American Journal of Mathematical and Computer Modelling
2017; 2(3): 84-87
http://www.sciencepublishinggroup.com/j/ajmcm
doi: 10.11648/j.ajmcm.20170203.12

Review Article

Study and Analysis of Topic Modelling Methods and Tools – A Survey

Himanshu Sharma[1], Arvind K. Sharma[2]

[1]School of CSE, Jaipur National University, Jaipur, India

[2]Dept of CSI, University of Kota, Kota, Rajasthan, India

Email address:
drarvindkumarsharma@gmail.com (A. K. Sharma)

To cite this article:
Himanshu Sharma, Arvind K. Sharma. Study and Analysis of Topic Modelling Methods and Tools – A Survey. *American Journal of Mathematical and Computer Modelling*. Vol. 2, No. 3, 2017, pp. 84-87. doi: 10.11648/j.ajmcm.20170203.12

Received: January 30, 2017; **Accepted**: February 18, 2017; **Published**: March 9, 2017

Abstract: Now days, topic models have been widely used to identify topics in text corpora. Topic modelling is a mechanism of extracting common topics which occurs among the collection of documents. Topic models are actually a suite of algorithms which uncover the hidden thematic structure in document collections. These algorithms shall definitely be help to develop new paradigms to search, browse and summarize large archive of texts. This paper presents a survey of various important topic modelling techniques and tools which highlights the probabilistic topic models. The primary aim of this paper is to help researchers who do not have a strong background in mathematics or statistics to feel comfortable with using topic modelling methods and tools in their research work. Apart from it, the merits and demerits of topic modelling methods are also summarized.

Keywords: Topic Models, Topic Modelling Methods, LSA, PLSA, LDA, CTM, Tools

1. Introduction

Today, the amount of text documents available on the Internet is vast and increasing with an explosive rate. Topic modelling provides an easy way to process large amounts of information efficiently. It also allows for individual search topics to be discovered. Text categorization has become one of the key techniques for managing and organizing those documents and also assists the information retrieval process in filtering the documents for a specific topic. The earliest work on topic modelling is performed by Papadimitriou, Tamaki, Raghavan, and Vempala [1], and Hofmann [2]. The technique was further developed by Blei, Ng, and Jordan [3]. There are a variety of different methods for topic modelling, using different sampling algorithms for word selection and topic creation. Examples of topic models include latent semantic analysis.

2. Topic Modelling– Background

Topic Modeling provides a convenient way to analyze big unclassified text. A topic contains a cluster of words that frequently occurs together. A topic modeling can connect words with similar meanings and distinguish between uses of words with multiple meanings. A document may have depth about a particular topic or multiple topics with different proportions. By using mathematical framework topic model examines each and every documents and discovers based on statistical words in each documents. Topic modelling is a generic term for several different algorithms that create topics based on documents. The topic modelling algorithm used in this thesis was latent dirichletallocation, therefore all explanations are of latent dirichlet allocation. It can be used in a variety of ways. To get a good understanding of how topic modelling works some simple examples of topic modelling will be discussed first. Topic models allow for discovering topic in a corpus. Often researchers will have a large collection of texts to read through. To get a good idea of the contents of a corpus, a topic model can provide a good overview [4]. The evolution of topics over time can be modelled to explore trends in language and ideas. The classifications of topic models are shown in Figure 1.

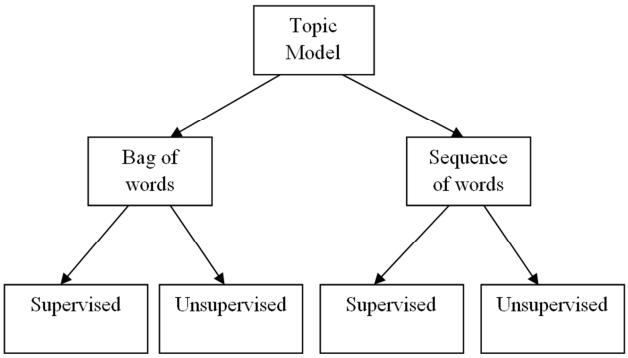

Figure 1. *The Classification of Topic Models.*

3. Literature Review

In this section, we cover evolution of topic models and what methods have been used in extraction of topics. There are some recent work which considers modelling with different types of topics. Many researchers have recently proposed new topic models which would learn more than one dimension of topics. Some of them are recorded here.

In [5]; D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation", proposed Topic modelling has become one of the most popular probabilistic text modelling techniques and has been quickly accepted by machine learning and text mining communities. It could automatically classify documents in a collection by a number of topics and represents every document with multiple topics and their corresponding distribution, it has ambiguity.

In [6]; H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, "Enriching text representation with frequent pattern mining for probabilistic topic modelling, "proposed frequent patterns were pre-generated from the original documents and then inserted into the original documents as part of the input to a topic modelling model such as LDA. The resulting topic representations contain both individual words and pre-generated patterns.

In [7]; X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," proposed the topical n-Gram model proposed in automatically and simultaneously discovers topics and extracts topically relevant phrases. It has been seamlessly integrated into the language modelling based IR task, compared with word representation, phrases are more discriminative and carry more concrete semantics. Since phrases are less ambiguous than words, they have been widely explored as text representation for text retrieval, but few studies in this area have shown significant improvements in effectiveness.

