# LITERATURE REVIEW ON PATTERN BASED TOPIC MODELING IN INFORMATION FILTERING

[1]Bencymol P P, [2]Manju K

[1]M-Tech Student, [2]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]College Of Engineering Cherthala, Alappuzha, India

---

***Abstract :*** Topic modeling is a major technique used in machine learning and natural language processing to discover topics that occur in the collection of documents. Topic models uncover the hidden semantic concepts in the corpora. There are several statistical methods for topic modeling like Latent Semantic Analysis(LSA), Probabilistic LSA(PLSA), Latent Dirichlet Allocation(LDA). Using these methods, we can generate multiple topics from the text corpora. Topic models can also be used for information filtering. But these methods cannot be directly applied in information filtering since topic models only includes group of words which has limited semantic meaning. So pattern based representation is used for information filtering. Pattern based topic modeling is an effective approach for information filtering.

***IndexTerms*** **- Topic modeling,LSA,PLSA,LDA,Information filtering,Pattern based representation.**

---

## I. INTRODUCTION

Digital information is an important part of our day to day life. It is an indivisible part of modern generation as well as old generation. Large amount of information are now available and accessible on the web. Each user interested in different information. For information access users emphasize on search engines. Search engines helps us to search for information based on terms. The problem of these term based search is the difference in semantic meaning of words. For example, if we search with the term ”Apple”, Apple is a fruit and also it is a brand. The search engine does not understand what the user is expected from that query. A better solution would be to divide documents into sections, with each section limited to the extent of a particular topic. And the boundaries between topic segments aligned with sentence boundaries for clarity. Natural language processing techniques can play a vital role in these difficulties. Topic modeling is a widely used natural language processing technique.

Topic modeling[1][2][3] is a frequently used technique for discovering the hidden semantic structure in documents. Topic model is a type of statistical model for discovering the abstract topics that occur in a collection of documents. Or it can be described as a method for finding a group of words (i.e topic) from a collection of documents that best represents the information in the collection. Topic modeling has become one of the most popular probabilistic text modeling 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. Term based[4] and phrase based[5] methods are generally used methodologies. But direct application of these methods have some limitations and also common people cannot understand the representation. This limitation can be overcome with the use of pattern based representation. Patterns carry more semantic meaning than words.

Using patterns rather than words, users can understand the topics easily. But user is interested in some of these topics. The available information need to filtered. Information filtering systems are widely used for this purpose. This system removes redundant or unwanted information from documents or other information streams based on user’s interest. There are many information filtering systems like pattern based topic models, maximum matched pattern etc. For the purpose of filtering, first the documents must be categorized as topics. Pattern based models are another approach, it uses pattern mining techniques to represent users interest. Patterns carry more semantic meaning than words which increases the understandability of topics. For this quality of the patterns need to be improved by developing advanced mining techniques. A good pattern should be free from redundancy and noise.

Latent Semantic Analysis(LSA)[6], Probabilistic LSA[7], LDA[3] are most popular topic modeling techniques. And extension to these LDA are also described in this survey. And it gives a brief review about some information filtering methods. They are A Two-stage Approach for Generating Topic Models[8], Pattern based topic model for information filtering (PBTM)[9], and maximum matched PBTM[10].

The paper is organized as follows. Section 2 gives a brief overview of topic modeling in information filtering and Section 3 concludes the paper.

## II. LITERATURE REVIEW

The information filtering process can be divided into two stages, topic modeling and information filtering. By using topic modeling methods, available documents are divided into a number of topics. Then information filtering is applied with the help of user interest model.

### 2.1 Methods of Topic Modeling

There are several methods for topic modeling and its representation. This section highlights on the current topic modeling techniques, methodologies and its features.

Papadimitriou et al. proposes Latent Semantic Analysis(LSA), arose from the problem of how to find relevant documents form search words. In the past, LSA is named as LSI(Latent Semantic Indexing) which is used for information retrieval and now LSA used for Natural Language Processing. The fundamental difficulty arises when words are compared to find relevant documents. LSA assumes that words that are close enough in meaning will occur in similar piece of text. A matrix containing word count per paragraph is constructed from a large piece of text. Singular Value Decomposition(SVD)[11] is used to reduce the number of rows. Then words are compared by taking the cosine of the angle between the two vectors formed from any two rows. Values close to 1 indicates very similar words and values close to 0 indicates dissimilar words.

