# Correlation between the Topic and Documents Based on the Pachinko Allocation Model

**Dr.C.Sundar**
*Associate Professor*
*Department of Computer Science & Engineering*
*Christian College of Engineering and Technology*
*Dindigul, Tamilnadu-624619, India*

**V.Sujitha**
*PG Scholar*
*Department of Computer Science & Engineering*
*Christian College of Engineering and Technology*
*Dindigul, Tamilnadu-624619, India*

## Abstract

Latent Dirichlet allocation (LDA) and other related topic models are increasingly popular tools for summarization and manifold discovery in discrete data. In existing system, a novel information filtering model, Maximum matched Pattern-based Topic Model (MPBTM), is used.Each topic is represented by patterns. The patterns are generated from topic models and are organized in terms of their statistical and taxonomic features and the most discriminative and representative patterns, called Maximum Matched Patterns, are proposed to estimate the document relevance to the user's information needs in order to filter out irrelevant documents.The Maximum matched pat-terns, which are the largest patterns in each equivalence class that exist in the received documents, are used to calculate the relevance words to represent topics. However, LDA does not capture correlations between topics and these not find the hidden topics in the document. To deal with the above problem the pachinko allocation model (PAM) is proposed. Topic models are a suite of algorithms to uncover the hidden thematic structure of a collection of documents. The algorithm improves upon earlier topic models such as LDA by modeling correlations between topics in addition to the word correlations which constitute topics. In this method the most accurate topics are given to that document. PAM provides more flexibility and greater expressive power than latent Dirichlet allocation.

**Keywords: Topic model, information filtering, maximum matched pattern, correlation, hidden topics, PAM**

## I. INTRODUCTION

Data mining is the process of analyzing data from different perspect practice of examining large pre-existing databases in order to generate new information. It is the process of analyzing data from different perspectives and summarizing it into useful information. Similarly text mining is the process of extracting data from the large set of data. It also refers to the process of deriving high quality information from text. Topic modeling has become one of the most popular probabilistic text modelling techniques and has been quickly accepted by machine learning and text mining communities. It can automatically classify documents in a collection by a number of topics and represents every document with multiple topics and their corresponding distribution. Topic models are an important tool because of their ability to identify latent semantic components in un- labeled text data. Recently, attention has focused onmodels that are able not only to identify topics but also to discover the organization and co-occurrences of the topics themselves. Text clustering is one of the main themes in text mining .It refers to the process of grouping document with similar contents or topics into clusters to improve both availability & reliability of the mining [2].

Statistical topic models such as latent dirichlet allocation (LDA) havebeen shown to be effective tools in topic extraction and analysis. These models can capture word correlations in a collection of textual documents with a low-dimensional set of multinomial distributions. Recent work in this area has investigated richer structures to also describe inter-topic correlations, and led to discovery of large numbers of more accurate, fine-grained topics. The topics discovered by LDA capture correlations among words, but LDA does not explicitly model correlations among topics. This limitation arises because the topic proportions in each document are sampled from a single dirichlet distribution.

In this paper, we introduce the pachinko allocation model (PAM), which uses a directed acyclic graph (DAG) [9], [11] structure to represent and arbitrary, nested, and possibly sparse topic correlations. In PAM, the concepts of topics are extended to be distributions not only over words, but also over other topics. The model structure consists of an arbitrary DAG, in which each leaf node is associated with a word in the vocabulary, and each non-leaf "interior" node corresponds to a topic, having a distribution over its children. For example, consider a document collection that discusses four topics: cooking, health, insurance and drugs. The cooking topic co-occurs often with health, while health, insurance and drugs are often discussed together [11]. A DAG can describe this kind of correlation. For each topic, we have one node that is directly connected to the words. There are two additional nodes at ahigher level, where one is the parent of cooking and health, and the other is the parent of health, insurance and drugs.

