International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 07 | July-2016 www.irjet.net p-ISSN: 2395-0072

# Similar Document Retrieval using Pattern-Based Topic Modelling

**NAYANA M K[1], SABEEHA K P[2]**

*[1]Mtech Student, Dept. of Computer Science and engineering, M E A Engg. College, Kerala, India*
*[2]Assistant Professor, Dept. of Computer Science and engineering, M E A Engg. College, Kerala, India*

---

**Abstract -** *Topic models can discover the hidden topical structure in document collections. Topic modelling algorithms are helping us to search, browse and summarize large archives of texts. A topic modeling is a form of text mining. It provides a way of identifying patterns in a corpus. It helps us to finding the topics that occur in a collection of documents. The topics that are discovered during topic modelling are represented by the distribution of words. A basic assumption is that the documents in the collection are all about single topic. Patterns are always more useful than single terms for describing documents. Selection of the most typical patterns from the huge amount of discovered patterns becomes crucial. A novel information filtering model is proposed here. In information filtering model user information needs are created in terms of multiple topics where each topic is represented by patterns.*

***Key Words***: **Topic Model, User Interest Modelling, Pattern Mining, Information Filtering**

## 1. INTRODUCTION

The main aim of information filtering is to remove or delete unwanted information and create user interest document. The user interest modelling is a process to understand the user's information needs based on the most relevant information that can be found and delivered to the user. In order to extract precise user's interests, traditionally, many term-based approaches [1] are used due to their efficient computational performance, as well as developed theories for term weighting, such as Rocchio, BM25, etc. But term-based approach suffer from the problems of polysemy and synonymy. Phrase-based approaches are more discriminative and should carry certain semantic meaning. However, the performance of using phrases in real applications is discouraging. The likely reasons could be phrases have inferior statistical properties to terms; and they occur in documents often with very low frequencies.

Topic modelling [2] such as Latent Dirichlet Allocation (LDA) [3] has become one of the most popular probabilistic text modelling techniques. And which has been quickly accepted by machine learning and text mining communities. The most impressive contribution of topic modelling is that it automatically classifies documents by a number of topics and represents every document with multiple topics and their corresponding distribution. The topic-based representation generated by using topic modelling can conquer the problem of semantic confusion compared with the traditional text mining techniques.

There are mainly two problems in directly applying topic modelling. First problem is limited number of topic that are predefined which are inadequate for document representation. Second problem is word model always generate frequent word set some word have meaning and some are not useful for document representation. The representation by single words with probabilistic distributions breaks the relationships between associated words. Therefore, topic modelling needs improved modelling users' interests in terms of topics' interpretations. In this work, a pattern-based topic model is proposed to enhance the semantic interpretations of topics.

The pattern based topic model can be considered as a "post-LDA" model because here patterns are constructed from the topic representation of the LDA model. When we are comparing pattern-based topic models with the word-based topic models we can analyse that the pattern-based topic model can be used to represent the semantic content of the user's documents more accurately. However, patterns in some topics can be huge and some patterns are not discriminative enough to represent specific topics.

In this work, to represent topics instead of using frequent patterns we proposed to select the most representative and discriminative patterns, which are called Maximum matched Patterns. A new topic model, called Maximum matched Pattern Based Topic Modelling (MPBTM) is proposed for document representation and document relevance ranking. The patterns in the MPBTM contained patterns are well structured so that the maximum matched patterns can be efficiently selected and used to represent and rank documents.

## 2. LITERATURE SURVEY

Two technical categories of baseline model include topic modelling methods, pattern mining methods. For the topic modelling category, the baseline models include Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Pattern-based Topic Model (PBTM). For pattern mining, the baseline models include Frequent Closed Patterns (FCP), frequent Sequential Closed Patterns (SCP) and phrases (n-Gram). An important difference between the topic modeling methods and pattern mining methods, the topic modeling methods consider multiple topics in each document in the

document collection and use patterns (e.g. PBTM and MPBTM) or words (e.g. LDA) to represent the topics, whereas the pattern mining method assume that the documents within one collection are about one topic and use patterns to represent documents directly. Literature Survey of these baseline models are given below.

### 2.1 Topic Modelling

The main aim of probabilistic topic modelling is to discover hidden topics from a large document collection. Topic modelling algorithms are used to analyze the words of the original texts to find out the themes that run through them, how those themes are connected to each other, and how they change over time.

- LSA, PLSA

In order to compress large amount of data into useful and manageable form, topic modelling is used. Latent Semantic Analysis (LSA) [4] uses a singular value decomposition of a collection, forming a reduced linear subspace. Another step to this concept is Probabilistic Semantic model, which is a generative data model. Almost all models that can be used are statistical mixture models, in which each word in a document form a mixture model, where the mixture components are multi-national random variables that can be viewed as a representation of topics. Topic modelling techniques can be generally divided into two categories, supervised and unsupervised; where bag of words and sequence words approaches are used respectively. In the field of Information retrieval, Document clustering and Summarization, uses an unsupervised bag of word technique, due to its simplicity. Whereas in the case of supervised models are used in supervised manner, using a pre-assigned labels for training set.

