Descriptive Naming & Summarization of large text using Topic Model-A Survey

Saumya Tripathi
tripathi.sam01@gmail.com
GLAU, MATHURA`

ABSTRACT

Due to the plethora of documents containing large scale of text that are available on web it sometimes gets difficult to go through each document to get the clear picture of what the text is depicting. In this paper, we are analyzing several techniques to evaluate Topic Model. A Topic Model is a very popular approach for representing and smoothing the content of documents. Here we will focus on uncovering the thematic structure of a corpus of document that will help in document classification and for compact document topic representation. We have gone through some of the famous topic model such as-Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (PLSI), Latent Dirichlet Allocation (LDA), Pachinko Allocation Model (PAM) where we encounter few issues such as Topic models are not proper for some SNS such as micro blogging and supervise learning techniques are designed for one-labeled corpus i.e. they are limiting the document to a single label.

1. INTRODUCTION

Document Classification is to automatically sort documents into sets of classes and categories. This task is supervised learning means it is useful for assigning tags and labels to documents. There are many methods for classifying text that are producing satisfactory results. There are few popular text classifiers used for text classification such as Support Vector Machine[5], Naïve Bayes Classifiers[8], Rocchio Classifier[10], Nearest Neighbors Classifier[7] from which SVM serves as a better classifier as observed by past researchers. Technique known as ‘Bag of Words’ which is used by all the above classifier for representing the text documents where document is the basic unit for text classification. ‘Bag of Word’ model consists of unordered collection of words. A global dictionary represent documents whereas number of words are represented by dimension of vector in the global dictionary. ‘Bag of Word’ model is an efficient way for the representation of documents. But it has a limitation that it is not able to discriminate the semantic meaning of two or many more
words that may form a phrase. To cope up with this limitation we need to include such n grams in the vector space and not add any additional changes in the classifier based on vector space models.

A phrase is a combination of words in order. It is observed when a word is combined with other word the semantic meaning of the word changes. For example, if we take a word ‘stream’ which is combined with other word ‘data’ then the semantic meaning of the combined word is changed. If the word ‘stream’ is used with ‘river stream’ that it means the flow of water whereas the former represent the flow of data. We can take one more example like ‘data mining’ and ‘gold mining’ means extraction of gold from mine and the former means mining of data from data warehouse. These are some of the example that contains a common word that depicts totally different meaning when combined with any other word.

To find phrases from the other document Latent Dirichlet Allocation(LDA)[12] is used which is able to find phrases from the document. But we are aware that over the time new data set with new entities and new structure are created which LDA cannot handle. It is a supervised model and it cannot take any information from itself or hierarchical labels. So, to deal with this limitation we use Hierarchical Latent Dirichlet Allocation(HLDA)[2].

SSHLDA(Semi Supervised hierarchical Latent Dirichlet Allocation) popularly known as Semi Supervised Hierarchical Topic Model[25], is very efficient in finding latent topics from the documents and take the information from the labels that are hierarchical to build corresponding topics. We use the hierarchical labels together and use the phrases with words in Vector Space Model. Studies show that ‘Bag of Phrase’ model has better performance that ‘Bag of Words’ model. In the next section we will discuss how other techniques are used to find better descriptive name and how it helps in summarizing the document.

2. PHRASE EXTRACTION

Most of the models use “bag of word” models to classify the documents but the inherent flaw in these models are the fact that they do not consider the discriminative power of words when used in combination. These models also lag because of the fact that qualitative meaning of the words is not accurately understood during the representation of the document. Therefore, to resolve this issue, Zhang[28] came up with the usage of multi-words to classify the documents instead of individual words.
An algorithm was developed that extracted multi words using the lexical tools and it helped in achieving an acceptable accuracy for classification. This further establishes the relation among the words and uses the domain knowledge to understand the meaning of individual words to classify the documents. Hotho et. al. used this concept that helped to combine the background knowledge with the relation among different words and classified the documents.

Another technique that is used commonly is linguistic unit enhanced representation. It employs to the usage of lexical and syntactic rules for phrases to retrieve the noun phrases, terminologies and set of entities from the documents to enhance the representation. Lewis[9] also compared word based and phrase based index mapping for representing tasks in document classification. But it refrained from giving better results as phrase indexing was not improving categorization because more phrases resulted in low frequencies. Other researcher also tried to improve the effectiveness of text retrieval by using multi-words. There is another method word sequence disambiguation which treats words as strings and it ignores the semantic in the document. Text is represented using this method by using words group using co-occurrences of similar word sequence extracted from a document. Li[3] used the technique of generalized suffix tree to extract frequent words to propose the CFWS clustering algorithm.