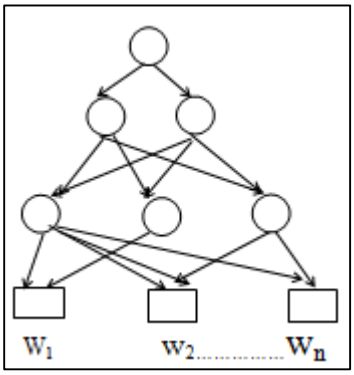

Fig. 1: DAG Structure

The DAG structure in the PAM is extremely flexible shows the Fig.1.

Topic modeling is a form of text mining, a way of identifying patterns in a corpus. In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract topics that occur in a collection of documents. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently. "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear equally in both. A document typically concerns multiple topics in different proportions: thus, in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each documents balance of topic is.

Formally, a topic is a probability distribution over terms. In each topic, different sets of terms have high probability, and we typically visualize the topics by sets.

Topic models provide a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings. Topic models are a suit of suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help us develop new ways to search, browse and summarize large archives of texts. A topic model takes a collection of texts as input and it will produce the set of topics for the document.

We can use the topic representations of the documents to analyze the collection in many ways. For example, we isolate a subset of texts based on which combination of topics they exhibit. Or, we can examine the words of the texts themselves and restrict attention to the politics words, finding similarities between them or trends in the language. Note that this analysis factors out other topics from each text in order to focus on the topic of interest. Both of these analyses require that we know the topics and which topics each document is about. Topic modelling uncovers this structure. They analyze the texts to find a set of topic patterns of tightly co-occurring terms and how each document combines them.

We formally define a topic to be a distribution over a fixed vocabulary. For example the genetics topic has words about genetics with high probability and the evolutionary biology topic has words about evolutionary biology with high probability. We assume that these topics are specified before any data has been generated. Now for each document in the collection, we generate the words in a two-stage process.

– Randomly choose a distribution over topics.
– For each word in the document

a) Randomly choose a topic from the distribution over topics in step.
b) Randomly choose a word from the corresponding distribution over the vocabulary.

The topic model is one of the techniques to uncover the thematic structure of the documents. Uncovering the topics within short texts, such as tweets and instant messages, has become an important task for content analysis applications. It gives flexible structure out of a corpus on the minimal criteria. A tool like MALLET gives a sense of the relative importance of topics in the composition of each document.

Topic modeling performs the topic proportions assigned to each document are often used, in conjunction with topic word lists, to draw comparative insights about a corpus. Boundary lines are draw around document subsets and topic proportions aggregated within each piece of the text. The corpus are first identified and fixed in the same set of proportions. Then the text mining techniques are applied to the text to give the correlated words. Topic modeling, instead allows us to step back from individual documents and look at larger patterns among all the documents. Topic models, probabilistic models for uncovering the underlying semantic structure of a document collection based on the original text. The topic model give the general way to identify the corpus in the given text document.

With the statistical tools we can describe the thematic structure of the given document. It is well understood and gives best and accurate topics for the given document.

## II. RELATED WORKS

Sailaja G and Prajna B proposed the effective frequent item set-based [8] document clustering approach. First, the text documents in the text data are preprocessed with the aid of stop words removal technique and stemming algorithm. The mined frequent itemsets are sorted in descending order based on their support level for every length of itemsets. Subsequently, we split the documents into partition using the sorted frequent itemsets. These frequent itemsets can be viewed as understandable description of the obtained partitions.

YuetingZhuang, HaidongGao,To ruling the probability [10] [5] of the each word in the document, the LDA method can be used. But the occurrences of the words in the each document are collected, based on that the high probability word is chosen for the topic model. The words are arranged in the decedent order then we can use the ranking method to take the words.

The maximum matched patterns [3] are collected in the given document. The patterns which represent user interests are not only grouped in terms of topics, but also partitioned based on equivalence classes in each topic group. The patterns in different groups or different equivalence classes have different meanings and distinct properties. Thus, user information needs are clearly represented according to various semantic meanings as well as distinct properties of the specific patterns in different topic groups.

Wei Li and Andrew McCallum Proposed a DAG-Structured Mixture Models of Topic Correlations: PAM uses the DAG [13] structure model. These methods easily model the words in the document. It will first construct the super topic. Then each super topic corresponds to one sub topic. The other word occurrences are collected. The leaves of the DAG represent individual words in the vocabulary, while each interior node represents a correlation among its children, which may be words or other interior nodes (topics).

