Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 69-73 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

TPMTM: Topic Modeling over Papers' Abstract

Than Than Wai[*], *Sint Sint Aung*

Web Mining, University of Computer Studies, Mandalay, ZIP Code 05071, Myanmar

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 05 February, 2018
Online: 08 March, 2018

Keywords:
Frequent pattern mining
Latent Dirichlet Allocation
Entropy

ABSTRACT

Probabilities topic models are active research area in text mining, machine learning, information retrieval, etc. Most of the current statistical topic modeling methods, such as Probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). They are used to build models from unstructured text and produce a term-based representation to describe a topic by choosing single words from multinomial word distribution. There are two main weaknesses. First, popular or common words are different topics, often causing ambiguity for understanding the topics; Second, lack of consistent semantics for single words to be represented correctly. To address these problems, this paper proposes a model (A Two-Phase Method for Constructing Topic Model, TPMTM) that combines statistical modeling (LDA) with frequent pattern mining and produces better presentations of rich topics and semantics. Empirical evaluation shows that the results of the proposed model are better than LDA.

1. Introduction

Topic models are Bayesian statistical models that are structured in accordance with a hidden theme, usually called unstructured data in a set of textual documents, topics with multiple distributions of words. Due to a collection of unstructured text documents, the topic model assumes that the collection of documents (corpus) has a certain number of hidden topics and that each document contains more than one topic in different sizes. Researchers have developed several topic models such as Latent Semantic Indexing (LSA) [1], Probability Latent Semantic Analysis (PLSA) [2] and Latent Dirichlet Allocation (LDA) [3]. Topic modeling automatically selects topics from the text and identifies topics over time [4], explores the connection between topics[5], supervised the topics [6], recommendation [7], and so on.

LDA or unsupervised generation probabilistic methods for modeling the document collection (corpus), is the most commonly used topic modeling method. The LDA, each document can be described as a probabilistic distribution for latent topics, and that the topic distribution of all documents is distributed a common Dirichlet prior. Within each topic in the LDA model is described as a probabilistic distribution over-represented as a probabilistic distribution of words and words distributions of topics distribute the same Dirichlet prior. Each latent topic in the LDA model is also distributions of topics distribute a common Dirichlet prior as well. A corpus D consists of M documents, with document d having N_d words ($d \in \{1,..., M\}$), LDA models D as stated in the following process[8].

LDA concludes the following generative process for each document w in a D corpus:

1. Choose N ~ Poisson (ξ).

2. Choose θ ~Dir (α).

3. For each of the N words w_n:

(a) Choose a topic z_n~ Multinomial (θ).

(b) Choose a word w_n from $p(w_n|z_n.\beta)$, a multinomial probability conditioned on the topic Z_n.

The discovered variables are words in documents although others are hidden variables (θ and ϕ) and hyperparameters (β and α).To provide hyperparameters and hidden variables, the probability of discovered data D and maximized as follows:

$$p\,(D|\alpha,\beta)= \prod_{d=1}^{M} \int p(\theta_d\,|\alpha)\Big(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(\omega_{dn}|z_{dn},\beta)\Big)d\theta_d \quad (1)$$

The LDA model has three levels of the representation. In corpus level, there are two parameters (α and β) that are involved

[*]Corresponding Author: Than Than Wai, Web mining, University of Computer Studies, Mandalay, ZIP Code 05071, Myanmar | Email: thanwai85@gmail.com

in the process of building a corpus. Document-level variables are the variables θ_d, which are sampled once per document. Finally, word-level variables $z_{d,n}$ and $w_{d,n}$ are the variables that are collected for each word in each document. The current statistical topic modeling techniques make multinomial distributions in words to represent topics in a given collection of texts. For example, Table 1 displays a sample of multinomial distributions used to describe the three themes of a scientific collection of publications.

Table 1: A Sample of Topic Presentation on AAAI Dataset

TopicId	Words
7	problem, result,order ,solver, present
12	algorithm,show, state,find,result
18	behavior ,system ,agent, develop ,result

From the above results in Table1, a sample of word distributions used to present three themes of a scientific paper collection of AAAI dataset. The term "result" is a general term and very general term in showing research papers in all different fields. The general words cause ambiguous to the topic presentation. So, a new model is required to solve these problems. The new method should take higher special representations and explore latent associations under multinomial word distributions.