Let X be a matrix representing elements (i, j) where i is the term and j is the document. Row in the matrix represent the vector, $t_i$ to the corresponding term. Similarly column represent vector, $d_j$ to the corresponding document. Now the dot product between two term vectors gives the correlation between the terms over the set of documents. The matrix product $XX^T$ contains all these dot product. Then by SVD,

$$X = U\sum V^T \quad (1)$$

U and V are orthogonal matrix and $\sum$ is a diagonal matrix. Singular values are computed for row vector and column vector. Now these singular values are used to rearrange the data. Latent Semantic Analysis has many nice properties that make it widely applicable to many problems.

- The documents and words are mapped to the same concept space. We can cluster documents and words in this concept space. This helps to retrieve documents based on words and vice versa.
- Compared to the original matrix, the concept space has vastly fewer dimensions. These dimensions have been chosen specifically because they contain the most information and least noise. So the new concept space is ideal for running algorithms such as testing different clustering algorithms.
- LSA can find things that may not be apparent to a more locally based algorithm since it is an inherently global algorithm that looks at trends and patterns from all documents and all words. LSA can also be fairly combined with a more local algorithm such as nearest neighbours to become more useful than either algorithm by itself.

There are a few limitations that must be considered when deciding whether to use LSA. Some of these are:

- LSA assumes a Gaussian distribution and Frobenius norm which may not able to solve all problems. For example: rather than a Gaussian distribution, words in documents seem to follow Poisson distribution.
- LSA cannot handle polysemy (words with multiple meanings) effectively. It assumes that the same word means the same concept which causes problems for words like book, mole, crane etc that have multiple meanings depending on which contexts they appear in.
- LSA depends heavily on SVD. As new documents appear, SVD become computationally intensive and hard to update. However recent work has prompted a new efficient algorithm which can update SVD based on new documents in a theoretically same sense.

Probabilistic Latent Semantic Analysis(PLSA) is an enhanced version of LSA, which is introduced by Jan Puzicha and Thomas Hofmann it in the year 1999. It is a statistical model applied for both information retrieval and filtering, natural language processing etc. Compared to standard LSA method which uses linear algebra and performs singular value decomposition of co-occurrence words , PLSA is based on a mixture decomposition derived from a latent class model. The main idea of PLSA is a statistical model known as aspect model, which is a latent variable model for co-occurrence data that associates an unobserved class variable. PLSA also introduces a joint probability, it is a product of probability of word to topic and probability of topic to document.

d = document index
c = word's topic drawn from P(c|d)
w = word drawn from P(w|c)
Both P(c|d) and P(w|c) are modeled as multinomial distributions. The parameters can be trained with Expectation Maximization(EM) algorithm.

Compared to LSA which assumes topics are orthogonal, an advantage of PLSA is that it allows topics to be non-orthogonal.
Few limitations of PLSA are: (1) It suffers from the problem of lack of parameters for the probability distribution over documents, so we cannot determine how to assign probability to a new document. (2) Another disadvantage is that number of parameters to probability of topics to documents grows linearly with number of documents. This leads to over fitting.

David M. Blei et. al., proposes LDA, which is a widely used topic modeling technique for topic modeling in documents. A corpus (text corpus is a large and structured set of texts) of document is given to the LDA model, each document is viewed as a mixture of topics that are present in the corpus. LDA is based on bag-of-words. In LDA, it considered word as the basic unit for segmentation. Each word in the document is attributable to one of the documents topics. When LDA starts its operation, It go through each document and randomly assign each word in the document to one of K topics (K is chosen beforehand). This random assignment gives topic representations of all documents and word distributions of all the topics. To improve the topic assignments,

- For each document d, go through each word w and compute:
  p(topic t| document d): proportion of words in document d that are assigned to topic t. p(word w| topic t): proportion of assignments to topict, over all documents d, that come from word w.
- Reassign word w a new topic t, where we choose topic t with probability p(topic t | document d) * p(word w | topic t).

This generative model predicts the probability that topic t` generated word w. On repeating the last step a large number of times, the system reach a steady state where topic assignments are pretty good. Even though LDA is the most widely used topic modeling technique, it has disadvantage and advantages. The main advantage of this method is LDA is easy to understand conceptually. And disadvantages are, The number of topics k, must known in advance and dirichlet topic distribution cannot capture the correlation among topics. Compared to LSA and PLSA, LDA is easy to understand

conceptually. But the limitation is that number of topics must be known in advance and Dirichlet topic distribution cannot capture the correlation among topics.

