A Performance Evaluation of Correlated and Dynamic Topic Modeling on a QA Dataset

Bindu K.R.[1*], Gowrikrishna G H[2], Latha Parameswaran[3]
Department of Computer Science and Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
[1*]j_bindu@cb.amrita.edu, [2]gowrihari16@gmail.com , [3]p_latha@cb.amrita.edu

***Abstract*—Topic modeling is a set of algorithms which is used to mine the data that is hidden inside a large collection of documents. In this paper we discusses about the Correlated topic modeling and dynamic topic modeling in detail and comparing their performance on a question answer dataset based on autism. Log Likelihood and Perplexity are the measures used for comparing the discussed topic modeling algorithms.**

Keywords— Topic Modeling, Correlated topic modeling, Dynamic topic modeling, perplexity, Log likelihood

1. INTRODUCTION

To handle the upsurge of collection of electronic documents in these days, there is a high need of systematizing, searching, indexing as well as browsing the documents automatically which leads to the necessity of recent tools and techniques. The mutual knowledge getting stored more in digitized form like in the form of images, web pages, news, books makes it very difficult to find out what we are searching for. Hence there is a high need of algorithms or programs that can recognize patterns in a large body of documents thereby extracting the topics from the texts in documents [10,12,13]. Topic Models are algorithms that can discover the topics that mainly spread through the large collection of documents and can further organize the collection of documents. Topic Models can otherwise be explained as a way of generalizing backward from a document collection so as to deduce the topics that might have generated them. The significance of the words in a text is not known by the topic modeling programs. Instead the assumption is that any part of text is comprised of words selected from probable word baskets wherein each basket relates to a topic.

The field of topic modeling consists of two categories wherein the primary category defines the field of methods in topic modeling and in the second category, the topics are modeled by taking time, an important factor into consideration. The first category comprises of Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM) [5]. The subsequent category includes different models like Topic Over Time (TOT), Dynamic Topic Models (DTM) and Multiscale Topic Tomography [5]. The paper compares the performance of Correlated Topic Model and Dynamic Topic Model on an autism dataset on the basis of a measure called perplexity analyzing which topic model operates better on a dataset.

2. RELATED WORK

A wide range of research has been done in the field of topic modeling in addition to Question Answering system since early 90s [9]. In this section, the previously done methods in the field of topic modeling and Question Answering system are being reviewed.

Rubayyi Alghamdi et.al [5] discussed the two main categories that come within the field of topic modeling. While the primary category discusses the topic modeling domain methods, the second category focuses on

modeling topics considering an essential factor namely time. The paper evidently explained the different algorithms together with the characteristics and limitations of each. The overview of topic modeling and comparison of the respective categories are described eventually. Out of the methods discussed in both the topic modeling categories, two algorithms have real time importance when applied to a dataset.

David M. B et.al [4] discussed mainly on the simplest topic modeling algorithm namely Latent Dirichlet Allocation (LDA). The main intuition behind the model LDA is of simulating documents in collection being as rising out of various topics wherein the distribution on top of a predetermined lexicon of terms is described as a topic. The paper further demonstrated on how to make use of the posterior distribution in LDA as an experimenting tool for large corpus of documents. In continuation, it described on how to efficiently estimate the posterior using the mean field variational approaches. The limitations of LDA are pointed out that led to suggestion of the relaxation on two of the inherent assumptions which LDA builds in order to discover maps of interconnected topics and simulate the topics that change over time.

Asli Celikyilmaz et.al [3] described a detailed survey of generative modeling for the task of question answering (QA) to grade candidate passages. The paper explored on the model Latent Dirichlet Allocation (LDA) for achieving ranking scores on the basis of a unique similarity measure used between a candidate passage and a natural language question given by the user. It described the construction of two modes introducing evaluation on the hidden characteristics of the corpus alongside the posed question. The paper also showed enhancement in the performance analysis of a QA ranking system using the new depiction of topical structures on question answer dataset.

David M. B et.al [2] discussed about Dynamic Topic Model (DTM) for analyzing progress of topics within a large pool containing documents on the basis of a factor namely time. The representation of topics is done by state space model. These models are used over the natural parameters in the distribution – multinomial distribution to evaluate the time evolution of subjects. To perform the approximate posterior inferencing on the hidden topics, Variational approximations on the basis of wavelet regression and Kalman filters are developed. The dynamic topic model developed in the paper provides both quantitative and qualitative results for a corpus in sequence. The demonstration of the model DTM is prepared by considering the OCR-ed archives taken from 1880 to 2000.

