Literature Survey on Topic Modeling

A. S. M. Ashique Mahmood
Dept. of CIS
University of Delaware
Newark, Delaware
ashique@udel.edu

Abstract

Topic Modeling refers to a suit of algorithms that gives us an insight of the 'latent' semantic topics or themes in a collection of documents. This survey provides a brief classification of different topic modeling techniques and an introductory overview of the most popular topic modeling technique LDA (latent Dirichlet Allocation) and some of its extensions. This survey also summarizes few applications of topic modeling in different fields of study.

1 Introduction

In this age of information technology, collective knowledge is continuously being digitized and stored in many forms such as blogs, news, research articles, social networks, webpages etc. We have fairly sophisticated search engines to find information from these collections. A typical search engine would let us search for terms and provide a list of documents that it deems most relevant to our search needs. But, each document has some 'themes' running through it which would give us a more general or zoomed-out view of topics that it is talking about. It would be useful to see through these themes on a broader level to decide the utility of the document. Topic modeling does this exact thing- models each document across a combination of hidden topics or themes and groups the words together that represent the topics.

Technically speaking, Topic modeling refers to a group of machine learning algorithms that infer the latent structure behind a collection of documents. The intuition behind topic models is that each document is comprised of some 'topics' or 'themes'. A "topic" is understood as a collection of words that represent the topic as a whole.

For example, consider the article in Figure 1, titled "Seeking Life's Bare (Genetic) Necessities". Different words are highlighted with different colors where each color represents a hidden 'topic'. A human can manually get the sense of topics by reading the document. Topic modeling tries to do the same- come up with the topics in a probabilistic way (topics are shown on the left of the figure) that captures the thematic structure of documents. Topic modeling techniques are statistical methods that analyze the words of documents and discover the topics that run through them. Knowing these topics would help us in many ways- judging the utility of the document according to needs, clustering documents based on similar topics etc.

This survey provides an introductory survey of topic modeling. Latent Dirichlet Allocation (LDA) is the simplest and most popular statistical topic modeling (Blei et el. 2003). Many other models came up as an extension of LDA. We would describe LDA first and then briefly describe few of its extensions. Topic modeling is used in various fields of study. We describe some of its known uses to depict the diverse use of topic modeling. Few other surveys of topic models already exist; among most significant are Daud et el. (2010), Blei (2012), Jelisavcic et el. (2012) etc. They are written from different perspectives. We hope that our survey would serve as a complement to those.

Rest of the paper is organized as follows. Section 2 gives a brief overview of topic modeling classifications. Section 3 describes LDA and few of its variants. Section 4 talks about applications of