According to William B. Cavnar et. al., N-Gram based text categorization [12], adapts the basic concepts of N-gram model. In this it uses N-word model. Text categorization is a fundamental process in document processing. It assigns the incoming document to some already existing category. N-gram is a contiguous sequence of n items from a given sequence of text. That is it is N-character slice of a given long string. It slices the string into a set of overlapping items. See an example using WORD:

eg:

bi-grams : _W , WO , OR , RD , D_

tri-grams : _ _W , WOR , ORD , RD_, D_ _

That is a string of length k, padded with blank character will have k+1 items in the slice set. in human language some words occur more frequently than other words. This idea can be expressed by Zipf0 s law stated as: The nth most common word in a text occurs with a frequency inversely proportional to n. This law indicates that there is always a set of words which dominates other words in terms of frequency of use. This is true if the words are about a particular subject. Document from the same category have similar N-gram frequency distributions. This system starts with a set of pre-existing text category. In the figure 1, Generate profile reads the incoming text and counts the occurrences of N-grams. First it eliminate digits and punctuations and other text are tokenized and are padded with sufficient blanks before and after the tokens. Then scan each token and generate N-grams, N= 1 to 5. Then it is hashed into a table to find the count. After completion it outputs all the N-grams with its count. Then sorts the output in the reverse order of frequency. It is plotted based on frequency and rank. Next process is measure profile distance, it takes two N-gram profiles and calculates the distance.

Figure 2.1: Architecture Diagram of N-gram based topic model

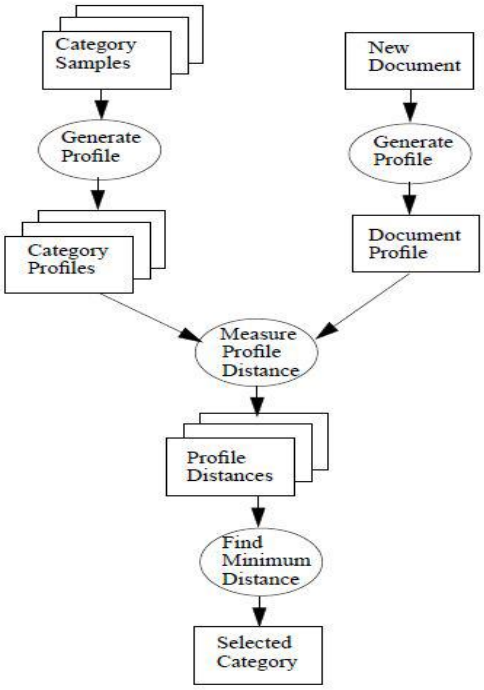
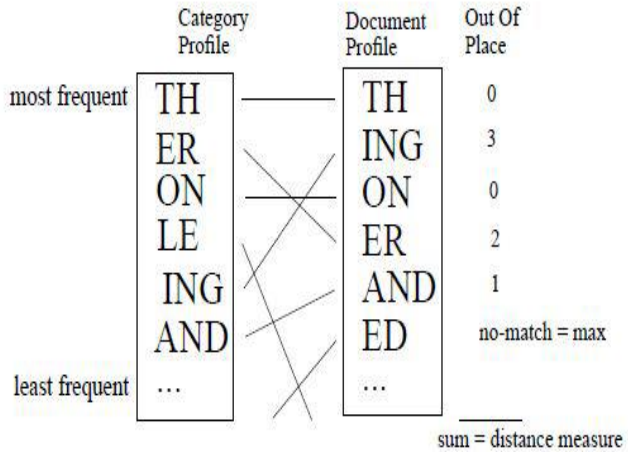



Figure 2 shows the operation for distance measure. It calculates the out-of-place measure between two profiles and summed up all the measures which gives the distance measure of two profiles. This operation is performed for all the profiles.

Figure 2.2: Calculating The Out-Of-Place Measure Between Two Profiles.




Find minimum distance simply takes the distance measure from all of the profile category to document category and determines the smallest one. This method is ideally suited for text from noisy sources like e-mail. And it has the ability to

work equally with short and long documents. The main disadvantage of this approach is that it needs a set of preexisting category. N-gram based method is best suited for text from noisy sources. It is able to work equally with short and long documents. But the disadvantage is that it needs a set of preexisting category to measure the profile distance.

According to Hearst: TextTiling : Segmenting Text into Multi-paragraph Subtopic Passages[13], is a text segmentation approach which uses topic IDs instead of words to represent n topics in the corpus. TextTiling construct topic sequence instead of sentences. Each sequence contains w topic IDs. Topic sequence is known as blocks. In this approach it uses k blocks, k is the block size. To find the similarity between two sequences called sequence gap, it uses k blocks. By using cosine similarity, similarity score is calculated. if the cosine similarity is close to 1, then it has high similarity and a value near 0 indicates the low similarity. To calculate the sharpness of the gap, depth score $d_i$ is calculated as:

$$d_i = 1/2\ (hl(i) - s_i + hr(i) - s_i) \quad (2)$$

The function hl(i) returns the highest similarity on the left side and hr(i) returns the highest similarity on the right side. Based on these depth score all local maxima points are calculated. Next step is to store these scores. If the number of segments n is given, n highest depth scores are used for segmentation. Otherwise a cut off function is used, which segments only if depth score is greater than $\mu - \sigma/2$, where $\mu$ is the mean and $\sigma$ is the standard deviation obtained from all the depth score.