A simple scheme is given by V.Kumar [15] for extracting phrases which is Latent Dirichlet Allocation in Topic Model. It uses phrases from the vector space in combination with words to represent the document. The technique ‘bag of words’ captures the discriminative power of similar linguistic words as an ordered pair. The main aim of researchers is to establish classifiers and deliver accuracy with better results. It gives the better results but cannot improve the relation between super topics and sub topics. It is unsupervised method also so it is unable to take any information from labels in hierarchy.

3. TOPIC MODEL

Topic model is categorized into four main types:

- Unsupervised hierarchical Topic Model
- Unsupervised non-hierarchical Topic Model
- Supervised hierarchical Topic Models
- Supervised non-hierarchical Topic Model
3.1 Flat Topic Model

3.1.1 Latent Semantic Analysis

LSA[1] which is the first topic model is studied on large scale which falls under the category of unsupervised non-hierarchical topic models which put light on automatic indexing and retrieval of text. The problem that hinders existing techniques of retrieval which try to match queries with words is discharged by them. These techniques finds the linguistic structure to get rid of Noise.

Decomposition using single values is done by Latent Semantic Indexing. In this technique a 2 D Array is constructed between term and document from which the documents that are more closely associated are kept together. Terms that are not similar to the document may not be added to the array.

3.1.2 Probabilistic Latent Semantic Analysis

LSA is applied for various domains to get a positive response in many other domains like automatic indexing but it has some limitations instead Probabilistic Latent Semantic Analysis (PLSA) [13] has firm foundations as it gives a proper generative model for data. It means that standard techniques can be applied for questionnaire like combination of model, control for complexity and model fitting.

PLSA is a great technique to distinguish between different types of words having different meaning and it deals with polysemes using its factor representation. But due to its number of parameter over fitting problem occurred.

3.1.3 Latent Dirichlet Allocation

The limitation of PLSA i.e overfitting is removed using Latent Dirichlet Allocation LDA[12] therefore to compensate the mixture of models that capture exchangeability of both words and documents for exchangeable representations of documents and words. This logical thinking promoted LDA model.

Bag of words model made it possible to mixer distributions for words that are single and helps to achieve rich model which have large structural units. Every unit is composed of distribution of probability over topics, it is known as topic modeling. There are variety of frame work available for developing syntactic framework of topic model.
3.1.4 **Correlated Topic Model**

The focus is on the restriction of the model that is nominated to date Correlated Topic Model [33] do not succeed to model the correlation between the topics. The correlation between latent topics and their subsets is normal. For example, an article on genes may be related to health and disease, but it may also be related to ultrasound and x Ray.

The use of one topic is not correlated if another topic is also present and the components of vector are independent to each other. This method leads to automatically assume unreal assumption. So the CTC is more flexible to topic proportions.

To model many topics per document unsupervised topic models are not sufficient. This is because they are not efficient enough to use the labels being observed in their learning. They have issued many modifications of LDA in their writings.

3.1.5 **Supervised Latent Dirichlet Allocation**

To deal with the shortcomings of CTC Supervised Latent Dirichlet Allocation [9SLDA0] [14] is used. In this the documents uses the same number of topics but each topic is having the mix of topics that are not similar.

This model provides document to response pair. to predict parameter and estimate algorithms. This technique is used to handle diversified response types. It helps in predicting the response variable that will help in the future to find unlabeled documents.

3.1.6 **Labeled Latent Dirichlet Allocation**

For multiple labeled corpus L-LDA ia used because it helps in supervising modern texts to word assignments. In this each label is representing only one topic and gives indirect response as compared to standard LDA. L-LDA is an extension for the LDA. It is preferred over LDA because it deals efficiently to multiple labeled documents. It efficiently uses Support Vector Machine and out performs it. In this each topic is driven from the word.

3.1.7 **Author Topic Model**
Probabilistic topic models uses authorship information in Author Topic Model[17]. It uses linguistics, authorship attributions and it focuses on joint author problems. It highlights the different problems dealing with the particular author whiting a piece of text.