Topic modelling have been started with Topic detection and Tracking (TDT) project [8, 9] and extensively studied in the literature. TDT is used to find and track topic from a sequence of news sequences and the technique followed is clustering based. Later came into existence of Probabilistic Latent semantic analysis (PLSA) [10], Latent Dirichlet Allocation (LDA) and their derivatives.

Bhei et al. [11], has proposed Dynamically Topic model (DTM) which enables to model topic over evolution of time. DTM is an enhancement to LDA. The advantage of DTM when compared to other probabilistic topic modelling algorithm is that it tracks topic over a period of time. Problem with DTM is that has fixed number of topics and a discrete notion of time.

Bhei et al. [12], has introduced hierarchical LDA that is the LDA's extension model is Hierarchical LDA and introduced by Bhei in 2003. LDA model a flat topic structure instead HLDA models tree of topics. Hierarchical LDA uses non-parametric Bayesian approach to model hierarchies. Tree of topic is constructed hierarchically of nodes by an algorithm. In topic tree model each and every node is represented by random number and has got corresponding word-topic distribution assigned to it. The tree can be traversed from the root till its leaves while sampling topics along the path [13].

4. Methods of Topic Modelling

In this section, some of the popular topic modelling methods are discussed which deal with words, documents and topics. This section provides a category that can be considered under the field of topic modelling. We discuss the area of methods of Topic Modelling, which has four methods such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM). Which are described one by one.

4.1. Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a method or a technique in the area of Natural Language Processing (NLP). The main goal of Latent Semantic Analysis (LSA) is to create vector based representation for texts 'to make semantic content. By vector representation (LSA) computes the similarity between texts 'to pick the heist efficient related words. In the past LSA was named as Latent Semantic Indexing (LSI) but improved for information retrieval tasking. So, finding few documents that are close to the query has been selected from many documents. LSA should have many aspects to give approach such as key words matching, Wight key words matching and vector representation depends on occurrences of words in documents. Also, Latent Semantic Analysis (LSA) uses Singular Value Decomposition (SVD) to rearrange the data.

4.2. Probabilistic Latent Semantic Analysis (PLSA)

Probabilistic Latent Semantic Analysis (PLSA) is an approach that has been released after LSA method to fix some disadvantages that have found into LSA. Jan Puzicha and Thomas Hofmann introduced it in the year 1999. PLSA is a method that can automate document indexing which is based on a statistical latent class model for factor analysis of count data, and also this method tries to improve the Latent Semantic Analysis (LSA) in a probabilistic sense by using a generative model. The main goal of PLSA is to identifying and

distinguishing between different contexts of word usage without recourse to a dictionary or thesaurus. It includes two important implications: First one, it allows to disambiguate polysemy, i.e., words with multiple meanings. Second thing, it discloses typical similarities by grouping together words that shared a common context [14].

4.3. Latent Dirichlet Allocation (LDA)

The reason of appearance of Latent Dirichlet Allocation (LDA) model is to improve the way of mixture models that capture the exchangeability of both words and documents from the old way by PLSA and LSA. This was happened in 1990, so the classic representation theorem lays down that any collection of exchangeable random variables has a representation as a mixture distribution—in general an infinite mixture [15]. There are huge numbers of electronic document collections such as the web, scientifically interesting blogs, news articles and literature in the recent past has posed several new challenges to researchers in the data mining community. Especially there is a growing need of automatic techniques to visualize, analyze and summarize these document collections. In the recent past, latent topic modeling has become very popular as a completely unsupervised technique for topic discovery in large document collections. This model, such as LDA [16].

4.4. Correlated Topic Model (CTM)

Correlated Topic Model (CTM) is a kind of statistical model used in natural language processing and machine learning. Correlated Topic Model (CTM) used to discover the topics that shown in a group of documents. The key for CTM is the logistic normal distribution. Correlated Topic Models (CTM) is depending on LDA [17].

The Analysis of these four topic modelling methods in summarised in the table-1 below.

Table 1. *The Analysis of Topic Modelling Methods [18].*

S.No.	Name of The Methods	Merits	Demerits
1.	Latent Semantic Analysis (LSA)	- LSA can get from the topic ifthere are any synonym words. -Not robust statistical background	- It is hard to obtain and to determinethe number of topics. - To interpret loading values withprobability meaning, it is hard tooperate it.
2.	Probabilistic Latent Semantic Analysis (PLSA)	-It can generate each word from asingle topic; even though variouswords in one document may begenerated from different topics. - PLSA handles polysemy.	At the level of documents, PLSAcannot do probabilistic model.
3.	Latent Dirichelet Allocation (LDA)	-Need to manually remove stopwords. -It is found that the LDA cannotmake the representation ofrelationships among topics.	- It becomes unable to model relationsamong topics that can be solved in CTMmethod.
4.	Correlated Topic Model (CTM)	-Using of logistic normaldistribution to create relationsamong topics. -Allows the occurrences of words in other topics and topic graphs.	- Require lots of calculation - Having lots of general words insidethe topics.