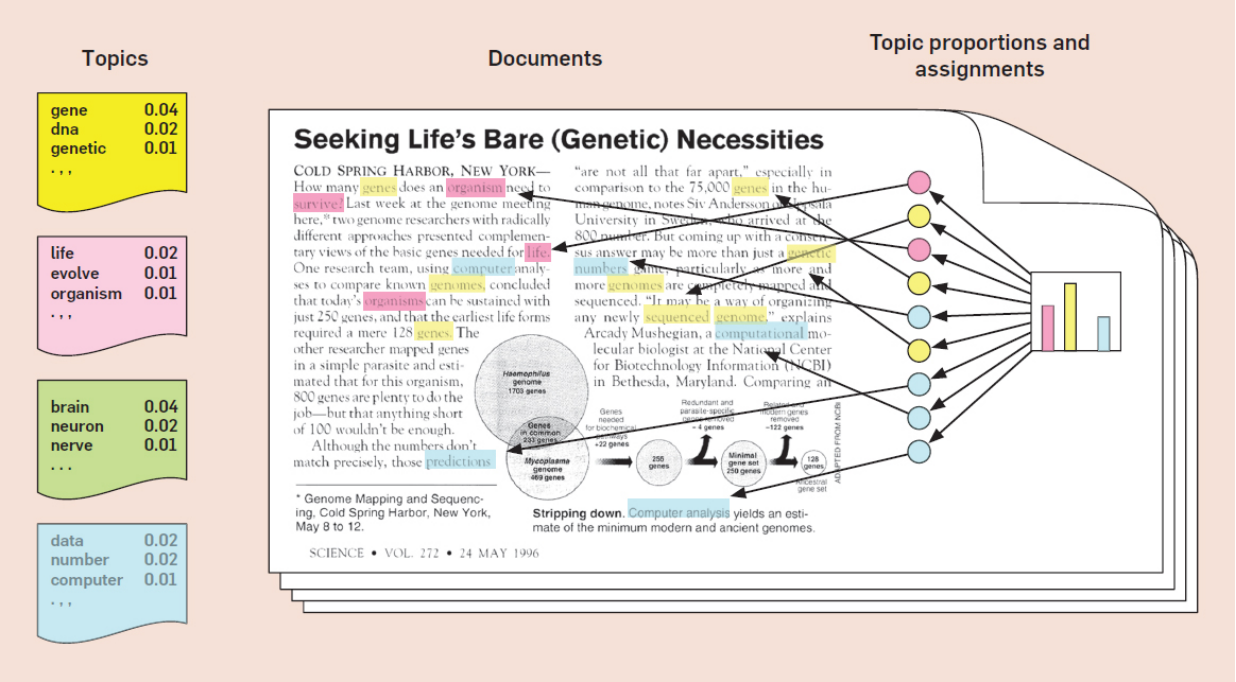

Figure-1: The intuition behind latent Dirichlet allocation. The document contains some number of topics, and the words for those topics highlighted in different colors. On the left, the assumed topics are shown. On the very right, topic proportions for this document are shown.

topic modeling. Section 5 concludes the paper with future research scopes of topic modeling.

2 Classification

Topic modeling is classified by the some important assumptions it makes about the documents (Blei 2012) (Jelisavcic et el. 2012). The first assumption is about the word ordering. A simpler and more useful approach is the "bag of words" model which neglects the word orderings. It sounds unrealistic but it performs reasonably well for finding coarse semantic structure of documents. There are topic modeling approaches which take the ordering into account. For example (Wallach 2006) assumes that topics generate words which are conditional on the previous word. The second assumption is about the use of in-domain knowledge. Using in-domain knowledge would improve topic quality but the model tends to grow complex. Third assumption is about the dependability on labeled data. The main idea behind topic modeling was to learn topics in an unsupervised manner so that it can be applied to a broad range of problems without the need of expensive labeled data. Mostly topic models are unsupervised. There are some semi-supervised and supervised models (Blei and McAuliffe, 2007) for problems that already have sufficient labeled data. Table-1 categorizes some topic modeling techniques based on the mentioned criteria.

Technique	*B/S*	*N/D*	*U/S*
pLSI (Hoffman 1999)	B	N	U
LDA (Blei 2003)	B	N	U
hLDA (Blei et el. 2003)	B	N	U
DTM (Blei and Lafferty, 2006)	B	N	U
CorrLDA (Blei 2006)	B	N	U
ATM (Rosen-Zvi et el., 2004)	B	N	U
sLDA (Blei 2007)	B	N	S
sCTRf (Zhu 2010)	S	D	S

Table-1: Some topic modeling technique with classifications. B/S is bag-of-words vs sequenced words. N/D is no in-domain-knowledge vs in-domain knowledge. U/S is unsupervised vs supervised.

3 Latent Dirichlet Allocation (LDA)

LDA is an unsupervised, non-parameterized and generative probabilistic topic modeling (Blei et el. 2003) which assumes that each document is a probability distribution of topics and each topic is a probability distribution of words from the document. LDA uses a "bag of words" approach, which treats each document as a vector of word

counts. It is a "generative" process because documents arose over time and topics are specified before any data is generated. Thus the main characteristic of LDA is that all the documents in the collection share the same set of topics, but each document contains those topics with different proportions. As this is a generative process, the documents are observed one by one while the *hidden structure*- the combination of topics, per-document topic distributions, and per-word topic assignment-persists. Hence, the main computational problem for LDA is to infer the hidden structure.

The generative process for LDA is defined as follows (Blei 2012):

1. Randomly choose a distribution over topics.
2. For each word in the document:

   i) Randomly pick a topic from the distribution over topics.
   ii) Randomly pick a word from the distribution over words associated with the chosen topic.

