International Journal For Technological Research In Engineering
Volume 5, Issue 10, June-2018 *ISSN (Online): 2347 - 4718*

SURVEY PAPER ON TOPIC MODELLING

Himani Trivedi
Faculty of Computer Engineering, LDRP-ITR, Gandhinagar, Gujarat.

Abstract: 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. This paper provides two categories that can be considered under the field of topic modeling. First one discusses the area of methods of Topic Modeling, which has four methods and can be considered under this category. These methods are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM). The second category is called Topic Evolution Model, it considers an important factor time. In this category, different models are discussed, such as Topic Over Time (TOT), Dynamic Topic Models (DTM), Multiscale Topic Tomography, Dynamic Topic Correlation Detection, Detecting Topic Evolution in scientific literatures, etc.

Keywords: Twitter, Social Media, Topic Modeling, Methods of Topic Modeling, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA).

I. INTRODUCTION

In recent years, social networks such as Facebook, Myspace and Twitter have become important communication tools for people Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To have a better way of managing the explosion of electronic document archives these days, it requires using new techniques or tools that deals with automatically organizing, searching, indexing, and browsing large collections. On the basis of today‘ s research of machine learning and statistics, it has developed new techniques for finding patterns of words in document collections using hierarchical probabilistic models. These models are called ―topic models. Discovering of patterns often reflect the underlying topics that are united to form the documents, such as hierarchical probabilistic models are easily generalized to other kinds of data; topic models have been used to analyze things rather than words such as images, biological data, and survey information and data [1]. The main importance of topic modeling is to discover patterns of word-use and how to connect documents that share similar patterns. So, the idea of topic models is that term which can be working with documents and these documents are mixtures of topics, where a topic is a probability distribution over words. In other word, topic model is a generative model for documents. It specifies a simple probabilistic procedure by which documents can be generated[2][3].

Recent studies in a variety of research areas show increasing interests in micro-blogging services, especially Twitter. Early work mainly focused on quantitative studies on a number of aspects and characteristics of Twitter. For example, Java et al. [6] studied the topological and geographical properties of Twitter’ s social network in 2007 and found that the network has high degree correlation and reciprocity, indicating close mutual acquaintances among users. Krishnamurthy et al. [7] studied the geographical distribution of Twitter users and their behaviors among several independent crawls. The authors mostly agree with the classification of user intentions presented by Java et al., but also point out evangelists and miscreants (spammers) that are looking to follow anyone. Weng et al. [19] studied the problem of identifying influential users on Twitter by proposing an extension of the PageRank algorithm to measure the influence taking both the topical similarity between users and the link structure into account. They also presented evidence to support the existence of homophily in Twitter. In their work, they utilized topic models (described below) to understand users’ interests[21],[22]. Among the research mentioned above and others, researchers wish to use messages posted by users to infer users’ interests, model social relationships, track news stories and identify emerging topics. However, several natural limitations of messages prevent some standard text mining tools to be employed with their full potentials. First, messages on Twitter (which are called “ tweets”) are restricted to 140 characters. This is substantially different from traditional information retrieval and web search. Second, within this short length, users invented many techniques to expand the semantics that are carried out by the messages. For example, when posting external URLs, users may use URL shortening services (e.g., http://www.bit.ly). In addition, users heavily use self-defined hash tags starting with “ #” to identify certain events or topics. Therefore, from the perspective of length (e.g., in characters), the content in messages is limited while it may convey rich meanings.

Topic models [4] are powerful tools to identify latent text patterns in the content. They are applied in a wide range of areas including recent work on Twitter (e.g., [5][8]). Social media differs from some standard text domain (e.g., citation network, web pages) where topic models are usually utilized in a number of ways. One important fact is that there exists many “ aggregation strategies” in social media that we usually want to consider them simultaneously. For example, on Twitter, we usually want to obtain topics associated with messages and their authors as well. Researchers typically only discuss one of them. Weng et al. [19] trained a topic model on aggregated users’ messages while Ramage et al. [13] used a slightly modified topic model on individual messages. Neither of them mentioned the other possibility.

Indeed, to our knowledge, there is no empirical or theoretical study to show which method is more effective, or whether there exists some more powerful way to train the models.

II. METHODOLOGY

In this section, we will introduce several methods to train topic models on Twitter and discuss their technical details. In this paper, we mainly consider two basic models: LDA and author-topic model [15]. We first briefly review these two models and then discuss their adaptation to Twitter.

Latent Dirichlet Allocation

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 [9]. 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 [10][12].