The LDA and other topic models are portions of the better field of probabilistic modeling. Generative probabilistic topic modeling is a method for unsupervised classification of documents, by modeling each document as a mix of themes and each theme as a mix of words. But there exist the problems of word uncertainty and semantic integrity [8].

Text mining is a technique that supports users' assets effective information from a variety of digital documents. Most text mining methods are keyword-based strategies that need single words to show the documents. Based on the theory that the phrase may have more linguistic meaning than the keyword, strategies for using phrases instead of keywords are also suggested. However, surveys have shown that phrase-based techniques are not always better than keyword-based techniques [9, 10]. Many strategies in the field of data mining are used in patterns to mine useful documents that yield encouraging results [11, 12].

Topic modeling provides a convenient way to analyze large classified text collections while extracting interesting features to express collections in text mining. Thus, it advances the proposed model to improve the accuracy and relevance of the topic's representations by using the techniques of text mining, especially frequent itemset mining techniques.

In this model, Latent Dirichlet Allocation (LDA) is integrated into data mining techniques and achieves successful accuracy for the collection documents (corpus). The proposed model is composed of two phases: 1) LDA is applied to accomplish first topic models and 2) the frequent itemset (pattern) mining method is applied to obtain further particular patterns to produce topics of the document collections. Furthermore, the frequent itemsets (patterns) often explain information about the structure of the relationship between words that provide topics that are understandable, relevant and broad.

2. Related Work

Probabilistic topic modeling is expanded to capture more interesting features [13], but they show topics through the distribution of multinomial words. The papers [14, 15] are a widespread way to express the linguistic meaning of the topics as mentioned in the introduction. The authors [16] show a way to calculate the similarities between given themes and a known hierarchy then select the most grant labels to show the topics. But, the weakness of existing methods is that they are strictly limited to resource candidates and are limited to linguistic coverage. The proposed model is a work extension originally presented at 16th IEEE/ACIS International Conference on Computer and Information Science, [17]. This paper [18] discusses the topic-related model phrase by Markov dependencies in word order based on LDA structure, related to this paper. The results provided on [19, 20, 21] show that the topics described by the phrases are easier to interpret than their LDA. But phrases may contribute low-level events in documents, which cannot be accomplished with efficient retrieval performance.

3. Phase1: Topic Presentation Propagation Using LDA

The Latent Dirichlet Allocation is an algorithm that automatically detects the themes that are present in the collection of documents. At the LDA, each document can be viewed as a mix of different topics. The LDA provides the topic using word distribution and representation of the document using the topic distribution. The description of the topic means which words are important to which the topic and representation of the document in which topics are important in the documents [22-26].

Let $D = \{d_1, d_2, ..., d_M\}$ be a corpus. Each document is considered as a mixture of themes and each theme can be defined as a distribution over frozen vocabulary words composed of documents using (1). In general, the proposed model has $\theta=\emptyset_1, \emptyset_2, \dots, \emptyset_V$ for all topics.

Example results of LDA, the topic presentation is shown in Table 2. At document level, table 3 shows LDA's example results, document representation and Table 4 also shows word-topic assignments results of LDA's example.

Table 2: Topic representations – probability distribution over words

Topic	θ
$\emptyset_7$	problem:$\frac{1}{3}$, result:$\frac{4}{12}$,order: $\frac{2}{12}$,solver$\frac{2}{12}$, present : $\frac{2}{12}$
$\emptyset_{12}$	algorithm$\frac{2}{15}$,show : $\frac{4}{15}$, state: $\frac{1}{3}$,find: $\frac{2}{15}$,result : $\frac{2}{15}$
$\emptyset_{18}$	behavior : $\frac{4}{15}$,system : $\frac{2}{15}$,agent : $\frac{1}{15}$, develop: $\frac{2}{15}$, result: $\frac{2}{15}$

LDA contributions are the representation of the topic that uses the word distribution that the words are important to what the topic matter is, and the representation of the document by distributing the topic which themes are important for a particular document. These representations are used for obtaining information, document classification, text mining, machine learning, etc.

Table 3: Document representation-probability distribution over topics

Document	z_7	z_{12}	z_{18}
d_7	0.385	0.123	0.108
d_{12}	0.464	0.118	0.073
d_{18}	0.305	0.11	0.098

Then, specified topics also indicate which words are important in which topics, similar to the representation on the subject. The proposed model performs word-topic emphasis on the LDA for more precise or more specific topic presentation for a given corpus.