As a generative probabilistic model, documents are seen arising from a generative process that has *hidden variables*. This process defines a joint probability distribution over both the hidden and observed variables. This joint distribution is used to infer the hidden variables given the observed variables. This conditional distribution is called the posterior distribution. In LDA, the observed variables are the words of the documents and the hidden variables are the topic structure.

More formally, the topics are $\beta_{1:k}$, where each β_k is the distribution over the vocabulary. The topic proportion for document *d* is θ_d , where $\theta_{d,k}$ is the topic proportion for topic k in document d. Topic assignment for the d-th document is Z_d , where $Z_{d,n}$ is the topic assignment for n-th word in document d. Observed words for document d are w_d , where $w_{d,n}$ is the n-th observed word for document d. This generative process is repeated N_d times where N_d is the total number of words in the document d. With these notations, the generative process for LDA is represented by the joint probability distribution of the hidden and observed variables:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right).$$

Figure 2 shows the graphical model (plate notation) for LDA.

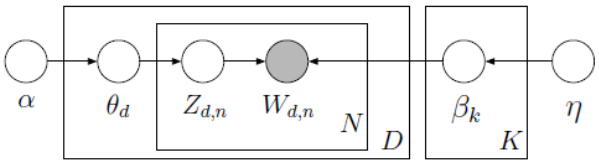

Figure-2: Plate notation for LDA

Latent Dirichlet allocation forms the basis of many other topic modeling techniques. Some of the popular ones are briefly introduced in the subsequent subsections.

3.1 Dynamic topic model

Dynamic Topic Model (DTM) is introduced by (Blei and Lafferty, 2006) as an extension of LDA that captures the evolution of topics over time in a sequentially organized corpus of documents. The notion of time was included using the meta-data of the documents. DTM exhibits the evolution of word-topic distribution which makes it easy to view the topic trends. Rather than a single distribution over words, a topic is now a sequence of distributions over words. We can glean the underlying themes of the collection and track how they have changed over time. Figure-3 shows a topic evolution that resulted from applying DTM on *Science* magazine articles starting from 1880 (Blei and Lafferty, 2006).

Fixed number of topics proves to be a disadvantage for DTM as many topics grow and die over time in a corpus (Blei and Lafferty, 2006).

3.2 Hierarchical LDA

Hierarchical LDA is introduced by (Blei et el. 2003) which builds a hierarchical topic model by combining the prior with the likelihood that is based on a hierarchical variant of latent Dirichlet allocation. This results in a flexible, general model

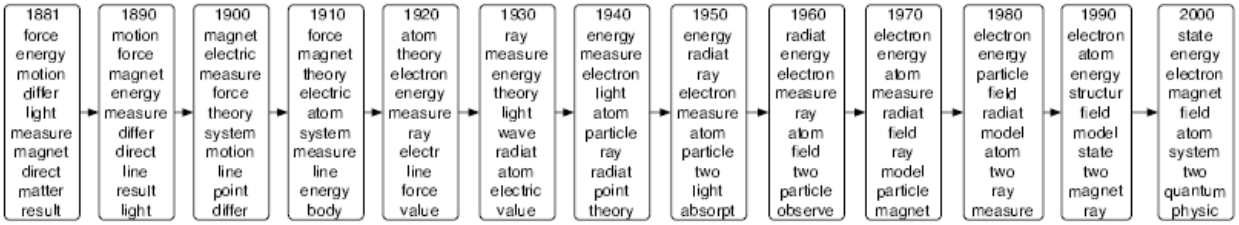

Figure-3: Evolution of topic "Atomic Physics" starting from 1881 to 2000.

for topic hierarchies that naturally accommodates the growing data as they become available. Like LDA, every node in the tree represents a random variable, and each has a word-topic distribution assigned. Documents are generated by traversing the tree from root to one of it's leaves and sampling for topics on the way. Blei et el. (2003) made two assumptions for hLDA. Firstly, they restricted the hierarchy to a certain depth *L*. Secondly, each document is associated with a single path although ideally it can mix over multiple paths.

