Topic Modelling

Shanthala Nagaraja, Kiran Yarehalli Chandrappa

Abstract*: *In the history of information and technology the knowledge which was generated is stored in the form of digital technology. In present day the search engines will search based on terms and extract the list of similar documents from many topics. In this paper, the proposed Topic Modelling techniques will search based on the group of words from each document. The aim behind proposed topic modelling techniques is to comprise the topics from each of the document. The hidden topics from the list of collected text documents can be extracted using proposed probabilistic topic modelling.

I. INTRODUCTION

To identify words from various documents and link all the documents which uses same pattern. The idea behind topic modelling is to work with the documents with mixture of topics where they use probability distribution over words. The documents are generated using probabilistic approach. Based on distribution each word from the document can be taken as a topic and generate new document by considering the distribution over topic. As per Seungil and Stephen, the words which are extracted can be represented as a histogram. The histogram is formed based on the words in a vocabulary and distribution over certain number of topics. By learning the distribution, a high dimensional histogram with least significance representation can be formed for each document. The various topic models, such as Probabilistic latent semantic analysis (PLSA), Latent semantic analysis (LSA), Correlated topic model (CTM), Latent Dirichlet allocation (LDA) have improved their accuracy in classifying the area of topic modelling.

[1] Initially to mine the topics from text corpora “Topic Modelling” techniques were used. These techniques reveal the hidden thematic structure during an assortment of documents and facilitate to make up new ways in which to browse, search and summarize massive archive of texts. a subject may be a cluster of words that regularly occur along. A topic modelling will connect words with similar meanings and create a distinction between uses of words with many meanings. Here we tend to gift a survey on journey of topic modelling techniques comprising Latent Dirichlet Allocation (LDA) and Non-LDA primarily based techniques and therefore the reason for classify the techniques into LDA and Non-LDA is that LDA has dominated the subject modelling techniques since its origin. We’ve used the three stratified classification criteria for classifying topic models that embody LDA and Non-LDA primarily based, bag-of-words or sequence-of-words approach and unattended or super- vised learning for our survey. Purpose of this survey is to explore the subject modelling techniques since Singular Value Decomposition (SVD) topic model to the newest topic models in deep learning. Also give the transient outline of current probabilistic topic models in addition as a motivation for future analysis.

[2] Two topic modelling algorithms are explained in this paper which are namely LSI & SVD and LDA. To find the statistical relationship among documents to genre ate topic ontology and ontology graph with less manual efforts. This kind of experimental analysis helps to showcase the effectiveness of proposed ontology for the user quires on topic modelling.

[3] The automatic generation of description for a picture is known as Image annotation, it is an important element in many search engines which are based on image to retrieve the applications based on image. It is costly and consumes time to generating an image database since the keywords which are used should be hand- coded. The images along with the occurring of text data that are produced by the mixture of latent topics are explained with the probabilistic model.

[4] This paper presents the novel task of best topic word selection, as a means of enhancing the interpretation and visual image of topic models. We propose variety of options meant to capture the most effective topic word, and show that, in combination as inputs to are ranking model, we are able to consistently achieve results on top of the baseline of merely choosing the highest-ranked topic word. This is the case each once coaching in-domain over different tagged topics for that topic model, and cross-domain, exploitation solely labeling from freelance topic models learned over document collection from different domains and genres.

II. TECHNIQUES FOR TOPIC MODELLING

To extract the topics from the collection of documents and link that document which have similar topic. The techniques for topic modelling are mentioned as below.

A. Unsupervised learning techniques with LDA based bag of words topic models

1) Latent Dirichlet Allocation (LDA): LDA becomes foundation for numerous latent factors discovery algorithms which was collectively known as Probabilistic topic models. This model upgrades into Bayesian graphical model by introducing priors on document-topic distributions than pLSI. Dirichlet prior distribution can be introduced in LDA which reduces the number of estimated parameters in pLSI which is represented in the graphical model for LDA.

Revised Manuscript Received on December 22, 2019.