Martin Riedl and Chris Biemann, proposes another method called TopicTiling[14] for text segmentation. TopicTiling is the simplified version of TextTiling algorithm. It assumes a sentence as the smallest basic unit. A coherence score $c_p$ is calculated between each position p between two adjacent sentences, which is calculated using cosine similarity. A value near to 0 indicates less similarity and a value near to 1 indicates higher similarity. In the next step, the calculated coherence score are plotted to find the local minima. These minima are used to segment the topic. But rather than using cp directly, a depth score $d_p$ is calculated for each minimum. The depth score measures the deepness of a minimum by looking at the highest coherence score on the left and right side. The depth score can be calculated as,

$$d_p = 1/2(hl(p) - c_p + hr(p) - c_p) \quad (3)$$

The function hl iterates through the left and returns highest coherence score and same is done in right side by hr. If the number of segments n is given, n highest depth scores are used for segmentation. Otherwise a cut off function is used, which segments only if depth score is greater than $\mu - \sigma/2$ , where $\mu$ is the mean and $\sigma$ is the standard deviation obtained from all the depth score. In this method the algorithm runtime is linear to the number of sentences. So compared to TextTiling, its runtime is less.

According to Zhiguo yo and Todd R John, Phrase based method is a form of topic modeling which uses phrase to represent topics. Extracting information from the large collection of document is important in many disciplines. LDA is a widely used method for topic modeling which is based on bag-of-words assumption and extension, N-gram based method is based on bag-of-N-grams are widely used methods. Phrase based method is another extension of conventional LDA, which is based on bag-of-key phrases. This method identifies semantic themes better than the previously stated methods.

***C-value method***

It is text summarization method which extracts key phrases that captures the summary of a collection of documents. in this method first it use three noun phrase regular expression filter to extract candidate phrases.

- Noun * noun
- (Adj | Noun) + noun
- ((Adj j Noun) + ((Adj j Noun) * (nounPrep)?) (Adj |Noun) * )noun

By using this, candidate phrase are extracted from the document. Then for each candidate phrase, c-value is computed based on its frequency.

**2.2 Pattern Based Representation And Filtering**

Pattern based approach uses pattern mining techniques to represent users interest. Patterns carry more semantic meaning than words. There are several approaches, which is based on pattern representation.

Yang Gao, Yue xu, Yuefeng Li, introduced two stage approach for generating topic models. In almost all topic models, the problem of word ambiguity and semantic coherence exist. To overcome these difficulties, a new method that extract more distinctive representation and discover the hidden association under multinomial word distribution are proposed. It is a two stage approach, in the first stage it generate topic representation. For topic generation LDA is used. The core idea of LDA is that every document is considered involving multiple topics and each topic can be defined as a distribution over fixed vocabulary of terms in the document. Second stage is the topic representation optimization. The topic representation generated by LDA includes single words. These individual words provides only limited information about the relationship between words. The semantic meaning of the resulted representation is also limited. To overcome this limitations, pattern based topic models are generated. First it generate a transactional dataset. From the transactional dataset, frequent patterns are generated. For a given minimal support threshold , an item set x in transactional dataset is frequent if $sup(x) \geq \sigma$. From this pattern based model, user interest U can be generated. To represent the user interest frequent patterns are used. Term based approach in this model suffers from problem of polysemy and synonymy. Low frequency of frequent pattern is another problem of this system. Advantage of this system is that it extract more distinctive representation and discover the hidden association.

Yang Gao, Yue xu, Yuefeng Li, introduced an advanced method of Two-stage Approach for Generating Topic Models to overcome the limitation of frequent pattern, which is known as PBTM. In this method to discover semantically meaningful and efficient patterns to represent topic and document. First generate a transactional dataset from the output of LDA and after that generate pattern based representation from the transactional dataset. Transactional dataset includes set of words without any duplicates. From the transactional dataset, frequent patterns are generated. For a given minimal support threshold, an item set x in transactional dataset is frequent is $sup(x) \geq \sigma$. From this pattern based model, user interest U can be generated. To represent the user interest frequent patterns are used. To represent user interest, frequent patterns are not useful since many of them are not necessarily useful. So more concise closed patterns are used. To understand the specificity of a pattern, specificity can be defined. Specificity of a pattern x can be defined as a power function of the pattern length with the exponent less than 1 denoted as,

$$Spe(x) = a|x|^m \quad (4)$$

Where a and m are constant real numbers, $0 < m < 1$.