The pachinko allocation model (PAM), which uses a directed acyclic graph (DAG) structure to represent and learn arbitrary-arity, nested, and possibly sparse topic correlations. In PAM, the concept of topics is extended to be distributions not only over words, but also over other topics. The model structure consists of an arbitrary DAG, in which each leaf node is associated with a word in the vocabulary, and each non-leaf "interior" node corresponds to a topic, having a distribution over its children. An interior node whose children are all leaves would correspond to a traditional LDA topic. But some interior nodes may also have children that are other topics, thus representing a mixture over topics. With many such nodes, PAM therefore captures not only correlations among words (as in LDA), but also correlations among topics themselves.

David Mimno, Wei Li and Andrew McCallum proposed a Mixtures of Hierarchical Topics with Pachinko Allocation: The proposed model PAM [5] model constructs the accurate topic for the given document. Everything will be considered as the mixture of the DAG structure [9]. The DAG will easily construct the topics because of the simple structure. The flexibility of attaining the good topics and also gaining the accuracy. In PAM the concepts are extended to distributions not only over words, but also over other topics.

Bettina Grun and Kurt Hornihave proposed a topic model techniques is a type of statistical model for discovering the abstract topics. It provides a simple way to analyze large volumes of unlabeled data. The topic modeling [7] such us LDA has given the good statistical results in single topic probability. But the correlations among the words are not finding in the LDA.A topic model dense a probabilistic procedure to generate documents as mixtures of a low-dimensional set of topics. Each topic is a multinomial distribution over words and the highest probability words brief summarize the themes in the document collection. As an elective tool to dimensionality reduction and semantic information extraction, topic models have been used to analyze large amounts of textual information in many tasks, including language modeling, document classification, information retrieval, document summarization, data mining and social network analysis.

## III. PROPOSED WORK

The proposed topic modeling technique is used to represent the documents as the meaningful information. The topics are given to the documents based on the word correlation. Each word in the documents is captured and the high probability word is chosen for representing the topics. Here the intermediate topics are first constructed and then the given the super-topic, we sample a sub-topic in the same way. Finally we sample the word from the sub-topic according to its multinomial distribution over words. The word correlation ismainly captured to give suitable topic for the document. Given the documents it will first check the similarity words and find the maximum matched patterns which are having the high probability. Based on the high probability the words arranged in the descendant order. After that the topic will be chosen using the topic model technique.

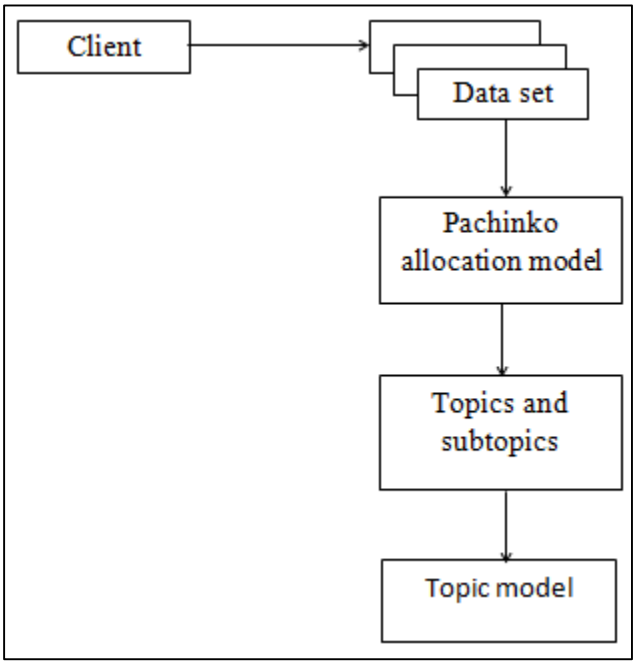


Fig. 2: Architecture diagram

The Fig.2 shows the working principle of PAM. First it will get the input from the user. The input contains the collection of documents. Then those documents are processed and it will be preprocessed to remove the uncorrelated words. The pachinko allocation model is applied to generate the new topics. The model constructs the DAG [9] structure to represent the topics. The root node occupy the topics leaf nodes are consider as super and sub topics. Based on that it will take the correlated words to emulate the new and accurate topic for the given document.