4. Phase2: Topic Presentation Expansion

Words with high probability in topic distributions are selected to represent topics in most LDA based applications. For example, the top 5 words for the 3 topics, as shown in Table 2, are: problem, result, order ,solver, present for topic 7, algorithm, show, state, find, result for topic 12 and behavior ,system ,agent, develop ,result for topic 18. So, they are likely to represent the general concepts or common concepts of the three topics and can not describe the three topics that are noteworthy. Furthermore, the words in representations on the topics formed by the LDA are individual single words. Single words provide very little information about relationships between words and very limited language definitions to make the topic matter clear.

In this section, we propose a method based on frequent patterns (itemset) mining techniques, detailed in the following sub-sections, aimed at reducing the above-mentioned problems.

4.1. Frequent Itemsets based on LDA

The frequent itemset mining is the method of mining data in a set of items or words in large data sets. These patterns are usually described in various forms such as frequent itemsets, sequential patterns, or substructures.

In this paper, the proposed model uses frequent itemsets. Typical itemset generally indicates that a set of items often happens together in a transaction dataset. The methods of using frequent itemsets are categorized into three basic skills: horizontal data format, vertical data format, and expected database strategy.

We use the vertical data format in the mine frequent itemsets in this paper and believe that frequent pattern mining based representation can be more meaningful and more accurate to describe topics. In addition, frequent pattern based representations contain structural information that shows the relationship between the words.

Create Transactional Dataset: The aim of the proposed frequent pattern-based method is to discover related words (i.e., frequent itemsets) from the words assigned by LDA to the topics. From this purpose, we build a set of words from each word-topic assignment instead of using the order of words, because, for frequent pattern mining, the frequency of a word within a transaction is less important. A topical document transaction is a set of words without any duplicates. Let $D=\{d_1,d_2,\ldots,d_M\}$be the original document collection, the transactional dataset for topic Z_j is defined as T_j .

For the topics in D, we can develop V transactional datasets. An example of the transactional datasets is described in Table 5, made from the example in Table 4.

Generate Frequent Pattern-based Representation: The basic idea of the proposed method is to reduce frequent itemsets generated from each transactional dataset T_j to represent Z_j. Here the frequent items in each transactional dataset are taken. Then documents associated with each item are converted into the vertical format and a number of documents containing each item are counted. Items are sorted according to their number of hits (i.e. number of documents of each item). Removing noise items that the number of hits is greater than the user provided threshold. If document-set is a subset of another documents-sets in each transactional dataset then items are merged into enhanced frequent itemsets. For example, $w_1{:}d_1d_2{:}3$, $w_2{:}d_1d_3{:}2$, $w_3{:}d_1d_4{:}3$, $w_1{:}d_2d_3{:}2$, $w_8{:}d_1d_2d_3{:}2$, $w_7{:}d_1d_2d_4{:}3$ can be compressed as the enhanced frequent itemsets $w_1,w_2,w_8{:}d_1d_2d_3{:}2$, w_1,w_3, $w_7{:}d_1d_2d4{:}3$ (format - "items" : "document" : "topic (transactional dataset)"). Such meaningless itemsets may harm document filtering tasks using frequent itemsets because it has duplicates the same frequent itemsets. So, we can regard the proposed method as lossless compression because we can cover all the removed frequent itemsets with the exact topic and it can effectively filter out the redundant itemsets. Finally, frequent itemsets in each transactional dataset are reconstructed. ("itemset" and "pattern" are interchangeable in this thesis). Frequent itemsets are the most widely used patterns created from transactional datasets to illustrate useful or interesting patterns. The main idea of the proposed frequent itemset method is to use of frequent patterns generated from each transactional dataset to represent topic Zj. For a given minimal support threshold δ, and itemset p in T_j is frequent if supp (p) >= δ where supp(p) is the support of p where the number of transactions containing p. Take T_7 as an example, which is the transactional dataset for topic Z_7. For a minimal support threshold δ = 2, all the frequent itemsets generated from are given in Table 6. Patterns represent words related to specific and recognizable meanings.

Table 6: The Frequent Itemsets Explored from T_7

Frequent Patterns	minimal support threshold δ
<result>	3
<algorithm,show,state,find>	2

5. Evaluation

We made experiments to evaluate the performance of the proposed method. In this section, we show the results of the evaluation.

