Middle-East Journal of Scientific Research
24 (Special Issue on Innovations in Information, Embedded and Communication Systems): 340-345, 2016
ISSN 1990-9233; © IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.IIECS.23182

A Latent Dirichlet Allocation Algorithm for Pattern-Based Topic Filtering

V. Vishnu Priya and S.K. Soma Sundaram

Department of CSE, PSNA College of Engineering
and Technology, Dindigul, TamilNadu, India

Abstract: The patterning provides a reusable architecture which speeds up many computer programs. It offers more characteristics meanings than the use of single words. Pattern based topic model can be recycled to represent the acceptable content of the user text more truthfully compared with the word based topic models. Patterns are continuously concluding to be more selective than single strings and are able to admit the inner relations between words. Pattern based topic filtering rendered by the some algorithms is imperfect to enact document, due to its partial number of dimensions. Each lexical is produced from a single topic; some other lexical in the document might be generated from different topics. Each text is represented as list of connecting proportions for the mixture of components. Topic modelling such as LDA was proposed to generate statistical model to represent numerous topic in a collection of documents. A novel information refining typical is maximum matched pattern-based topic model is suggested for filtering information needs are generated in terms of multiple topics. The success of the proposed model is finding the most relevant information to users primarily appear from its accurately acceptable delegation to represent documents and also accurate classifications of the thesis at both text matching and group level.

Key words: Filtering · Latent Dirichlet Allocation · Maximum Matched Pattern based Topic Model · User Interest Model

INTRODUCTION

Pattern mining algorithms depends on developing data mining algorithms to find out interesting, surprising and functional pattern in databases [1]. Pattern mining algorithms can be applied on various types of data such as transaction databases, sequence databases, streams, spatial data, graphs, etc. The goal is to discover all patterns whose frequency in the basis dataset exceeds a user specified threshold. Database model filtering that helps you to create mining models that use subset of data in a mining structure [2]. The Pattern is always thought to be more discriminative than single terms and are able to inner relations between words. Pattern based topic filtering used to filter out the irrelevant document and gives relevant document from the collection of documents. Since patterns carry more semantic meaning than terms. In many [3] patterns-based methods only the presence and absence of the patterns in the documents are considered. Even if the pattern occurs multiple times in the documents to be filtered equal importance is considered. Some data mining techniques have been produced to remove redundant and noisy patterns for improving the quality of the discovered patterns such as maximal patterns, closed patterns, master patterns etc., some of which have been used for representing user information needs in information filtering systems [1]. It plays an essential role in many data mining tasks discovered toward finding interesting patterns in datasets. Pattern-based representations are more meaningful and more accurately represents topics than word-based representation [2]. Next Generate Pattern Enhanced Representation, the basic idea of the proposed pattern-based method is to use frequent patterns generated from each transactional dataset to represent. In this paper, we propose to select the most representative and discriminative patterns, which are called Maximum matched patterns to represent topics instead of using frequent patterns. A new topic model, called MPBTM is proposed for document representation and document relevance ranking. The patterns in the MPBTM are well structured so that the maximum matched patterns can be efficiently and effectively selected and used to represent and rank documents.

Corresponding Author: V. Vishnu Priya, Department of CSE, PSNA College of Engineering and Technology, Dindigul, TamilNadu, India.

Related Work: Information filtering System gets user interest or user information needs based on the 'user profiles'. Information filtering systems expose users to the information that are more relevant to them. In the process of information filtering main objective is to rank the documents based on its relevance[1]. If D is the collection of incoming documents the process of information filtering is a mapping Rank(d):D_R where rank(d) represent the relevance of the document d.

Text Filtering can be considered as the document ranking process. Most popular term-based models include tf*idf,Okapi,BM5 and various weighting scheme for the bag of words representation. These models suffer from the problem of polysemy and synonymy and have the limitation of expressing semantics. so more semantic features such as phrases and patterns are extracted to represent the documents[2].Many effective algorithm like Apriori, Fp-tree, are developed to extract the frequent patterns. In many the number of these patterns will be huge to process. So more precise or relevant patterns Topic models techniques have been incorporated in the frame of language model and have achieved successfully retrieval results which have opened up a new channel to model the relevance of a document.