- LDA

LDA is probabilistic topic model which considers probability distribution functions for assigning words in a document to particular topic. The underlying instinct behind LDA is, documents are mixture of multiple topics. For example document named as computer science, can have topics such as data structure, algorithms, theory of computation, computer network etc means documents are mixture of topics. These topics are distributed over document in equal or unequal proportion. There are mainly two types of variables in LDA as hidden variables and observed variables. Observed variables are words within documents. While hidden variables describes topic structure. More precisely data arises from hidden random variables and these variables form topic structure. The process of inferring hidden structure from document is accomplished by computing posterior distribution. This distribution is conditional distribution of hidden variables in documents. The word 'Dirichlet' in Latent Dirichlet Allocation is a distribution that is used to draw per document topic distribution i.e. it specifies how topics are distributed in particular document. In generative process this output of dirichlet distribution is used to assign words of documents to different topics.

- PBTM

Yang Gao, Yue Xu and Yuefung Li, 2015, [5], proposes a two-stage model for modelling the documents in a collections. One of the main discriminative feature of this model is, it combines the data mining techniques to statistical topic modelling to generate discriminative and pattern based representations for modelling topics in documents. In the first stage, it generates word distributions over topics for documents in the collection whereas in the stage second stage ,it uses the topic representations that are generated in the first stage for representing the topics by using term weighting method and pattern mining methods. The pattern based and discriminative term based representations generated in the second stage are more accurate and efficient than the representations generated by typical statistical topic modelling method LDA. Another important feature of this representation is, patterns carry more structural and inner relationship within each topic.

### 2.2 Pattern-based Modelling

Frequent pattern is the most flexible type of pattern. Frequent pattern mining ignores order of words and allows gaps between words. With a flexible threshold, frequent pattern mining can generate variable number of patterns as needed. Using many patterns can improve the discriminative power of models. However, using too many patterns may decrease the performance of models. Frequent sequential pattern considers `order' of words. Thus frequent sequential pattern mining can give us more discriminative meaningful phrases because it can distinguish phrases whose semantic meanings changes in different order.

- FCP

H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, proposed Enriching text representation with frequent pattern mining for probabilistic topic modeling [6]. Here frequent patterns are pre-generated from the original documents and it is then added into the original documents as part of the input to a topic modelling model such as LDA. The resulting topic representations contain both individual words and pre-generated patterns. To remove redundant patterns, pattern compression methods such as closed pattern and compressed pattern are used. Pattern compression can filter out meaningless patterns and let us use only important ones.

- SCP

The work proposed by N. Zhong, Y. Li, and S.-T. Wu, shows effective pattern discovery for text mining. Author studied an efficient and effective pattern discovery method which includes the processes of pattern deploying and

pattern evolving. It is helpful to improve the effectiveness of using generated patterns for finding relevant information. In this research work, an effective pattern discovery method has been proposed to minimize the low-frequency and misinterpretation problems for text mining. In this work a Pattern Taxonomy Model is considered. There are two main stages in PTM [7]. The first stage describes how to extract useful patterns from text documents, and the second stage is then how to use these discovered patterns to improve the effectiveness of a knowledge discovery system. In PTM, first of all they divide a text document into a set of paragraphs and treat them as an individual transaction, which consists of a set of words (terms). At the subsequent phase, to find frequent patterns from these transactions apply data mining method and generate pattern taxonomies. To get relevant pattern a pruning process is applied next sequential pattern mining algorithm named SPMining is used here.

## 3. PROPOSED SYSTEM

Representation generated by pattern based LDA carry more meaning than the word based representation. A new topic model, called Maximum matched Pattern Based Topic Modelling (MPBTM) is proposed for document representation and document relevance ranking.

### 3.1 Maximum matched Pattern-Based Topic Model (MPBTM)

Maximum Matched Pattern-based Topic Model [8] consists of topic distributions describing topic preferences of each document or the document collection and pattern-based topic representations presenting the semantic meaning of each topic. Here a structured pattern-based representation is proposes and in which discovered patterns are organized into meaningful groups, called equivalence classes. The equivalence class is based on their taxonomic and statistical features. With this equivalent class, the most representative patterns can be identified which will benefit the filtering of relevant documents. In this system a new ranking method determines the relevance of recent documents based on the proposed model and especially the structured pattern-based topic representations. The maximum matched patterns are the largest patterns in each equivalence class that exist in the incoming documents, and used to evaluate the relevance of the incoming documents to the user's interest.