5.1. Datasets

Two datasets are used in experiments, containing abstracts of papers published in the AAAI from 2013 to 2014 and NSF from 1990 to 2003. The two datasets contain 548 and 129000 abstracts. The abstracts are obtained from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml) and by using Porter's stemmer in Java (http://www.tartarus.org/~martin/PorterStemmer).

Table 4: Word-topic Assignments

Document	z_7		z_{12}		z_{18}	
	Proportion of Topic	***Terms***	***Proportion of Topic***	***Terms***	***Proportion of Topic***	***Terms***
d_7	0.385	problem, result, order ,solver, present	0.123	algorithm ,state, find ,algorithm	0.108	behavior , system ,agent
d_{12}	0.464	algorithm,show, state,find,result	0.118	spars ,distribute ,find	0.073	class, complex
d_{18}	0.305	algorithm,show, state,find,result	0.11	algorithm, state	0.098	behavior, system ,system

Table 5: Transactional datasets

Document	$z_7(T_7)$		$z_{12}(T_{12})$		$z_{18}(T_{18})$	
	Proportion of Topic	***Terms***	***Proportion of Topic***	***Terms***	***Proportion of Topic***	***Terms***
d_7	0.385	problem, result, order ,solver, present	0.123	algorithm,state, find	0.108	behavior , system ,agent
d_{12}	0.464	algorithm,show, state,find,result	0.118	spars,distribute, find	0.073	class, complex
d_{18}	0.305	algorithm,show, state,find,result	0.11	algorithm,state	0.098	behavior, system

5.2. Settings

Firstly, we are preparing datasets from the UCI Machine Learning Repository and preprocessing of all documents by removing stop and stemming words.

Secondly, we apply the LDA model to construct a topic model with V = 20 topics for each data collection, using the MALLET topic modeling toolkit (http://mallet.cs.umass.edu//index.php). Our experiments show that an inadequate number of topics will mainly lead to abundant expanded patterns in the topic model. We run Gibbs sampling for 1000 iterations, the LDA hyperparameters are α = 50 / V and β = 0.01.

Thirdly, topical transaction datasets for optimizing topic representations are developed.

Finally, frequent itemsets based on topic representations using the proposed method presented in Section 4 are developed. We used 10% of the documents for testing purpose and trained the resting model at 90%.

5.3. Baseline Model

In order to compare the suggested method, the LDA chose the baseline model in the experiments. Examples of the results of the two models (the LDA model and the frequent itemset based on model) are shown in Table 7. The top 10 terms or frequently itemsets in each of the topic presentations produced by two models are shown in Table 7.

Table 7: Topic Presentations of All Models Using the AAAI Dataset

Topic8		**Topic4**	
LDA	***Frequent itemsets***	***LDA***	***Frequent itemsets***
change	algorithm,method	parallel	markov
system	gener	semant	agent
data	differ	complex	popular,demonstre
unsupervise	propos	algorithm	analyze,learn
input	mani	mathemat	design
compact	consid	condit	relationship,topic
benefit	effici	represent	tempor
superior	sever,train	tool	recommend,exploit
state	imag	regular	adapt
trust	target,problem	improve	time

5.4. Results

The objective of the proposed approach and other current topic modeling methodologies is to show the topics of a collection of documents as specific potential. For existing topic modeling methods and the proposed methodology, the topic representations are word distributions or patterns with probabilities. The more specific selected words or patterns are in the representation of the topic, the more precise representation of the topic matter becomes. The performance of the proposed method is evaluated by the use of information entropy. The higher the entropy, the more the proposed model is disordered.

Table 8: Comparison of All Models in Information Entropy Using a collection of documents of AAAI and NSF

Datasets	**Latent Dirichlet Allocation**	**Frequent Itemsets (patterns)**
AAAI	4.86159	2.68171
NSF	28.3747	20.7532

From the above results in Table 8, the suggested method based on frequent itemsets has lower entropy rates than the baseline model. Thus, it can make more exact presentations of the topics of a corpus.

6. Conclusion

This paper proposes a model to produce more discriminative and semantic rich representations for modeling topics in a given collection of documents. The main contribution of this paper is the novel approach of incorporating the pattern mining method and topic modeling method (Latent Dirichlet Allocation) to produce representations based on the pattern for modeling the topics. The test results show that representations based on patterns are more specific than representations developed by the Latent Dirichlet Allocation. In the future, we will examine the structure of the patterns and discover the relationship between words that represent topics at a more granular level.

