I J C T A, 9(28) 2016, pp. 175-179
© International Science Press

SUB-TOPIC MODELING-A HIERARCHY MODEL FOR TOPIC CORRELATIONS

R.Uma* and B.Latha**

Abstract: Probabilistic Latent semantic Analysis (pLSA) is one of the commonly used techniques for determining the co-occurrence of data in a document. This method however makes no assumptions on how the mixture weights are generated. Latent Dirichlet Allocation (LDA) is yet another most widely employed topic modeling approach. But it provides co-relation among distribution of words but not topics. PAM(Pachinko Allocation Model) provides a flexible alternative to LDA by capturing correlations between topics in the corpus. We propose a generative probabilistic model that describes generation of words in a document based on hierarchy of latent variables with correlation of topics and without losing the statistical relationship of the underlying topics we also describe that by considering the correlation between the topics, better retrieval performance can be achieved. Using this proposed technique we show enhanced retrieval performance in classification of news articles, email, etc. In this paper we focus on bag of words model however the described method can also be applied to other structures like n -grams etc.

Keywords: information retrieval, bag of words, latent variable, correlation, topicmodelling, document clustering

1. INTRODUCTION

The growth of electronic information has led to the development of information age. This explosive growth of information has led to the development of methods for effectively organizing and Indexing the data. Once the information is organized and indexed properly, information retrieval becomes a easy task. Electronic information is available in many forms text, audio, video, image etc. Even though there are many forms of information available text becomes the primary means of information. In this paper we have discussed a technique for effectively retrieving information from text document. Topic models are being extensively used for this purpose. Topic model is the one in which a document is a mixture of topics, and topic is a probability distribution of words. A document in a topic model is generated by a probabilistic procedure. Many Topic models are available currently. In this paper we have proposed a new model which considers the correlation between the topics.

PLSI (Probabilistic Latent Semantic Indexing): PLSI is a statistical model that models every word as a sample to the mixture in the document. This probabilistic model represents the document as a mixing proportion of topics, with each word generated from a topic. But PLSI introduces the problem of increasing number of parameters with increase in size of the text corpus.

Latent Dirichlet Allocation (LDA): Probabilistic topic models are based upon the idea that the documents are made up of a collection of topics with each topic being a probability distribution of words. Latent Dirichlet Allocation is a widely used topic modeling approach that is used to derive word correlation under topics in a document. Latent Dirichlet Allocation is a generative three level hierarchical Bayesian model in which each document corpus is viewed as a mixture of topics ,where topic probabilities give

* Research Scholar, Anna University, Chennai, Tamil Nadu, India &1Associate Professor, Department of Computer Science and Engineering, Sri Sai Ram Engineering College, Chennai, Tamil Nadu, India, **Email:** umavina11@gmail.com

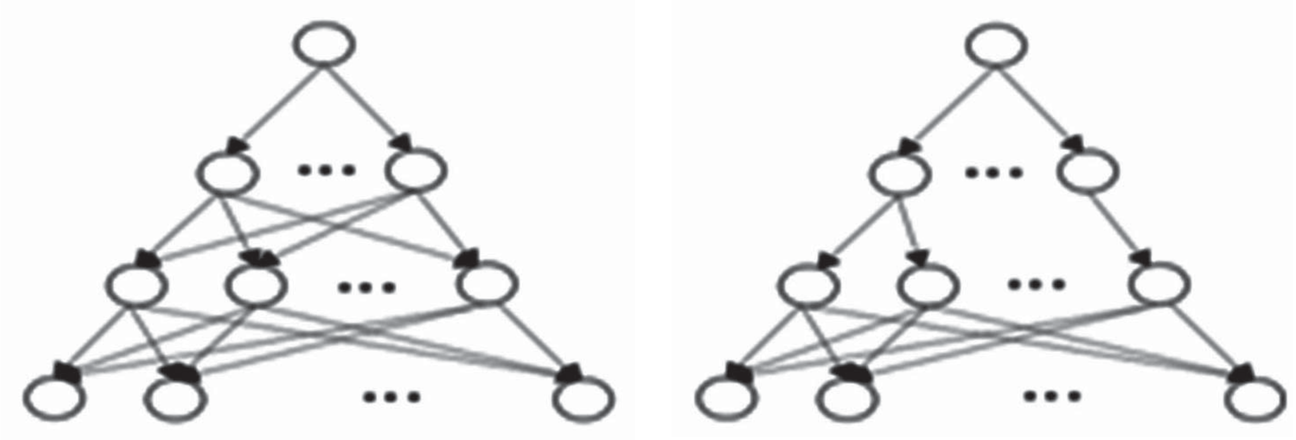
** Professor, Department of Computer Science and Engineering, Sri Sai Ram Engineering College, Chennai, Tamil Nadu, India. **Email:** sivasoorya2003@yahoo.com