5. Tools for Topic Modelling

There are various text analysis software tools available for use and which work upon on different methods of topic modelling. Each one requires different skills for use. Some of the more popular software tools for Topic Modelling are discussed in this section.

- *'R' Tool*

R is a free to download, statistical software package. It runs on Windows, MacOS and UNIX systems. R has different packages available to download and install for different purposes. It is command-line based but there are packages such as RStudio that provide a graphical user interface for running an analysis. There is also a large library of literature on how to use R and run different packages. There are three packages capable of doing topic modelling analysis. They are mallet, topic models, and LDA.

- *Mallet*

The mallet package in R is based on the Mallet software package but rather than being capable of running a large selection of text analysis algorithms it can only do topic modelling analysis. It provides an interface to the Java implementation of latent dirichlet allocation [19]

- *Gensim*

Gensim is a Python based program that allows for more customisation. The website has a comprehensive tutorial on how to run a topic model analysis. It runs latent semantic analysis, latent dirichlet allocation and random projections. The preferred file formats are plain text files [20].

- *LDA-C*

LDA-C is a C implementation of latent dirichlet allocation software. It was one of the earlier software programs available so does not have as many options to tailor the analysis. It runs a basic latent dirichlet allocation model [21].

- *GibbsLDA++*

GibbsLDA++ is a C and C++ implementation of latent dirichlet allocation. It uses Gibbs sampling for fitting the model. GibbsLDA++ is also an older program so it does not have as many features as more recent programs. There is little room for customisation [22].

6. Conclusion

This paper starts with the description of a topic model, with a focus on the understanding of topic modelling. This survey

paper presented most popular topic modelling methods and tools in text mining. Firstly, it has discussed a background of topic modelling and general idea about the four topic modelling methods including Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM). Also the merits and demerits of these four methods are presented. This paper opens a new door for the researchers in the topic modelling domain.

References

[1] Papadimitriou, C. H., Tamaki, H., Raghavan, P., &Vempala, S. (1998). Latent semantic indexing: A Probabilistic Analysis, Paper presented at the Proceedings of the Seventeenth ACM Sigact-Sigmod-Sigart Symposium on Principles of Database Systems.

[2] Hofmann, T. (1999), Probabilistic Latent Semantic Indexing, Paper presented at the Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), Latent Dirichlet Allocation, the Journal of Machine Learning Research, 3, 993-1022.

[4] Rebecca Katherine Abey, The Statistics of Topic Modelling, University of Canterbury, 2015.

[5] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, "Mining Frequent Patterns With Counting Inference, " ACM SIGKDD Explorations Newsletter, Vol.2, No.2, pp.66–75, 2000.

[6] M. J. Zaki and C. J. Hsiao, "CHARM: An Efficient Algorithm for Closed Itemset Mining", in Proceedings, SDM, Vol.2, 2002, pp.457–473.

[7] X. Wei and W. B. Croft, "LDA-based Document Models for Ad-Hoc Retrieval", in Proceedings 29th Annual International, ACM SIGIR Conf. Res. Develop. Information Retrieval, 2006, pp.178–185.

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

[9] Mark Steyvers, Tom Griffiths, "Probabilistic Topic Models", In Landauer.

[10] Zhu, Jun and Eric P Xing, "Conditional Topic Random Fields", Forbes. Ed. Johannes Fürnkranzand Thorsten Joachims.

[11] A. Gruber, M. Rosen-Zvis and Y. Weiss, "Hidden Topic Markov Models", in Artificial Intelligence and Statistics, 2007.

[12] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, "Integrating Topics and Syntax", In Advances in Neural Information Processing Systems 17, Vol.17, 2005, pp. 537-44.

[13] M. Divya, et al., "A Survey on Topic Modelling", International Journal of Recent Advances in Engineering & Technology (IJRAET), Volume-1, Issue - 2, 2013.

[14] Hofmann, T., Unsupervised learning by probabilistic latent semantic analysis, Machine Learning, 42 (1), 2001, 177-196.

[15] Blei, D. M., Ng, A. Y., and Jordan, M. I., -Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 2003, 993-1022.

[16] Ahmed, A., Xing, E. P., and William W., -Joint Latent Topic Models for Text and Citations, ACM New York, NY, USA, 2008.

[17] RubayyiAlghamdi et al., A Survey of Topic Modeling in Text Mining, International Journal of Advanced Computer Science and Applications, Vol.6, No.1, 2015.

[18] Lee, S., Baker, J., Song, J., and Wetherbe, J. C., -An Empirical Comparison of Four Text Mining Methods, Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010.

[19] Mimno, D. (2015). Package 'mallet' Packages.

[20] Řehůřek, R., &Sojka, P. (2011), Gensim–Python Framework for Vector Space Modelling, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic.

[21] Blei, (2012), Topic Modelling and Digital Humanities, Journal of Digital Humanities, 2 (1), 8-11.

[22] Phan, X. -H., & Nguyen, C. T. (2007), GibbsLDA++: AC/C++ Implementation of Latent Dirichlet Allocation (LDA): Technical report.