Shanthala Nagaraja, Pursuing Ph. D BNM Institute of Technology, College of Engineering, Bangalore, India.

Dr. Kiran Y. C, Professor, BNM Institute of Technology, College of Engineering, Bangalore, India.

2) Hierarchical Latent Dirichlet Allocation (HLDA): HLDA is an extension for LDA model that uses topic of tree rather than flat topic. A non-parametric Bayesian model is used by HLDA for identifying the repeated topics or words. Every node in a tree is linked as a topic whereas topic is scattering of words.

3) Author-Topic Model (ATM): ATM is a first generative model for probabilistic approach and an expansion for LDA. Metadata which is present in each document is used for extracting the topic. The two variables which are connected with each words as an author and topic. The learning of author-topic and topic-word distribution can be achieved with Markov chain Monte Carlo (MCMC).

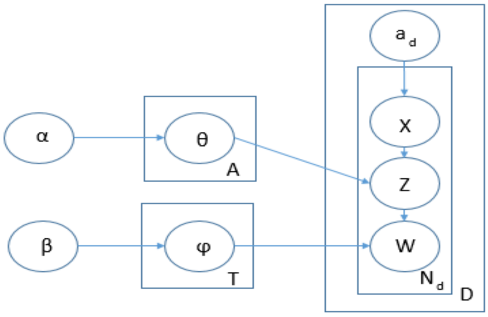


Fig. 1. Graphic model for ATM

4) Dynamic Topic Model (DTM): This method is used for detecting a topic in a sequentially organized documents based on evolution of. This uses a sequential distribution of words instead of single distribution. To overcome inferences which have come from sampling methods and non- conjugacy, the best option is to use Variational Wavelet Regression or Variational Kalman filtering.

5) Correlated Topic Model (CorrTM): To overcome in-ability from LDA, CorrTM. This model is able to produce complex structure of topics and generates a covariance matrix which can be used to form a topic graph. The factorized distribution of a topics can be formed using Mean variational algorithm. This model is very much expressive compared to.

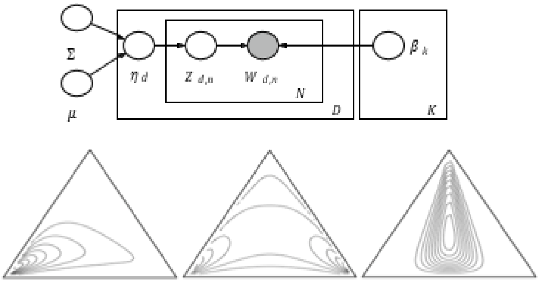


Fig. 2. Graphic model for CorrTM

6) Bigram LDA Topic Model (BLTM): To construct the model based on N-gram model which helps to come over from bag-of-word approach. By measuring the previous words probability, the next word can be predicted. By explaining the distribution of words over content and topic BLTM expands LDA models.

B. Supervised learning LDA based bag of words topic model

1) Supervised Latent Dirichlet allocation (SLDA): In SLDA, a response variable has added for each document. In order to find latent topics that would best predict the response variables for future unlabeled documents by jointly model the responses and documents. To estimate the parameters which are unknown it is preferred to use Mean field variation inference approach along with EM algorithm.

2) Dirichlet Multinomial Regression (DMR): DMR has been trained using Expectation–Maximization (EM) sampling algorithm. With no additional coding DMR is able to incorporate arbitrary types like discrete, categorical and continuous features.

3) Supervised Citation Network Topic Model (SCNTM): It produces a vectors using Griffiths-Engen-McCloskey (GEM) distribution based on probability whereas the base as Pitman-Yor process (PYP). The author information from each document is used for supervised learning process, Markov chain Monte Carlo (MCMC) is used as one of the learning process for this kind of models.