a brief representation of the document. LDA although discovers correlations among words it doesn’t provide any correlation between topics.

Pachinko Allocation Model(PAM): The Pachinko Allocation Model uses a DAG structure to determine the topic co-relations thereby extending the distribution also over to topics. By using an arbitrary graph structure and with many nodes, PAM captures the co-relations among topics which is limited to only words in case of LDA. However, the major limitation of PAM is that with the increase in the number of nodes, the complexity of the DAG increases which thereby rises the time complexity to capture the relationship.

2. PROPOSED TECHNIQUE

In this paper, we propose a generative probabilistic sub-topic modeling approach. The proposed method assumes that the document is a mixture of random topics, with each topic characterized by a collection of sub-topics.



(a) Four-Level PAM

(b) Sub-Topic Modeling

Figure 1. Sub-Topic Modeling approach consisting of the document as the root node and the word vocabulary as leaf nodes.

The model consists of a tree structure with the document as the root node and its children being the topics. The leaf nodes are the words which are distributed under a sub-topic. A sub-topic may be distributed under one or

more topics. By extending the hierarchy to a sub-topic node, we will not only be able to capture the co-relationship among topics, but also reduce the time complexity relative to that of PAM. In this paper, we also demonstrate the results of the proposed sub-topic modeling approach in comparison to that of LDA and PAM describing its improved performance over the former two approaches.

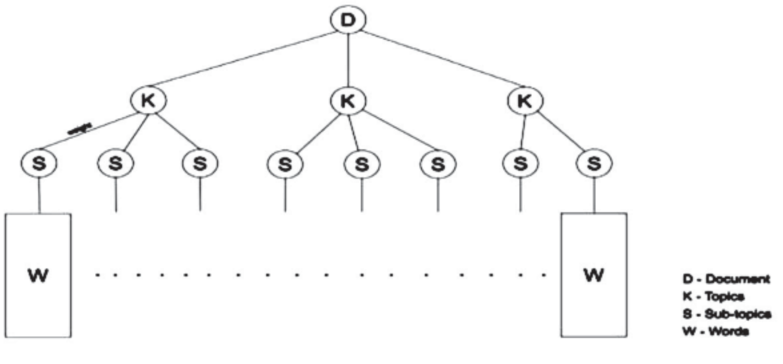


Figure 2. Topic co-relation in Sub-Topic Modeling. Each topic’K’ is connected to one or more sub-topics.

2.1 The Model

In this section, we explain the sub-topic modeling approach, its terminologies and the estimation algorithm. A document can be viewed as a mixture of random topics containing a bag of words distributed under each topic. Each word w_i comes under a particular topic, *i* referring to the index of the word in the vocabulary which is generated by Gibb's sampling from a sub-topic over the topic distribution.