There are mainly two phases in this model. First one is Document training phase and second one is Document filtering phase. During the document training phase user interest modelling is done. Four steps are proposed to generate the Topic based user interest model. First Topic modelling algorithm named LDA applying to each documents. LDA results number of words under different topics. Next construct a new transactional dataset from the result of LDA, which removes duplicate words. Then from that transactional dataset mine frequent patterns by using efficient pattern mining algorithm. Pattern carry more information than single words. In the filtering stage, incoming documents are passes through topic modelling, pattern mining and finally the MPBTM selects maximum matched patterns, instead of using all discovered patterns. Then compare the incoming document pattern with training documents pattern. From that we can find out Maximum matched patterns and which is used for estimating relevance of incoming documents.

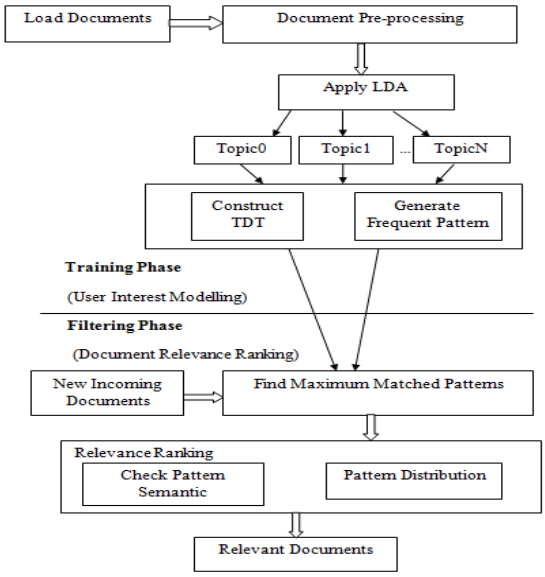


**Fig -1:** Maximum matched Pattern-based Topic Model (MPBTM)

**Table -1:** Comparison of Different Topic Model Methods

| Topic Model | Characteristics | Limitations |
|---|---|---|
| LSA | LSA can get from the topic if there are any synonym words. | It is hard to obtain and to determine the number of topics. |
| PLSA | Handles polysemy. | At the level of documents PLSA cannot do probabilistic model. |
| LDA | Need to manually remove stop words. | It is unable to find the relationships among topics. |
| Pattern based model | Represent the semantic content of the user's documents more accurately. | Many times the patterns are not discriminative enough to represent specific topics. |

## 4. Conclusion

Topic modelling is one of the most popular probabilistic text modelling techniques. Its importance is quickly accepted by machine learning and text mining communities. They are very useful in information filtering, document ranking, content-based feature extraction and modelling tasks, such as information retrieval and recommendations. The topics that are discovered during topic modelling are represented by the distribution of words. A basic assumption for these approaches is that the documents in the collection are all about single topic. But in reality this is not necessarily the case. Patterns can convey more semantic meaning than single words. Here topic modelling in the field of information filtering is mainly considered. The Maximum matched Pattern Based Topic Modelling (MPBTM) can consider multiple topics within a document. The patterns in the MPBTM are well structured so that the maximum matched patterns can be efficiently selected and used to represent and rank documents.

**REFERENCES**

[1] Ronen Feldman1, Moshe Fresko, Yakkov Kinar,Yehuda Lindell, Orly Liphstat, Martin Rajman, Yonatan Schle1, Oren Zamir, "Text Mining at the Term Level", April 2000

[2] Sawant Ganesh S. Kanawade Bhavana R., "A Review on Topic Modeling in Information Retrieval" International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume 1 Issue 1 (April 2014).

[3] N. A. Y. Blei, D. M. and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, p. 993-1022, 2013.

[4] D. L. Thomas K Landauer, Peter W. Foltz, "An introduction to latent semantic analysis," Discourse Processes, pp. 259-284, 1995.

[5] G. Y. L. Y. . L. B. Xu, Yue, "A two-stage approach for generating topic models," Advances in Knowledge Discovery and Data Mining, pp. 221-232, 2012.

[6] Hyun Duk Kim, Dae Hoon Park, Yue Lu, ChengXiang Zhai, "Enriching Text Representation with Frequent Pattern Mining for Probabilistic Topic Modeling" ASIST , 2012 October , pp. 26-31

[7] Rohini Y. Thombare, Shirish.S. Sane , "Effective Pattern Deploying Approach in Pattern Taxonomy Model for Text Mining", International Journal of Computer Applications, P. No 0975 – 8887, 2013.

[8] Yang Gao, Yue Xu, and Yuefeng Li, "Pattern-based Topics for Document Modelling in Information Filtering,"in Knowledge and Data Engineering,2015.