Graph based and network based models are generally used to represent and analyses relations between authors as a base. Many authors like Erten[2003] McCann[16] used many techniques based on the bibliometrics, graphs and social networking. It analyzed co-authors and citations based on each ones visualization in the literature. White&Smyth [34] for analyzing co-author graphs used page Rank style ranking.

### 3.1.8 Prior Latent Dirichlet Allocation

For multi label documents to classify large number of labels in documents Prior LDA [27] is used. There are three models available using LDA framework. There is Flat model and other two models. The second one is Prior label, it is used when frequency in label is occurred and the third one is dependency label which is used when there is dependency between the labels. All the three models are compared and tested.

This model improves the performance other than any other model. This model is useful for large scale having multi label problems. This model is better than SVM.

### 3.1.9 Partially Latent Dirichlet Allocation

The challenges that occurred in partially supervised text mining the model Partially LDA(PLDA)[18] made use of unsupervised learning. But, it used labels provided by humans. There is a similarity between PLDA and LDA is that PLDA uses words from specific mix of latent topics where the topic is represented as the distribution over the words.

### 3.2 Hierarchical Topic Model

There are complex models that are probabilistic which are frequently increasing with high rate , e.g. vision, IR, biomedical and bioinformatics. The present tools for statistical modeling are unvarying in this concern mainly old school techniques that help in the continuing flow of data to unbounded sets. There are many models to solve this drawback.
3.2.1 Hierarchical Pachinko Allocation Model

An unusual member was added to the PAM and used the task of hierarchical topic modeling. Each internal note in the DAG is used in HPAM[22] hierarchical PAM. This model is good when there is no limit to a single hierarchy. This model can be better understood through example: suppose there are three fields of computer science like Computer Vision, Artificial Intelligence and Machine Learning all the three are associated to Artificial Intelligence.

3.2.2 Hierarchical Dirichlet Processes

Hierarchical model technique is used in Hierarchical Dirichlet Process (HDP) [26]. This tool is very efficaciously for Bayesian Classification. This type of model are very useful to deal with general multi-tasking learning and to learn the learning and it also helps in building the notion that is helpful in random learning.

The most common example in this context is the problem of Gaussian means, here we will use K-mean and Gaussian mean and all the data will be used using Gaussian distribution.

3.2.3 Hierarchical Concept Topic Model

Here the author is proposing a general framework for combining data driven topics and semantics concepts. Hierarchical Concept Topic Model[32] takes the best feature from other two approaches. To get the best result one needs to extend the framework to hierarchical concept Topic Model. It is very obvious that how documents can be retrieved from the corpus and how ontologies can be matched. Remarkable work is done on how to use data set to create ontology dictionary i.e., how ontologies can be matched to set of data and how they will use it for methodologies or data set. Broad area is proposed where we can use general purpose probabilistic model to combine concepts and topics in a single framework.

3.2.4 Hierarchical Latent Dirichlet Allocation

Directed Acyclic Graph[DAG] is mostly used which is used to construct relations in hierarchy. Blei 920040 has proposed a model HLDA, this model learns simultaneously both the hierarchy. We can extract topic hierarchy from large collection of documents by using this algorithm. Few more
models are proposed like HLLDA[24] and HSLDA[23] that can deal with the relations from labels in data hierarchy.

### 3.2.5 Hierarchical Supervised Latent Dirichlet Allocation

Here the main focus was on unstructured text that is categorized manually. There are examples but not necessarily related to web pages and are curated in hierarchical directories where the data is arranged in the order of ranking. Here the author has suggested that the labels might not be inaccurate. So one must improve similarities between the documents for the purpose of Information Retrieval. He also suggested how to combine information from two sources to create a single model that will categorize new documents automatically.

### 3.2.7 Semi Supervised Hierarchical Latent Dirichlet Allocation

HLLDA is proposed by the author to handle the problems of HLLDA. To cover up the problems of HLLDA, SSHLDA[25] is proposed in which latent topics and observed topics are proposed. It can explore latent topics in data space and can extend the hierarchy of observed topics. We use SSHLDA for phrase extraction. This is a better technique for phrase extraction. It used phrases from the ordered pair of words

### 4. CLASSIFICATION

It has been observed since last few years that for efficient classification and for access to relevant data a content based management of data is required. Classification of documents is done by assigning predefined categories to them. It helps in providing conceptual view and has an important application in real world. Generally, nowadays news item are classified according to geographical codes and technical papers on the basis of their technical domain.