David M. B and John D. Lafferty [1] discussed about the main limitation of LDA of not being able to model topic correlation, a real time need. This limitation occurs because of the usage of Dirichlet distribution which is used for modeling the variability amongst topic proportions. The paper developed the correlated topic model (CTM), wherein the factor, correlation among proportions of topic is exhibited through logistic normal distribution. For posterior inferencing, the model developed uses mean field variational algorithm. To model the correlation in CTM, logistic normal is used. The use of logistic normal is extended to a model where the description of the hidden composition of subjects associated with every document is achieved. In CTM, the parameter namely covariance matrix in the logistic normal is proposed to represent such correlations. Furthermore, simulating correlation leads to improved predictive distributions and makes the system more realistic while analyzing document collections.

Domain specialized information retrieval in the areas like social science, bio medicine and blogosphere was implemented by Fautsh et.al in [6]. Initially, retrieval procedures that are standard were focused. But later on domain specific problems started taking in to consideration. David M. B [4] discussed about the Probabilistic model to tackle the collection of corpus. Introduction to Information Retrieval by Chistopher D.Manning [7] discussed on the basic knowledge of designing the Question Answering System. It provided an insight of the basic concept of Information Retrieval.[11]

3. PROPOSED ALGORITHM

The previous works in the area of Question Answer system are based on the simplest topic modeling

algorithm, LDA [3]. However, the Dirichlet distribution used in LDA for independent assumptions regarding the topic proportions is not of real importance and the greatest limitation of LDA. But in most of the text collections, it is likely to think that the subclasses of the primary hidden subjects are extremely interrelated. Finding the time evolution of hidden topics in a sequential collection of documents is another important aspect in real time. In this paper the limitations of LDA are considered and consequently two real time important algorithms namely DTM and CTM are implemented in a Question Answer dataset.

The proposed algorithm is to analyze and evaluate the performance of two topic modeling algorithms namely dynamic and correlated topic model on a question answer (QA) dataset. The measures that are used to analyze the performance measures are perplexity and log-likelihood. As both of the models taken into consideration are of real time importance, the dataset used for the analysis of the models is an autism dataset which is also of real time value. In the present world, so many people are becoming victims of the grievous disease, autism. At the same time the people are unaware of the disease, the cause, the precautions and the medications that are available. In such cases, the question answering can provide as better platform in paving a path for those who need to know about the details of the disease from those who are already aware of it.

The proposed system focuses on more domain specific knowledge as it is more available than general world knowledge. A domain specific Question answer dataset is taken where the primary aspect is to identify the compound terms. The system takes an input of a domain specific question answer dataset on which after preprocessing steps, the two topic modeling methods are applied. Out of the topic modeling algorithms known till date, the topic modeling algorithms applied on the dataset are of much relevance in real time.

The basic preprocessing steps are used for parsing the dataset and to create the dictionary. The preprocessing steps include:

A. Tokenization

The procedure of shattering a flow of text into tokens for example words, phrases etc. The token list thus provides as the input for additional processing like analyzing or text mining.

B. Stop words removal

Stop words are common words which are usually of minor value in selecting documents that match a query. There is a high need of entirely excluding stop words from the vocabulary. The approach for deciding a stop list is by sorting the words by the overall amount of times every single term figures in the document collection, ultimately taking the most repeated terms equal to a stop list, which are then discarded in the course of indexing.

C. Stemming

To describe the reduction of injected or derived words to their word stem, a term called stemming is used in linguistic morphology. It is not necessary that the stem has to be matching to the root of the term. Even though a stem isn't an acceptable root, usually the interrelated words get mapped to the similar stem. For doing the process of stemming, Porter stemmer has been used in order to remove morphological endings in words. The main use is normalization of term which is used for mapping each term in the vocabulary of the document to its root.

D. Removal of alphanumeric characters

The alphanumeric characters that are present in the dataset are removed since they do not help in selecting documents.

E. Term Frequency

The add up of the amount of time a word t appears in a document d. Term frequency is calculated by the equation [7]:

$$tf(t,d) = 0.5 + \frac{0.5 * f(t,d)}{max f(w,d) : w \in d} \quad \ldots\ldots\ldots\ldots (1)$$

