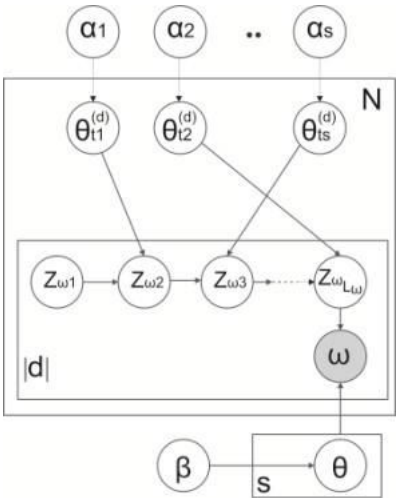

Figure 8. PAM Plate notation

The major drawback of LDA is its inability in modeling correlated data as it uses single Dirichlet distribution to sample the topic proportions in each document. PAM model uses DAG representation to generate a document, in which each interior node is a having Dirichlet distribution over its children. Using PAM, more complicated correlated data can be expressed efficiently.

D.5. Author Topic Model

The Author Topic Model (ATM) [15] is a generative model for authors and documents that reduce the generation of documents to a simple series of probabilistic steps. Each author is associated with a mixture of topics, where topics are multinomial distributions over words. The words in a collaborative paper are assumed to be the result of a mixture of the authors' topics mixtures.

ATM is a consequent model from LDA to find topic distributions corresponding to each author based on metadata. Along with modeling of document-topic and topic-word distributions, it addresses author-topic distributions with the use of Markov Chain Carlo algorithm. Figure 9 is a representation of ATM plate notation.

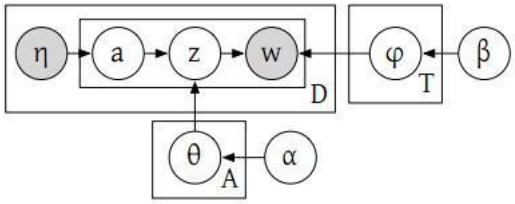

Figure 9. ATM Plate notation

IV. CONCLUSION

In NLP, extracting the theme of textual data is a very important task. Topic modeling is one such research area which addresses this problem. Probabilistic topic models constructed an explicit framework to solve topic modeling problems. In this study, we presented some classical probabilistic topic models like LSA, PLSA, and LDA along with their limitations. We also discussed LDA variants and their performance.

V. REFERENCES

[1] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Knowledge discovery through directed probabilistic topic models: a survey," Frontiers of Computer Science in China, vol. 4, no. 2, pp. 280–301, Jun. 2010.

[2] David M. Blei. Introduction to Probabilistic Topic Models. Communications of the ACM, 2011

[3] Steyvers, M. and Griffiths, T., Probabilistic Topic Models. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), A handbook of Latent Semantic Analysis. Hillsdale, NJ: Erlbaum, 2007

[4] Jelisavcic, V., Furlan, B., Protic, J., & Milutinovic, V. M., "Topic Models and Advanced Algorithms for Profiling of Knowledge in Scientific Papers", 35th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'2012), 1030–1035.

[5] Evangelopoulos, N., Zhang, X., and Prybutok, V. Latent semantic analysis: Five methodological recommendations. European Journal of Information Systems 21, 1 (Jan. 2012), 70–86, 2012.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, vol. 41, pp. 391–407, 1990.

[7] Hofmann, T., Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd ACM SIGIR Conference on Research & Development on Information Retrieval, Berkeley, CA, USA, 1999.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, Jan. 2003.

[9] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. Suppl 1, pp. 5228–5235, Apr. 2004.

[10] D. Blei, T. Gri, M. Jordan, and J. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," 2003.

[11] D. M. Blei and J. D. Lafferty, "Dynamic Topic models," in Proceedings of the 23rd international conference on

Machine learning, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 113–120.

[12] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 424–433.

[13] David M. Blei, John D. Lafferty, "A Correlated Topic model of Science", Annals of Applied Statistics 2001, Vol. 1, No. 1, 17 -35, 2007.

[14] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in Proceedings of the 23rd international conference on Machine learning, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 577–58.

[15] M. R. Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," ACM Trans. Inf. Syst., vol. 28, no. 1, pp. 1–38, Jan. 2010.