Document classification combines Machine Learning and Information Retrieval. Naive Bayes Classifier[8], Rocchio[10], Nearest Neighbor Classifier[7], Support Vector Machine (SVM)[5] are popular text classifiers.

#### 4.1 Rocchio Classifier

Uden [10] gave the better understanding of the word weight. It is going to process the topics that are non-relevant if the contents of the document match with the query of the users an dis done for
limited period of time so the class vector will be optimized. It has got some limitations like it is not
efficient if there are more number of documents.

4.2 Naïve Bayes Classifier

E. Keogh [8], state that naive Bayes is one of the quickest ways to classify data. It discards irrelevant
information and considers real time data and streaming data efficiently, producing great results.

4.3 Support Vector Machine

C.j.c.Burgs et. al. [5], came forward with a view that SVM is also an efficient way of pattern
recognition. Since it uses simple geometry, therefore a lot of scope is left for further research. SVM
is a supervised learning model that uses the regression to analyse the data.

4.4 Nearest Neighbor Classifier

V.Hautamaki et. al. [7], suggested that nearest neighbor considers pattern recognition as a major tool
for classification of data. It discards redundant category and comparisons.

5. CONCLUSION

This paper lists out various techniques used for phrase extraction along with their advantages and
disadvantages. The major problem in multi word extraction is the limitation on word limit. If the
query exceeds the word limit, then no satisfactory results are provided. To solve this issue, various
researchers have come forward with their techniques and methodologies. Each one of them has
achieved success to a large extent. This survey paper tells about what are the major techniques used
till now for phrase extraction along with their flaws.

6. REFERENCES

[1] Thomas K Landauer, Peter W. Foltz & Darrell Laham,” Latent semantic analysis: An
introduction to latent semantic analysis,” Quantitative Approaches to Semantic Knowledge
Representations, - Issue 2-3: Volume 25, 1998.
[2] Chang, Jonathan, and David M. Blei. "Hierarchical relational models for document networks." The Annals of Applied Statistics 4, no. 1, pp. 124-150, 2010.

[3] Li, Wei, David Blei, and Andrew McCallum. "Nonparametric bayes pachinko allocation." arXiv preprint arXiv, pp. 1206-5270, 2012.

[4] Diederich, Joachim. "Rule extraction from support vector machines: An introduction." Rule extraction from support vector machines. Springer, Berlin, Heidelberg, pp. 3-31, 2008.

[5] Byun, Hyeran, and Seong-Whan Lee. "Applications of support vector machines for pattern recognition: A survey." International Workshop on Support Vector Machines. Springer, Berlin, Heidelberg, 2002.

[6] Andreas Hotho, Robert Jäschke, Christoph Schmitz, Gerd Stumme, “Information Retrieval in Folksonomies: Search and Ranking,” European Semantic Web Conference, pp. 411-426, 2006.

[7] Hautamaki, Ville, Ismo Karkkainen, and Pasi Franti. "Outlier detection using k-nearest neighbour graph." Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Vol. 3. IEEE, 2004.

[8] Keogh, Eamonn J., and Michael J. Pazzani. "Learning the structure of augmented Bayesian classifiers." International Journal on Artificial Intelligence Tools 11.04, pp. 587-601, 2002.

[9] David D. Lewis, “Naive (Bayes) at forty: The independence assumption in information retrieval,” European Conference on Machine Learning, pp. 4-15, 1998.

[10] Van Uden, Mark. "Rocchio: Relevance feedback in learning classification algorithms." Proceedings of the ACM SIGIR Conference. 1998.

[11] Kjell, Bradley. "Authorship determination using letter pair frequency features with neural network classifiers." Literary and Linguistic Computing 9.2, pp. 119-124, 1994.

[12] Blei, D.M., Ng, A.Y. and Jordan, M.I., “Latent dirichlet allocation,” Journal of machine Learning research, 3(Jan), pp. 993-1022, 2003.

[13] Hofmann, Thomas. "Probabilistic latent semantic analysis." arXiv preprint arXiv, pp. 1301.6705, 2013.
[14] Li, X., Ma, Z., Peng, P., Guo, X., Huang, F., Wang, X. and Guo, J., "Supervised latent Dirichlet allocation with a mixture of sparse softmax," Neurocomputing, 312, pp. 324-335, 2018.