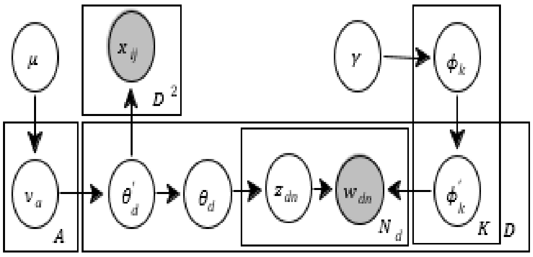


Fig. 3. Graphic model for SCNTM

C. Unsupervised learning LDA based sequence of words topic models

1) Sequential Latent Dirichlet Allocation (seqLDA): It is introduced based on novel variant for LDA based on sequential structure. It is again divided in to multiple segments and each segment is linked with following segments. To bind the sequence of LDA models first order Markov chain is very useful. Gibbs sampling algorithm was used for parameter based estimation of Poisson Dirichlet process(HPDP).

D. Unsupervised learning for non-LDA based bag of words topic models

1) Latent Semantic Analysis (LSA):

To extract the hidden meaning of text LSA can be useful method. LSA generates a document with matrices which will have most frequent words occurring in the document. By applying a Single Value Decomposition (SVD) method a matrix is further divided into group of three matrices which are V, U and S whereas:

- U Indicates the term eigenvectors matrix,
- S Indicates the diagonal matrix of singular values,
- V Indicates the document eigenvectors matrix.

III. PROPOSED METHODOLOGY

A. LSI & SVD based Semi-Automatic Ontology Learning

Two approaches are examined for semi-automatic ontology learning using LSI & SVD and LDA in this paper. Domain specific ontology is used to retrieve the information. Since it is tedious to construct an ontology, so it is better to make the process automatic. The documents must be read and processed to make the domain specific ontology process automatic.

Where the set of words is considered as a document. LSI is known as vector space approach. To replace document spaces with lower dimensional concept, space an LSI statistical reference is used. LSI uses singular value decomposition(SVD) for replacing the document spaces. To collect the frequency of meaningful terms we need to remove unimportant words like (the, is, a), it is called as pre-processing of text. The next step is to normalize by measuring the weight of each term.

B. LDA based Semi-Automatic Ontology Learning

This process deals with a constructive way of developing ontology which helps to overcome the hindrances produced by old collection. The next latent topic is extracted after completion of correcting the generated raw test data. Since it is easy to calculate the range of topics in LDA, to identify the latent topics from the text corpus which is generated from the previous sequence when compared to the pLSI and LSI. It will become a choice whether to further identify a specific filed or not since the topic has been identified already therefore it saves time. To measure the cold and hot topics (θj) over time (years) LDA can significantly be used. To analyze the topic dynamics per year one must know the words assigned to the topic. To collect the documents from mails, internet and newsgroups LDA can be used significantly.

The objective is to use data mining models for analyzing the words which are obtained from emails and other SNA (Social Network Analysis). To connect the hidden words, it is advisable to use above mentioned tools in parallel. LDA models have higher precision than TF-IDF models. To understand a problem in topic from real world the mentioned two approaches are used one is “Weighted edge graph model” and “Directed graph models”. LDA uses Bayesian model to reduce complexity. To classify the topics and words by detecting the fraud emails helps us for security purpose which can be achieved using LDA models. LDA model along with Map Reduction method will reduce time and cost. Labelling of the documents is not required for LDA models and hence we can all it as unsupervised approach, whereas the extra information like title, author name is used as a metadata. The dimensionality issues can be observed by text corpus using LDA model.

Ontologies are classified in to three types which are formal ontologies, terminological ontologies, prototype-based ontologies. The process of learning ontologies is again divided in to six sub-categories which are learning terms, relations, synonyms, rules, concept hierarchies, concepts.

C. Steps for achieving the Topic Modelling using LDA

Step 1: To segregate the set of verbs, nouns and adjectives which are present in the input data, the input data must be processed and tokenized and also all the words like the, in, at, that, which, and on should be removed and this kind of words are called as stop words.

Step 2: Resulting output from the step 1 is then subjected to LDA modelling algorithm which gives set of words as output. The output contains the set of words which are related to each other and are highlighted as different topics.