References

[1] Blei, David M. "Probabilistic topic models." Communications of the ACM 55, no. 4 (2012): 77-84.

[2] Blei, David M., and D. Jon. "McAuliffe. supervised topic models." Advances in Neural Information Processing Systems 20 (2007): 121128.

[3] Blei, David M., and John D. Lafferty. "A correlated topic model of science." The Annals of Applied Statistics (2007): 17-35.

[4] Blei, David M., and John D. Lafferty. "Dynamic topic models." In Proceedings of the 23rd international conference on Machine learning, pp. 113-120. ACM, 2006.

[5] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[6] Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. "Indexing by latent semantic analysis." Journal of the American society for information science 41, no. 6 (1990): 391.

[7] Fürnkranz, Johannes. "A study using n-gram features for text categorization." Austrian Research Institute for Artifical Intelligence 3, no. 1998 (1998): 1-10.

[8] Gao, Yang, Yue Xu, Yuefeng Li, and Bin Liu. "A two-stage approach for generating topic models." In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 221-232. Springer, Berlin, Heidelberg, 2013.

[9] Gupta, Shivani D., and B. P. Vasgi. "Effective Pattern Discovery and Retrieving Relevant Document for Text Mining."

[10] Hofmann, Thomas. "Unsupervised learning by probabilistic latent semantic analysis." Machine learning 42, no. 1 (2001): 177-196.

[11] Lau, Jey Han, David Newman, Sarvnaz Karimi, and Timothy Baldwin. "Best topic word selection for topic labelling." In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 605-613. Association for Computational Linguistics, 2010.

[12] Lau, Jey Han, Karl Grieser, David Newman, and Timothy Baldwin. "Automatic labelling of topic models." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 1536-1545. Association for Computational Linguistics, 2011.

[13] Magatti, Davide, Silvia Calegari, Davide Ciucci, and Fabio Stella. "Automatic labeling of topics." In Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on, pp. 1227-1232. IEEE, 2009.

[14] Moran, Kelly, Byron C. Wallace, and Carla E. Brodley. "Discovering Better AAAI Keywords via Clustering with Community-Sourced Constraints." In AAAI, pp. 1265-1271. 2014.

[15] Náther, Peter. "N-gram based Text Categorization." Lomonosov Moscow State Univ (2005).

[16] Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34, no. 1 (2002): 1-47.

[17] Wai, Than Than, and Sint Sint Aung. "Enhanced frequent itemsets based on topic modeling in information filtering." In Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on, pp. 155-160. IEEE, 2017.

[18] Wang, Chong, and David M. Blei. "Collaborative topic modeling for recommending scientific articles." In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 448-456. ACM, 2011.

[19] Wang, Xuerui, Andrew McCallum, and Xing Wei. "Topical n-grams: Phrase and topic discovery, with an application to information retrieval." In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, pp. 697-702. IEEE, 2007.

[20] Westergaard, D., Stærfeldt, H.H., Tønsberg, C., Jensen, L.J. and Brunak, S., 2017. "Text mining of 15 million full-text scientific articles." bioRxiv, p.162099.

[21] Wu, Sheng-Tang, Yuefeng Li, and Yue Xu. "Deploying approaches for pattern refinement in text mining." In Data Mining, 2006. ICDM'06. Sixth International Conference on, pp. 1157-1161. IEEE, 2006.

[22] Zeng, J., 2012. "A topic modeling toolbox using belief propagation." Journal of Machine Learning Research, 13(Jul), pp.2233-2236.

[23] Zhang, W., Ma, D. and Yao, W., 2014. "Medical Diagnosis Data Mining Based on Improved Apriori Algorithm." JNW, 9(5), pp.1339-1345.

[24] Zhao, Wayne Xin, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. "Topical keyphrase extraction from twitter." In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, pp. 379-388. Association for Computational Linguistics, 2011.

[25] Zhu, X., Ming, Z., Hao, Y. and Zhu, X., 2015, June. "Tackling Data Sparseness in Recommendation using Social Media based Topic Hierarchy Modeling." In IJCAI (pp. 2415-2423).

[26] Zhu, J., Wang, K., Wu, Y., Hu, Z. and Wang, H., 2016. "Mining User-Aware Rare Sequential Topic Patterns in Document Streams." IEEE Transactions on Knowledge and Data Engineering, 28(7), pp.1790-1804.