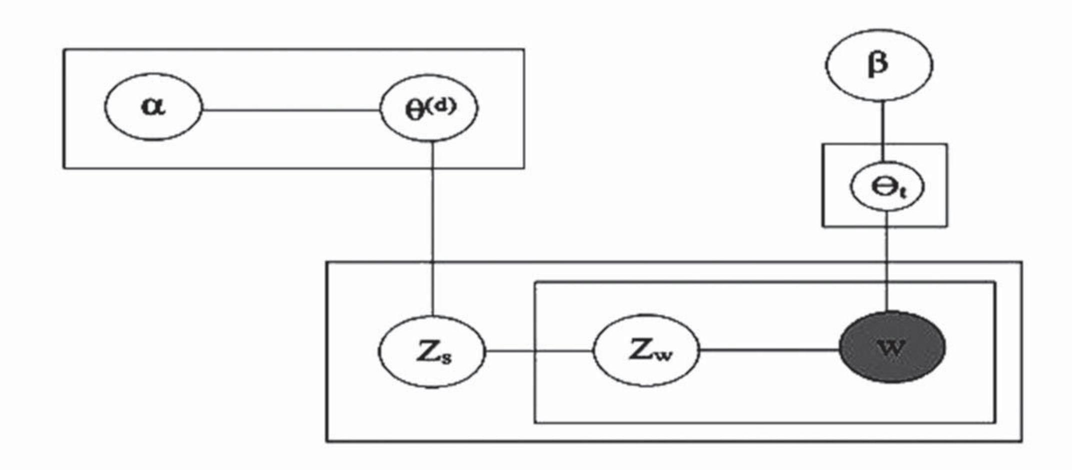


Figure 3. Modeling Approach

Let us consider that P(*z*) be the probability distribution of topics *z* in the document and P(*s*|*z*) is the distribution of the subtopics over the topics and P(*w*|*s*) is the subtopic-words distribution. We assume that $P(S_{i=j})$ is the probability that the j^{th} subtopic was sampled for the i^{th} word and also $P(w_i|S_{i=j})$ is the probability of the word w_i under the sub-topic *j*. The generative process for the document can be viewed as follows,

1. Sample $\Theta^{(d)}_{t1}$,.. $\Theta^{(d)}_{tz}$ where $\Theta^{(d)}_{t1}$ is the multinomial distribution of the topic *ti* over its sub-topics.
2. Sample $\Theta^{(t)}s_1$,... $\Theta^{(t)}_{sz}$ where $\Theta^{(t)}_{sz}$ is the multinomial distribution of the sub-topic s_j over its children.
3. For each word *w*, compute the following,
   (a) Compute the distribution of the words with the sub-topics (i.e.,) P(word wi|subtopics *i*)
   (b) Sample the path of the sub-topic si through its parent topic ***ti*** according to the multinomial distribution of how much the sub-topics s(n-1)overall other sub-topics of the topic ***ti***.
   (c) Estimate the word sample wi from $\Theta^{(t)}s_i$

This produces a hierarchical structure with the root node as the document D, its children being the topics T, and each topic has a number of sub-topics. The leaf nodes of the tree being the words W. A word may come under one or more sub-topics and each topic may have multiple sub-topics

$$P(d, t^{(d)}, \Theta^{(d)} |\alpha) = \pi^{s}{}_{i=1} P(\Theta^{(d)} |\alpha_i)\cdot \pi_w \Sigma_{t(w)} (\pi_{i=2}{}^{L}P(t_{w(i)}|\Theta^{(d)} t_{w(i-1))\cdot}P(w|\Theta^{(\delta)}{}_{twLw}))$$

3. EXPERIMENTAL OBSERVATION

In our experimental study, we use L3S-GN1 dataset to train the models and likelihood comparison. To determine the likelihood with other models like LDA and PAM, we utilize an empirical likelihood approach by sampling from the models and estimating the empirical distributions. We use 20% of corpus as test data and the other 80% for training.