3.3 Correlated topic model

LDA cannot model the correlations among topics. For example the topic "genetics" is more likely to be similar to "disease" than to "astronaut". LDA fails to depict this correlation of topics. Correlated topic model (CTM) is introduced by Blei and Lafferty (2007) which is an extension of LDA that can model the correlations among topics.

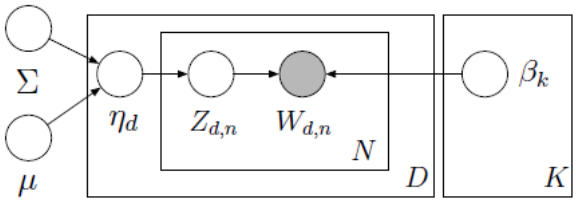

Figure-4: Plate notation for correlated topic model

CTM provides a graph representation of topic relationships whereas LDA imposes a mutual independence among topics. For posterior inference in this model, they employed a mean-field variational algorithm to form a factorized distribution of the latent variables, parameterized by free variables which are called the variational parameters (Blei and Lafferty, 2007). These parameters are chosen such that the K-L divergence between the approximate and true posterior is small. Figure-4 shows the CTM in plate notation.

3.4 Author-topic model

The author-topic model (ATM) is an extension of LDA, first proposed by Rosen-Zvi et. el (2004) and later expanded by Rosen-Zvi et. el (2010). Using the meta-data present in the documents, ATM models the topics distribution corresponding to each author in the corpora. Under this model, each word in a document is associated with two latent variables: an author and a topic. Similar to LDA, each author is seen as a distribution of topics and each topic is seen as a distribution of words. But unlike LDA, along with the words, authors are also the observed variables. The main intuition behind the author-topic model is to allow us to explicitly include authors in document models, providing a general framework to predict at the level of authors as well as the level of documents.

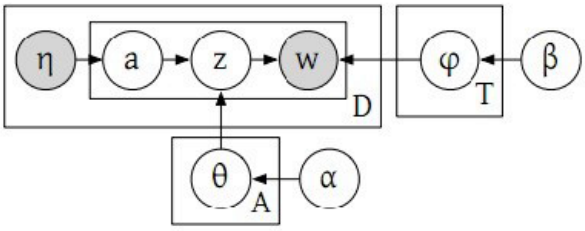

Figure-5: Plate notation for author-topic model

4 Applications of topic modeling

Topic modeling is used in diverse fields of study-from biomedical domain to scientific knowledge discovery to social media analysis etc. We took the opportunity to highlight some of them here.

Topic modeling on unstructured nursing notes for ICU patients is used to stratify the risk and mortality prediction for the hospital. Hierarchical Dirichlet Processes (HDP), a non-parametric topic modeling technique, is used to automatically discover "topics" as shared groups of co-occurring UMLS clinical concepts (Lehman et el. 2012). This topic structure significantly improves the performance of SAPS-I algorithm in hospital mortality prediction.

Drug with similar side effects are likely to be effective for same disease. So, a probabilistic topic model was constructed from the warnings, precautions and adverse effects from the drug labels (Bisgin et el. 2012). Then using this topic distribution, similarity of drugs was found. This allows the reposition of drugs, meaning drugs with more adverse effects can be replaced with safer alternatives.

Chen et el. (2012) conducted a study to estimate the functional groups in human gut microbiome using probabilistic topic modeling. Each microbial sample is considered as a "document", which is comprised of "functional groups" (which are thought as latent topics). The "functional elements" are considered as the "words" of the document. Thus a topic modeling uncovers functional groups in each sample.

Wearable wireless sensor devices track low-level physical activities (such as walking, sitting etc.) and high-level physical activities are comprised of these low-level activities. Kim et el. (2011) introduced the idea of topic modeling as a means of finding 'latent topics', considered as the high-level physical activities, from these low-level activities.

Bisgin et el. (2011) used LDA topic modeling on the drug labels(i.e., Boxed Warning, Warnings and Precautions, Adverse Reactions) to generate 100 topics, each associated with a set of drugs grouped together based on the probability analysis. Each topic (drug groups) was linked to specific adverse events or therapeutic application. Potential adverse effect of drugs could be identified from the topics.