Closed patterns are more effective and efficient to represent topics rather than frequent patterns. But only using closed patterns may affect the effectiveness of document filtering since the closed patterns may not be present in new incoming documents. To overcome this problem Yang Gao, Yue xu, Yuefeng Li proposes another method which uses maximum matched patterns. In this method first it generates the transactional dataset form the output of LDA and generates frequent patterns like in the previous methods. Then it generates a pattern equivalence class. An equivalence class is defined as, let x be a closed item set and G(x) consists of all generators of x in the transactional dataset. Frequency of patterns in the equivalence class is same and classes are exclusive to each other. Relevance of each document is computed to filter out irrelevant documents. For this significance of maximum matched patterns are calculated. A maximum matched pattern is pattern which is a proper subset of the document and subset of an item set x, which is an element in the equivalence class and subset of the document. Pattern specificity is computed to estimate the pattern significance. Pattern significance is the summation of specificity and corresponding support of the matched patterns.

## III. CONCLUSION

A survey on topic modeling in information filtering is highlighted in order to enlist the growing number of research in this area. Topic modeling uncover the hidden semantic meaning and topics in the text corpora. Traditional term based approach of topic modeling suffers from the problem of polysemy and synonymy. But pattern based methods are more effective than term based model. Maximum matched patterns provide pattern enriched topic models for information filtering. Pattern based topic modeling has growing number of applications. By understanding these methods new potential ideas can be evolved and leads to more research opportunities.

## REFERENCES


**[1]** X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval, in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 178185.

**[2]** C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles, in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2011, pp. 448456.

**[3]** D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation, J. Mach. Learn. Res., vol. 3, pp. 9931022, 2003.

**[4]** Manoj kumar, D. K yadav, Vijay Kumar Gupta. Frequent term based text document clustering: A new approach in International Conference on Soft Computing Techniques and Implementations (ICSCTI), 2015.

**[5]** Todd R J, Zhiguo Yu, Ramakanth Kavuluru. Phrase Based Topic Modeling for Semantic Information Processing in Biomedicine, in International Conference on Machine Learning and Applications, 2013.

**[6]** Hong T, TuTuoi T, PhanKhu P. An adaptive Latent Semantic Analysis for text mining, in International Conference on System Science and Engineering (ICSSE), 2017.

**[7]** T. Hofmann. Probabilistic latent semantic indexing: in Proceedings of 22nd Annu. International ACM SIGIR Conf. on Res. Develop. Inform. Retrieval, vol. 15, no. 4, pp. 368 399, 1997.

**[8]** Yang Gao, Yue Xu, Yuefeng Li, Bin Liu. A Two-stage Approach for Generating Topic Models:Advances in Knowledge Discovery and Data Mining, PADKDD13. New York, NY, USA: Springer, 2013, pp 221232.

**[9]** Yang Gao, Yue Xu, Yuefeng Li. Pattern-based Topic Models for Information Filtering: in International Conference on Data Mining Workshops, 2013.

**[10]** Yang Gao, Yue Xu, and Yuefeng Li. Pattern-based Topics for Document Modelling in Information Filtering: in ieee transactions on knowledge and data engineering, ol. 27, no. 6, june 2015.

**[11]** Ali Sekmen Akram, Aldroubi Ahmet, Bugra KokuKeaton Hamm. Matrix resconstruction: Skeleton decomposition versus singular value decomposition, in International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS), 2017.

**[12]** William B. Cavnar and John M. Trenkle. N-Gram-Based Text Categorization: in Proceedings of the Fifth USPS Advanced Technology Conference, Washington D.C., 1993.

**[13]** Martin Riedl, Chris Biemann. Text Segmentation with Topic Models, JLCL 2012 Band 27 (1) 47-69.

**[14]** Martin Riedl and Chris Biemann. TopicTiling: A Text Segmentation Algorithm based on LDA: in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 916, Las Cruces, NM, USA.

**[15]** Ning Zhong, Yuefeng Li, and Sheng-Tang Wu. Effective Pattern Discovery for Text Mining: in Proceedings of ieee transactions on knowledge and data engineering, vol. 24, no. 1, january 2012.

**[16]** M.Divya, K.Thendral, Dr.S.Chitrakala. A survey on topic modeling: in Proceedings of International Journal of Recent Advances in Engineering and Technology (IJRAET), Volume-1, Issue - 2, 2013.