### *A. Dataset Preprocessing*

First of all the user have to load the dataset. Preprocessing includes apply stop words and stemming in the dataset. Because very often the number of patterns in some of the topics can be huge and many of the patterns are not discriminative enough to represent specific topics. The system present to integrate data mining techniques with statistical topic modelling techniques to generate a pattern-based topic model to represent documents and document collections. The preprocessing mainly used to remove the words which are not related to those topics. So very often the words in the document are preprocessed and then only the words which are used to represent the suitable topics.

### *B. LDA*

LDA supports a very strong foundation for generating semantics in terms of topic representation and topic distribution. Find the support value for the words in the file. And sort the data in descending order. In word topic assignment the sorted words are splitted into three parts. And find the probability value for the splitted data. The probability value will be used to calculate the words which are closely to that topic. The word which is having the high probability is taken for constructing the topics.

### *C. Pachinko Allocation Model*

There are many words present in the document dataset. In that the word probability is used for the topic selection for the dataset. The intermediate topics are calculated in all the documents dataset. The intermediate topics are the words in the document having high probability. It is connected to all the leaf nodes related to the intermediate topic. The leaf nodes are the low probability words in the document.

Pachinko allocation models documents as a mixture of distributions over a single set of topics, using a directed acyclic graph to represent topic co-occurrences. Each node in the graph is a Dirichlet distribution. At the top level there is a single node. Besides the bottom level, each node represents a distribution over nodes in the next lower level. The distributions at the bottom level denote distributions over words in the vocabulary.

### *D. Topic Modelling*

A topic model is a type of statistical model for discovering the abstract topics. It provides a simple way to analyze large volumes of unlabeled data. The topic modeling such us LDA has given the good statistical results in single topic probability. But the correlations among the words are not finding in the LDA. Finally, based on the sub and super topics it will collect the words. The correlated words are sorted in the descending order. The accurate topics are given based on the collected words.The leaf words

are connected to the intermediate nodes (the high probability words in the document). The document is a mixture of collection of topics. The Probability is calculated for the all the topics in every document. Based on the high probability the topic is fixed for each document in the dataset. This method is more accurate than the latent dirichlet approach.

## IV. EXCREMENTAL ANALYSIS AND RESULTS

For different document collections, the number of topics involved in the collections can be different. Therefore selecting an appropriate number of topics is important. As Table 5 shows, the result of the PAM with 5 or 10 topics achieves relatively the best performance for this particular dataset. When the topic number rises or reduces, the performance drops. Especially when the topic number rises to 15, the performance drops dramatically, although still outperforms most of the baseline models In the experiments we discuss below, we use a fixed four-level hierarchical structure for PAM, which includes a root, a set of super-topics, a set of sub-topics and a word vocabulary. For the root, we always assume a fixed Dirichlet distribution with parameter 0.01. We can change this parameter to adjust the variance in the sampled multinomial distributions. We choose a small value so that the variance is high and each document contains only a small number of supertopics, which tends to make the super-topics more interpretable. We treat the sub-topics in the same way as LDA and assume they are sampled once for the whole corpus from a given Dirichlet with parameter 0.01. So the only parameters we need to learn are the Dirichlet parameters for the super-topics, and multinomial parameters for the sub-topics.