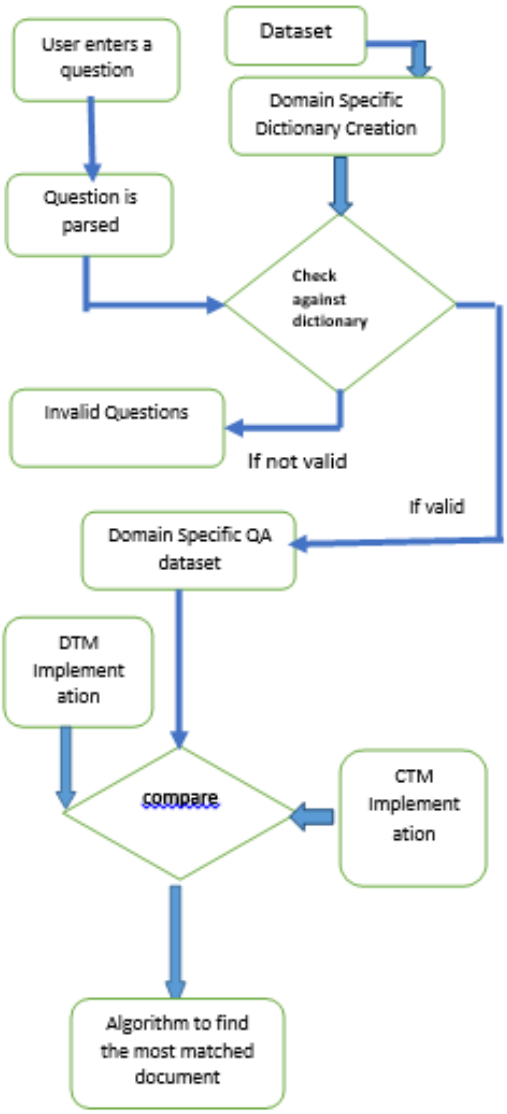

Figure.1. Architecture Diagram

Figure 1 shows the detailed workflow of the proposed work discussed in the paper.

F. Document Frequency

Document frequency is the count of the number of documents where the term occurs.

G. Inverse Document Frequency

It is a degree of the informativeness of the word which means whether the occurrence of the word is usual or infrequent. It can be further explained as the document's logarithmic fraction that contained the term calculated by the way of dividing the entire number of the documents by the entire number of the documents that contain the word. Inverse document frequency is calculated by the equation [7]:

$$idf(t,D) = log \frac{N}{df(t,D)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

H. Term-Frequency Inverse Document Frequency

It is a value that shows the importance for a term in a document of a corpus. Mostly the usage of TF-IDF is as a weighting factor in areas of Information Retrieval. The value of tf-idf upsurges with the amount of time a word is seen in a document, which is offset with the occurrence of the term in the collection of documents. TF-IDF is calculated by the equation [7]:

$$tfidf(t,d,D) = tf(t,d) * idf(t,D) \quad \ldots\ldots\ldots\ldots (3)$$

I. Normalization

Information retrieval methods deal with those documents having varying dimensions in a collection of texts. It is used for fairly retrieving documents of every lengths. Normalization of a token is the procedure of tokens being canonicalized because of which matches arise despite apparent differences in token's character sequences.

On completion of the preprocessing steps, tf-idf is calculated and normalization of the documents is done following the generation of a document-term matrix or term-document matrix describing the occurrence of terms which occur in a group of documents. The rows of a doc-term matrix point to documents of the group and columns point to terms. Inorder to determine the denomination that every entry of the matrix has to take, various schemes exist. One among those schemes is tf-idf.

Once the document-term matrix was generated, the corpus collection is used for dynamic and correlated topic modeling. The Dynamic Topic Model (DTM) is

used for analyzing the evolution of topics in large collection of documents on the basis of the factor namely time. The topics are represented by state space model. To model the topic proportions being document specified, state space models are applied on the regular parameters of logistic Gaussian distribution [2]. In DTM, it is supposed that the data available gets divided by a factor namely time slice. The documents contained in each slice is modeled using K – component topic model [2] where the evolution of topics of t takes place from the topics of slice t develop from that of t-1. Mean parameterization [2] is usually used for representing multinomial distribution. With the known parameters in the multinomial distributions, a method of expectation-maximization is used to update the unknown or hidden parameters namely number of topics In a sequential corpus, the generative process of a slice t is as follows [2]:

1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$
3. For each document:
    a. Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
    b. For each word:
        i. Draw $Z \sim \text{Mult}(\pi(\eta))$
        ii. Draw $W_{t,d,n} \sim \text{Mult}(\pi(\beta_{t,z}))$

Because of the non conjugacy of Gaussian model, variational methods are used for the approximation of posterior inference [2]. The notion behind variational method is the optimization of the distribution's free parameters which makes it close to true posterior [2]. The parameters are topics, topic proportions and topic indicators. Finally, time dynamics is incorporated into the variational approximation of natural parameters. The parameter that plays a key role in DTM is time slices which records the time slices given by the user. Based on the value for time slice, the modeling is performed on the sequential corpus. The dataset considered includes data from four consecutive years.