Latent Dirichlet Allocation (LDA) is an Algorithm for text mining that is based on statistical (Bayesian) topic models and it is very widely used. LDA is a generative model that tries to mimic what the writing process is. So it tries to generate a document on the given topic. It can also be applied to other types of data. There are tens of LDA based models including: temporal text mining, author- topic analysis, supervised topic models, latent Dirichlet co-clustering and LDA based bioinformatics [11], [18]. In a simple way, the basic idea of the process is, each document is modeled as a mixture of topics, and each topic is a discrete probability distribution that defines how likely each word is to appear in a given topic[16]. These topic probabilities provide a concise representation of a document. Here, a "document" is a "bag of words" with no structure beyond the topic and word statistics. LDA models each of D documents as a mixture over K latent topics, each of which describes a multinomial distribution over a W word vocabulary[14],[15]. Figure 1 shows the graphical model representation of the LDA model. The generative process for the basic LDA is as follows:

For each of Nj words in document j
1) Choose a topic zij ~ Mult(θj)
2) Choose a word xij ~ Mult(φzij)

Where the parameters of the multinomials for topics in a document θj and words in a topic φk have Dirichlet priors [13]

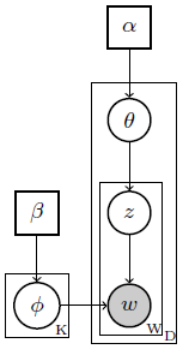

Fig 1: A Graphical Model Representation of LDA

Overview of other methods in topic modelling

TABLE I. THE CHARACTERISTICS OF TOPIC MODELING METHODS [17]

Name of The Methods	Characteristics
Latent Semantic Analysis (LSA)	LSA can get from the topic if there are any synonym words. Not robust statistical background.
Probabilistic Latent Semantic Analysis (PLSA)	It can generate each word from a single topic; even though various words in one document may be generated from different topics.PLSA handles polysemy.
Latent Dirichelet Allocation (LDA)	Need to manually remove stop words. It is found that the LDA cannot make the representation of relationships among topics.
Correlated Topic Model (CTM)	Using of logistic normal distribution to create relations among topics. Allows the occurrences of words in other topics and topic graphs.

III. CONCLUSION

Model topics without taking into account time will confound the topic discovery. In the second category, paper has discussed the topic evolution models, considering time. Several papers have used different methods of model topic evolution. Some of them have used discretizing time, continuous-time model, or citation relationship as well as time discretization. Topic models are useful in exploring large-scale specialized corpora in a bottom-up way.This leads to insights into - how they change over time - how they change within papers, and - how each text is characterized in terms of topics. Topic models have been extensively researched in machine learning and computational linguistics, and a number of extensions have been proposed. topic models using n-grams, correlated topic models that

allow correlation between, dynamic topic models that account for the chronological change of keywords within topics, automated ways to identify the optimal number of topics, automated ways to compute coherence of each topic.

REFERENCES

[1] D. Blei and J. Lafferty. Topic models. Text Mining: Theory and Applications, 2009.

[2] D. M. Blei and J. D. Mcauliffe. Supervised topic models. Advances in Neural Information Processing Systems 21, 2007.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993– 1022, 2003.

[4] J. Chang, J. Boyd-Graber, and D. M. Blei. Connections between the lines: augmenting social networks with text. In KDD ' 09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 169– 178, 2009.

[5] T. L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101:5228– 5235, 2004.

[6] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In WebKDD/SNA-KDD ' 07: Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web mining and Social Network Analysis, pages 56– 65, 2007.

[7] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In WOSP ' 08: Proceedings of the First Workshop on Online Social Networks, pages 19– 24, 2008.

[8] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: joint models of topic and author community. In ICML ' 09: Proceedings of the 26th Annual International Conference on Machine Learning, pages 665– 672. ACM, 2009.

[9] C. D. Manning, P. Raghavan, and H. Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.

[10] A. McCallum, X. Wang, and N. Mohanty. Joint group and topic discovery from relations and text. In Statistical Network Analysis: Models, Issues and New Directions, volume 4503 of Lecture Notes in Computer Science, pages 28– 44. Springer-Verlag, Berlin, Heidelberg, 2007.

[11] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In KDD ' 08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 542– 550, 2008.

[12] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In WWW ' 08: Proceedings of the 17th International Conference on World Wide Web, pages 91– 100, 2008.

[13] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In International AAAI Conference on Weblogs and Social Media, 2010.

[14] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In EMNLP ' 09: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 248– 256. Association for Computational Linguistics, 2009.

[15] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. ACM Transactions on Information Systems, 28(1):1– 38, 2010.

[16] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In UAI ' 04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pages 487– 494, 2004.

[17] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In WWW ' 06: Proceedings of the 15th International Conference on World Wide Web, pages 377– 386, 2006.

[18] H. M. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems, 2009.

[19] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In WSDM ' 10: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pages 261– 270, 2010.

[20] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In KDD ' 09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 937– 946, 2009.

[21] W.-T. Yih and C. Meek. Improving similarity measures for short segments of text. In AAAI' 07: Proceedings of the 22nd National Conference on Artificial Intelligence, pages 1489– 1494, 2007.

[22] H. Zhang, C. L. Giles, H. C. Foley, and J. Yen. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In AAAI' 07: Proceedings of the 22nd National Conference on Artificial Intelligence, pages 663– 668, 2007