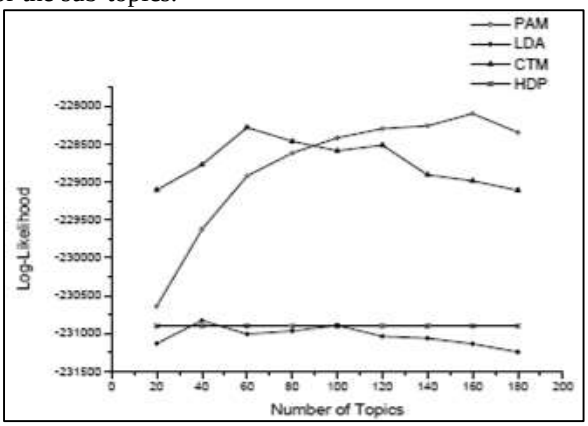


Fig. 3:Likelihood comparison with different numbers oftopics: the results are averages over all samples in 10 different Gibbs sampling

We show the log-likelihood on the test data in Figure 3, averaging over all the samples in 10 different Gibbs sampling. Compared to LDA, PAM always produces higher likelihood for different numbers of sub-topics. The advantage is especially obvious for large numbers of topics. LDA performance peaks at 40 topics and decreases as the number of topics increases. On the other hand, PAM supports larger numbers of topics and has its best performance at 160 sub-topics.

## V. CONCLUSION

In this paper, we have presented pachinko allocation,a mixture model that uses a DAG structure to capture arbitrary topic correlations. Each leaf in the DAG is associated with a word in the vocabulary, and each interior node corresponds to a topic that models the correlation among its children, where topics can be not only parents of words, but also other topics. The DAG structure is completely general, and some topic models like LDA can be represented as special cases of PAM.
Compared to other approaches that capture topic correlations
such as hierarchical LDA and correlated topic model, PAM provides more expressive power to support complicated topic structures and adopts more realistic assumptions for generating documents.The future work is to extend the some of the other methods to describe the topic models efficient than PAM.

## REFERENCES

[1] V.Sujitha, C.Sundar,2015, "Correlation Between the topic and the Documents Based on the Pachinko Allocation Model",Journal of Applied Science and Research.,pp.50-55.
[2] Yang Gao, YueXu, and Yuefeng Li, 2015, "Pattern-based Topics for Document Modelling in Information Filtering, IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 6.
[3] B.Prajna, G. Sailaja,2014," A Novel Similarity Measure for frequent Term Based Text Clustering on high dimensional data" IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10.
[4] S.DurgaBhavani, S.Murali Krishna ,2014"Performance Evaluation of an Efficient Frequent Item sets-Based Text Clustering Approach" Vol. 10 Issue 11 (Ver. 1.0).
[5] Fei Wu, HaidongGao, Siliang Tang, YuetingZhuang, , Yin Zhang and Zhongfei Zhang,2013," Probabilistic Word Selection via Topic Modeling" IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 6.
[6] EnhongChen ,HuiXiong, Haiping Ma, LinliXu,2012,"Capturing correlations of multiple labels: A generative probabilistic model for multi-label learning".
[7] Christopher S. Corley, Elizabeth A. Kammer, Nicholas A. Kraft,2012, "Modeling the Ownership of Source Code Topics" IEEE.

[8] S.DurgaBhavani, S.Murali Krishna,2010, " An Efficient Approach for Text Clustering Based on Frequent Itemsets" European Journal of Scientific Research ,ISSN 1450-216X Vol.42 No.3 , pp.385-396.
[9] Andrew McCallum, Wei Li,2008,"Pachinko Allocation: Scalable Mixture Models of Topic Correlations"Journal of Machine Learning Research.
[10] Ivan Titov, Ryan McDonald,2007 "Modeling Online Reviews with Multi-grain Topic Models".
[11] Andrew McCallum, David Mimno, 2006,"Pachinko Allocation:DAG-Structured Mixture Models of Topic Correlations".
[12] A. McCallum, D. Blei, and W. Li, 2007, "Nonparametric Bayes pachinko allocation", In UAI.
[13] A. McCallum ,W. Li,2008, "Mixtures of hierarchical topics with pachinko allocation", In ICML.
[14] David M. Blei,2006," Introduction to Probabilistic Topic Models".
[15] David M. Blei ,2007 John D. Lafferty," Correlated Topic Models"
[16] Bettina Grun, Kurt Hornik,2011," topicmodels: An R Package for Fitting Topic Models", Journal of Statistical Software.
[17] M. Steyvers and T. Griffiths, 2007,"Probabilistic topic models," Handbook Latent Semantic Anal, vol. 427, no. 7, pp. 424–440