The main limitation of LDA of not being able to model topic correlation, a real time need is solved by introducing correlated topic model. This limitation occurs because of the usage of the Dirichlet distribution which is used to model the variability amongst the topic proportions. To find out the covariance structure among the topic proportions, logistic normal distribution [9] is used in a significant role. The distribution presumes that the parameter η is in normal distribution and is mapped with the simplex [1]. With the covariance matrix in the Gaussian distribution, correlation is modeled by logistic normal joining the components of random variable. The assumption in CTM is that a document of N-word results from a generative process as follows [1]:

1. Draw $\eta \mid \{\mu, \sum\} \sim \mathcal{N}\{\mu, \sum\}$
2. For $\eta \in \{1, \ldots, \mathcal{N}\}$:
    a. Draw topic allocation $Z_n \mid \eta$ from $\text{Mult}(\int(\eta))$
    b. Draw word $W_n \mid \{z_n, \beta_{1:K}\}$ from $\text{Mult}(\beta_{zn})$

where 'μ' is mean, '∑' is covariance matrix and '$\beta_{1:K}$' represents topics. The process uses logistic normal distribution rather than Dirichlet to draw topic proportions. For predictive distributions with better result, CTM is used as correlated topics which are also associated with probable topics. The model developed uses mean field variational algorithm [1] for finding out posterior inference which is a central challenge. As the logistic normal is not conjugate towards multinomial, posterior inference is done in CTM. The logistic normal not being conjugate makes the computation of topic assignment's log probability difficult. When a document collection is given, the parameter estimation in CTM is carried out by the attempt of maximizing the likelihood in a document corpus. For the maximization, variational expectation-maximization [1] is used. The step is done for maximizing the bound of a collection's log probability. Within E-step, the bound regarding variational parameters is maximized through the performance of variational inference [1] in each document. Within M-step, the bound regarding model parameters is maximized. As correlation is considered, maximum topics are found out and merged that leads to the better retrieval of answers for questions posed by users.

After the implementation of the two modeling algorithms on the corpus, when a user inputs a query, initially it is parsed following the retrieval of the answers for the particular query based on both the modeling algorithms.

4. Results and Discussions

The dataset includes 4000 files related to autism. When a user inputs a question, answer retrieval happens for both the topic modeling algorithms.

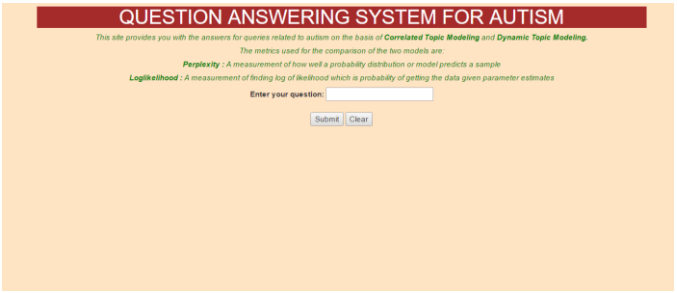

Figure.2. QA interface

Figure.2. shows the interface where users can input questions related to autism. In Figure.3 the retrieved answers along with the perplexity and log likelihood of the models for the particular query are shown.

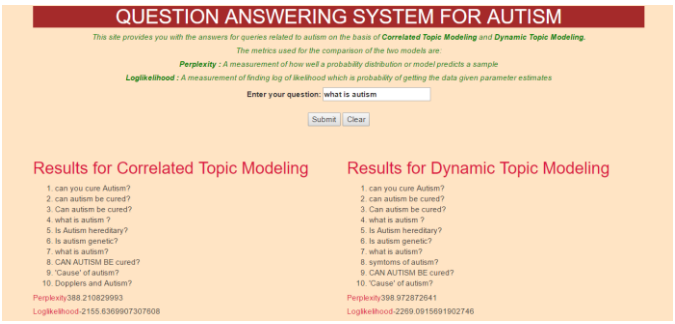

Figure.3. Answers Retrieved

The ratios taken for training and testing of the corpus are 50-50, 60-40, 70-30, 80-20 and 90-10. Based on the answers that both the models retrieve given a query by user, the better model can be identified. The measures used for comparison of the performance of the models CTM and DTM are perplexity and log-likelihood. These measures are used to identify which model predicts the sample in a better way. With increased value of perplexity, the analysis is that performance of the model is poor. From the analysis, correlated model performs better than dynamic model.