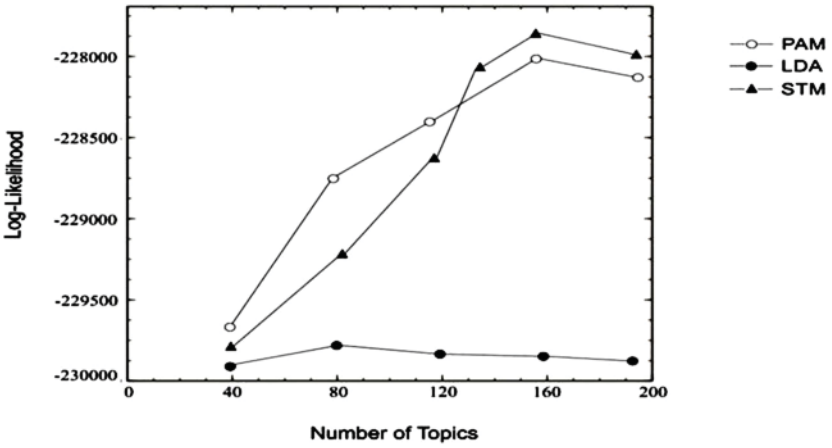


Figure 4. Performance Evaluation of Proposed Approach

We use a set of 40 topics and 120 to 140 sub-topics for the study. Our proposed method results in a higher likelihood in comparison with LDA and PAM and the performance gets better with large datasets. For each class, we train an individual model and estimate the likelihood using the test data.

Example: Here we have provided an example of the use of STM discovery on real world data of over 5000 documents from TREC corpus with 325467 word tokens. Based on the Dirichlet parameter, we determine the correlations between the super-topics in the data. Unlike PAM, we generate an acyclic graph with a collection of sub-topics under a single super-topic.

3.1 Human Evaluation

In the experiments, we discussed above, over 10 human evaluators are provided with a set of topics generated using STM, PAM and LDA. Based on the human evaluation of 40 topics presented to them, the evaluators found that the topics generated by STM to be better than the ones by PAM and LDA.

Document Clustering: Document Clustering techniques are commonly used to facilitate topic modeling. STM along with Document Clustering enables us to form document clusters based on topics that are shared among the documents. Unlike LDA that extracts a single set of topics from a document, STM facilitates the mechanism to group documents based on the topic behaviors.

Table 1. Clustering

Document Cluster			
Topic 1	*Topic 2*	*Topic 3*	*Topic 4*
Election	People	Russian	University
President	War	Turks	Education
Democracy	World	People	Unemployment
Country	History	Soviet	Professor
Voter	Years	American	Information
People	Time	Jews	Literacy
Government	Peace	Chinese	Effects

4. CONCLUSION

In this paper, We introduced a new technique for topic modeling based on sub-topic correlations and from the experimental observations, it is evident that the proposed method performs better than the existing technique. Each leaf node corresponds to a word, that is associated with several sub-topics. This approach can be viewed as a special case of LDA, whereas it provides less ambiguous approach on capturing topic correlations.

References

1. Blei, D., Ng, A.,& Jordan, M. (2003). "Latent Dirichlet Allocation". Journal of Machine Learning Research, 3, 993-1022.
2. Li, Wei; McCallum, Andrew (2006) "Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations". Proceedings of the 23rd International Conference on Machine Learning.
3. Blei, D.,& Laerty, J. (2006). "Correlated topic models". In Advances in neural information processing systems .
4. Hofmann, Thomas (1999). "Probabilistic Latent Semantic Indexing". Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval.
5. T. Grifths and M. Steyvers (2002) "A Probabilistic approach to semantic representation". In Proceedings of the 24th Annual Conference of the Cognitive Science Society.
6. C. Papadimitriou, H. Tamaki, P. Raghavan,& S. Vempala (1998), "Latent Semantic Indexing: A Probabilistic analysis". In Proceedings of the ACM Conference on Principles of Database Systems (PODS), pages 159-168,Seattle.
7. Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts (2013), "Recursive Deep Models for Semantic Compositionality Over a Sentiment Tree bank". Conference on Empirical Methods in Natural Language Processing (EMNLP).
8. Lawrie, D., Croft, W., & Rosenberg, A. (2001). "Finding topic words for hierarchical Summarization". Proceedings of SIGIR'01 (pages. 349-357).
9. Adler Perotte, Nicholas Bartlett, No`emie Elhadad, Frank Wood "Hierarchically Supervised Latent Dirichlet Allocation".
10. D. Blei and J. McAuli_e (2008), "Supervised Topic models". Advances in Neural Information Processing, 20:121-128.
11. Justin Grimmer, (2009) "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases". December 3.
12. D Mimno, W Li, A McCallum (2007), "Mixtures of hierarchical topics with pachinko allocation". Proceedings of the 24th International Conference on Machine Learning, Corvallis.