Step 3: To assign the words which falls under specific topic and to create an ontology using LDA topic modelling a tool called Protégé is used. Since LDA can’t deal with CTM so one topic can’t be connected to other topic directly.

Step 4: To retrieve the results for user queries, one need to process the quire based on entered user input and also based on its type. SPARQL query is used to detect the word and topic.

D. Assumption of LDA

Based on the statistics obtained from the corpus the assumptions for LDA are listed. The “Bag-of-Words (BoW)” assumption as below:

- LDA does not care about the order of documents.
- LDA assumes that count of topics is known and constant.
- LDA assumes that the topic should be distributed based on vocabulary.

IV. RESULTS AND CONCLUSIONS

One of the blocker for LDA is it can’t be used to model the similarities among topics but only capture the similarities among words which is due to Dirichlet distribution. Since the topic correlation is common in the real world one can ignore the topic correlation and limit the ability of LDA there by one can predict the new data with high likelihood. This issue can be solved using Pachinko Allocation Model (PAM), which gives a correlation among topics by using Directed Acyclic Graph (DAG).

In this paper we proposed two classification of topic modelling under text mining. The four topic modelling approaches has been discussed under the first category which are Probabilistic latent semantic analysis (PLSA), Latent Dirichlet allocation (LDA), Latent semantic analysis (LSA), and Correlated topic model (CTM). The difference between the above mentioned four methods is explained in terms of theoretical results, characteristics and limitations. In this paper we described the high level approach of topic modelling in text mining, and also the applications based on the above mentioned techniques are also described. Out of the above mentioned four techniques we have also discussed the disadvantage of old method and the advantages of new method over the old one.

The discussions based on topic evolution model will fall in second category, which model topics by taking time in to consideration. Many papers use multiple methods to model a topic.

Retrieval Number: B11241292S19/2019©BEIESP
DOI: 10.35940/ijitee.B1124.1292S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

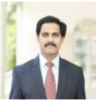
In that some of them uses discretized time method and continuous time method to model a topic, and apart from those two few of them uses employ citation relationship along with the discretized time model, therefore time is considered as an important factor to model a topic.

REFERENCES

1. Bijendra Kumar1, Satish Chand2, Deepak Sharma1, “A Survey on Journey of Topic Modeling Techniques from SVD to Deep Learning”, Department of Computer Engineering, Netaji Subash Institute of Technology, Sector-3, Dwarka, New Delhi, 110078, India.
2. Khalid Alfalqi, Rubayyi Alghamdi, “A Survey of Topic Modeling in Text Mining”, Information Systems Security CIISE, Concordia University Montreal, Quebec, Canada.H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
3. Monika Rani*, Amit Kumar Dhar and O. P. Vyas, “Semi-Automatic Ter- minology Ontology Learning Based on Topic Modelin”, Department of Information Technology, Indian Institute of Information Technology, Allahabad, India.
4. Best Topic Word Selection for Topic Labelling”, JeyHanLau, David- Newman, SarvnazKarimi and TimothyBaldwin, NICTA Victoria Re- Search Laboratory, Dept of Computer Science and Software Engineer- ing, University of Melbourne, Dept of Computer Science, University of California.

AUTHORS PROFILE

Shanthala Nagaraja, completed my B.E from SJMIT, Chitradurga, India, in the year of 2000, M. S from BITS Pilani, India, in the year 2006, and currently working in Toshiba Software India Pvt Ltd as a Senior program manager also pursuing Ph. D in BNMIT college of Engineering, Bangalore, India. My area of interests are Machine Learning techniques and Text mining.



Dr. Kiran Y. C, Completed my B.E from Adichunchanagri Institute of Technology, Chikmagalur, India in the year of 1997, M. Tech from Sri Jayachamarajendra College of Engineering, Mysore, India in the year of 2003 and Ph. D from Jain University, Bangalore, India in the year of 2015 and currently working as a Professor in BNMIT college of Engineering, Bangalore. My area of interest is Image processing and pattern recognition and supervising over five Ph.D. scholar’s.