Table.1. Log likelihood values

Test %	Train %	No. of test doc	No. of train doc	CTM log likelihood	DTM log likelihood
50	50	2196	2196	-2155.636	-2269.091
40	60	1756	2636	-2663.825	-2804.026
30	70	1316	3076	-3468.609	-3651.167
20	80	876	3516	-5243.828	-5519.829
10	90	436	3956	-10298.66	-10851.87

Table.2. Perplexity Values

Number of Topics	CTM perplexity	DTM perplexity
10	389.154172991	398.902818504
20	389.154172991	398.015161651
40	389.154172991	397.081113417
60	389.154172991	396.063806646
80	389.154172991	395.962911859
100	389.154172991	395.025698629
120	389.154172991	394.698851743
140	389.154172991	393.843855128
160	389.154172991	393.686225908
180	389.154172991	393.02034585
200	389.154172991	393.154172991

In Table.1. The log likelihood values obtained for both DTM and CTM topic models. In Table.2. The perplexity values for both the models were obtained.

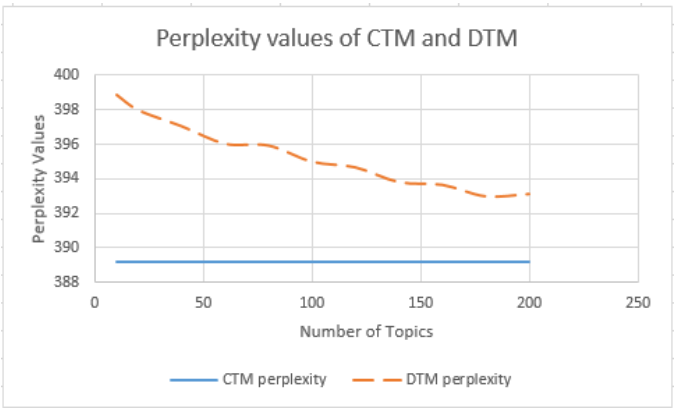

Figure.4. Plot of perplexity values

The plot in Figure.4. shows the perplexity values of CTM and DTM for different number of topics.

Comparing the plot of perplexity values of two models, it is analyzed that CTM performs better than DTM as less value for perplexity causes a model to predict a sample in a better manner.

5. CONCLUSION

In the paper, the topic modeling algorithms implemented are used to find out not only the hidden topics but also the correlation between the topics as well as the time evolution of the same over the years. On comparing the two models considered here namely DTM and CTM on the basis of better predictivity over a sample, CTM performs better. The conclusion is drawn with the help of two metrics namely perplexity and log likelihood.

6. REFERENCES

[1] Blei, David, and John Lafferty. "Correlated topic models." Advances in neural information processing systems 18 (2006): 147.

[2] Blei, David M., and John D. Lafferty. "Dynamic topic models." In Proceedings of the 23rd international conference on Machine learning, pp. 113-120. ACM, 2006.

[3] Celikyilmaz, Asli, Dilek Hakkani-Tur, and Gokhan Tur. "LDA based similarity modeling for question answering." In Proceedings of the NAACL HLT 2010 Workshop on Semantic Search, pp. 1-9. Association for Computational Linguistics, 2010.

[4] Blei, David M. "Probabilistic topic models." Communications of the ACM 55, no. 4 (2012): 77-84.

[5] Alghamdi, Rubayyi, and Khalid Alfalqi. "A Survey of Topic Modeling in Text Mining." Editorial Preface 6, no. 1 (2015).

[6] Fautsch, Claire, and Jacques Savoy. "Adapting the tf idf vector-space model to domain specific information retrieval." In Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1708-1712. ACM, 2010.

[7] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Vol. 1, no. 1, p. 496. Cambridge: Cambridge university press, 2008

[8] J. Aitchison. The statistical analysis of compositional data. Journal of the Royal Statistical Society, Series B, 44(2):139–177, 1982.

[9] Hirschman, Lynette, and Robert Gaizauskas. "Natural language question answering: the view from here." natural language engineering 7, no. 4 (2001): 275-300.

[10] Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. "Reading tea leaves: How humans interpret topic models." In Advances in neural information processing systems, pp. 288-296. 2009.

[11]. Vikas K Vijayan; K. R. Bindu; Latha Parameswaran "A comprehensive study of text classification algorithms" 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 13-16 Sept. 2017, Copyright © 2017, IEEE.

[12]. Rajasundari T., Subathra P., Kumar P.N." Performance analysis of topic modeling algorithms for news articles" Journal of Advanced Research in Dynamical and Control Systems 2017 Special Issue 11 175 183.

[13]K. R. Bindu, L. Parameswaran, K. V. Soumya**,** "Performance Evaluation of Topic Modelling Algorithms with an application of Q & A Dataset "International Journal of Applied Engineering Research, vol. 10, pp. 23-27, 2015