With topic modeling, Hisano et el. (2013) showed how news affect the stock market activity. Abnormal market activity can be explained by the flow of news and thus help in estimating trading. Words that represents the topic distributions extracts important pieces of information that influence stock market.

Hong and Davison (2010) studied the effectiveness of topic modeling on microblogging texts, namely twitter. They investigated whether the character limit put by twitter (140 characters) affects the traditional topic models. Hong and Davison took two problems to look at: predicting popular twitter messages and classifying twitter users and corresponding messages into topical categories. They found that aggregating short twitter messages by users and training on them yields better classification results.

5 Conclusion

A survey on topic modeling is presented in order to emphasize the growing number of research to discover the latent topics in a text corpora. Motivation of this survey was to provide an introductory overview of the most popular topic models and hence, to encourage new research alleys regarding topic modeling. The growing number of applications of topic modeling in various fields of study hints that this is extremely useful and more rigorous research opportunities are possible to uncover new potentials. For example, visualization of the topics and user-interacting interface could bring more utilities out of topic modeling.

References

Bisgin H., Liu Z., Kelly R., Fang H., Xu X., Tong W. 2012. Investigating drug repositioning opportunities in FDA drug labels through topic modeling. BMC Bioinformatics. 2012; 13(Suppl 15): S6.

Bisgin H, Liu Z, Fang H, Xu X, Tong W. 2011. Mining FDA drug labels using an unsupervised learning technique--topic modeling. BMC Bioinformatics. 2011 Oct 18;12 Suppl 10:S11.

Blei, D., Ng, A., and Jordan, M. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022.

Blei M. David. 2012. Probabilistic Topic Models. Communications of the ACM, Vol. 55 No. 4, Pages 77-84.

Blei, D., McAuliffe, J. 2007. Supervised topic models. In Neural Information Processing Systems (2007).

Blei, D., Lafferty, J. 2006. Dynamic topic models. In International Conference on Machine Learning (2006), ACM, New York, NY, USA, 113–120.

Blei, D., Gri, T., Jordan, M., and Tenenbaum, J. 2003. Hierarchical topic models and the nested chinese restaurant process. Seventeenth Annual Conference on Neural Information Processing Systems (NIPS 2003).

Blei, D., Lafferty, J. 2007. A correlated topic model of Science. Ann. Appl. Stat., 1, 1 (2007), 17–35.

Chen X, He T, Hu X, Zhou Y, An Y, Wu X. 2012. Estimating functional groups in human gut microbiome with probabilistic topic models. IEEE Trans Nanobioscience. 2012 Sep;11(3):203-15.

Daud, A., Li, J., Zhou, L., Muhammad, F. 2010. Frontiers of Computer Science in China. June 2010, Volume 4, Issue 2, pp 280-301

Hisano R, Sornette D, Mizuno T, Ohnishi T, Watanabe T. 2013. High quality topic extraction from business news explains abnormal financial market volatility. PLoS One. 2013 Jun 6;8(6):e64846.

Hong, L., Davison, B. 2010. Empirical study of topic modeling in Twitter. In Proceedings of the First Workshop on Social Media Analytics (SOMA '10). ACM, New York, NY, USA, 80-88.

Jelisavcic, V., Furlan, B., Protic, J., Milutinovic, V. 2012. Topic models and advanced algorithms for profiling of knowledge in scientific papers. MIPRO, 2012 Proceedings of the 35th International Convention, vol., no., pp.1030, 1035, 21-25 May 2012.

Kim S, Li M, Lee S, Mitra U, Emken A, Spruijt-Metz D, Annavaram M, Narayanan S. 2011. Modeling high-level descriptions of real-life physical activities using latent topic modeling of multimodal sensor signals. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:6033-6.

Lehman, LW., Saeed M., Long W., Lee J., Mark R. 2012. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. AMIA Annu Symp Proc. 2012;2012:505-11. Epub 2012 Nov 3.

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

Rosen-Zvi, M., Griffiths, T., Steyvers, M, and Smyth, P. 2004. 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.

Wallach, H. 2006. Topic modeling: Beyond bag of words. In Proceedings of the 23rd International Conference on Machine Learning (2006).