[15] Kumar, Ravi, and K. Raghuvbeer. "Legal document summarization using latent dirichlet allocation." Int J Comput Sci Telecommun 3, pp. 114-117, 2012.

[16] McCann, Donald W. "A neural network short-term forecast of significant thunderstorms." Weather and Forecasting 7.3, pp. 525-534, 1992.

[17] Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P., "The author-topic model for authors and documents," arXiv preprint arXiv, pp. 1207.4169, 2012.

[18] Ramage, D., Manning, C.D. and Dumais, S., "Partially labeled topic models for interpretable text mining," In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 457-465, 2011.

[19] Zhang, H., Chen, B., Cong, Y., Guo, D., Liu, H. and Zhou, M., 2020. Deep autoencoding topic model with scalable hybrid Bayesian inference. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[20] Mutschke, P., "Enhancing information retrieval in federated bibliographic data sources using author network based stratagems," In International Conference on Theory and Practice of Digital Libraries (pp. 287-299). Springer, Berlin, Heidelberg, 2001.

[21] Erten, Y.M. and Tomur, E., "A layered security architecture for corporate 802.11 wireless networks," In 2004 Symposium on Wireless Telecommunications, pp. 123-128, IEEE, 2004.

[22] Mimno, D., Li, W. and McCallum, A., "Mixtures of hierarchical topics with pachinko allocation," In Proceedings of the 24th international conference on Machine learning, pp. 633-640, 2007.

[23] Perotte, Adler, et al. "Hierarchically supervised latent Dirichlet allocation." Advances in neural information processing systems 24, pp 2609-2617, 2011.
[24] Mao, Xian-Ling, et al. "Ehlda: a supervised hierarchical topic model." Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, Cham, pp. 215-226, 2015.

[25] Mao, Xian-Ling, et al. "SSHLDA: a semi-supervised hierarchical topic model." Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, 2012.

[26] Ren, Lu, David B. Dunson, and Lawrence Carin. "The dynamic hierarchical Dirichlet process." Proceedings of the 25th international conference on Machine learning, pp. 824-831, 2008.

[27] Liu, B., Zhang, P., Lu, T. and Gu, N., 2020. A reliable cross-site user generated content modeling method based on topic model. Knowledge-Based Systems, 209, p.106435.

[28] Wen Zhang, Taketoshi Yoshida, Xijin Tang, "Text classification using multi-word features," IEEE International Conference on Systems, Man and Cybernetics, 2007.

[29] Mao, Xian-Ling, et al. "SSHLDA: a semi-supervised hierarchical topic model." Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, 2012.

[30] Qian, X.D., “Fast latent semantic index using random mapping in text processing,’ In 2008 International Conference on Wavelet Analysis and Pattern Recognition (Vol. 2, pp. 788-792). IEEE, 2008.

[31] Wen Zhang, Taketoshi Yoshida, Xijin Tang, “Text classification using multi-word features,” IEEE International Conference on Systems, Man and Cybernetics, 2007.

[32] Bazan, Jan G. "Hierarchical classifiers for complex spatio-temporal concepts." Transactions on rough sets IX. Springer, Berlin, Heidelberg, pp.474-750, 2008.

[33] Blei, D.M. and Lafferty, J.D., “A correlated topic model of science,” The Annals of Applied Statistics, 1(1), pp.17-35, 2007.
[34] Stephen A White, P. Smyth, “Algorithms for estimating relative importance in networks,” KDD ’03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 266–275, August 2003.

[35] Hu, N., Zhang, T., Gao, B. and Bose, I., 2019. What do hotel customers complain about? Text analysis using structural topic model. Tourism Management, 72, pp.417-426.

[36] Newman, Mark EJ. "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality." Physical review E 64.1 ,016132,2001.

[37] Rortais, A., Barrucci, F., Ercolano, V., Linge, J., Christodoulidou, A., Cravedi, J.P., Garcia-Matas, R., Saegerman, C. and Svečnjak, L., 2020. A topic model approach to identify and track emerging risks from beeswax adulteration in the media. Food Control, 119, p.107435.

[38] Wallach, H., Mimno, D. and McCallum, A., “Rethinking LDA: Why priors matter,” Advances in neural information processing systems, 22, pp.1973-1981,2009.