The research proposes an alternative approach for relevance feature discovery in text documents. It presents a method to find and classify low-level features based on both their appearances in the higher-level patterns and their specificity [4]. It also introduces a method to select irrelevant documents for weighting features. It is based on an innovative technique for finding and classifying low-level terms based on their appearances in the higher-level features and their specificity in a training set [5]. It also introduces a method to select irrelevant documents called that are closed to the extracted features in the relevant documents in order to effectively revise term weights.

A simple solution to alleviate the sparsity problem is to aggregate short text into lengthy pseudo documents before training a standard topic model. The problem is to deal with simplify the topic model by adding strong assumptions on short texts [6]. Topic models are widely used to uncover the latent semantic structure from text corpus and the effort of mining the semantic structure in a text collection can be dated from latent semantic analysis which employs the singular value decomposition to project documents into a lower dimensional space, called latent semantic space[7]. The topic model contains cluster of words with similar meanings and text, it contains different terms of topic modelling. It also include model topics with taking into account time based on user interest model and it will cofound the topic discovery. Further it has been mentioned in the some of the applications that have been in these methods [8]. The Comparison of different topic model features is essential to design a new proposal for information filtering based on user interest model. All of these models considers the time as a most vital factor.

The MPBTM model achieve the excellent performance is mainly because we creatively incorporate pattern mining techniques into topic modelling to generate pattern based topic models which can represent user interest models in terms of multiple topics. Mostly, the topics are represented by patterns which bring concrete and precise semantics to the user interest models. This is because the proposed maximum matched patterns are most representative quality patterns for modelling user's interest and relevance of documents.

Methodology

Latent Dirichlet Allocation: A Latent Dirichlet Allocation is a powerful learning algorithm for automatically joint clustering words into "topics" and documents into mixture of topics. It is a generative model that allows set of observations to be explained by unobserved groups that explain why some parts of the data are similar [1]. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. In LDA, each document may be viewed as a mixture of topics. For example, an LDA model might have topics that can be classified as Computer Desktop_related and Network_related [9]. A topic has probabilities of generating various words, such as monitor, keyboard and mouse, which can be classified and interpreted by the as "Computer Desktop_related". Naturally the word Computer Desktop itself will have high probability given this topic [10]. A lexical word may occur in several topics with a different probability, however with a different typical set of neighbouring words in each topic [11]. Each document is assumed to be characterized by a particular set of topics as shown in the Figure 3.1. This is similar character to the standard bag of words model assumption and makes the words exchangeable.

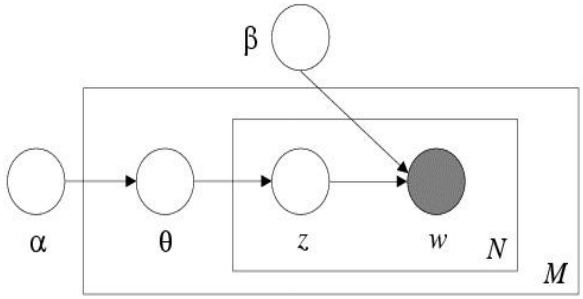

Fig. 1: Graphical Model Representation Of LDA

In this Latent Dirichlet Allocation the graphical model representation of LDA in these boxes are plate representing replicates [12]. The Outer plates represent the documents and the inner plates represents the repeated inner choice of topics within a documents. This model is based on following notations and terminology.

- A word is the basic unit of discrete data, defined to an item to from a vocabulary indexed by{1...v}.We represents words using unit-basis vectors that have a single component equal to one and all other components equal to zero.
- A document is a sequence of N words denoted by a w={w1,w2,w3...wN}
  A corpus is a collection of M documents denoted by D={ w1,w2..wM}

Algorithm

Input*:* a collection of positive training documents D;
minimum support sj as threshold for topic Zj; number of topics V
Output*:* U_E={E(Z1),…..E(ZV)}
1: Generate topic representation f and word-topic assignment zd;i by applying LDA to D
2: U_E:=□
3: for each topic Zj _ [Z_l,Z_v] do
4: Construct transactional dataset $+_j$ based on □ and zd,i
5: Construct user interest model X_Zj for topic Z_j using a pattern mining technique so that for each pattern X in X_{Zj}, supp(X) > □j
6: Construct equivalence class EðZjÞ from XZj
7: U_E:=U_E □{E(Z_j)}
8: end for

Topic Modelling: A Topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections [6]. Topic models provide a convenient way to analyze large of unclassified text. A topic contains a cluster of words that frequently occur together. A topic model contains a collection of text as input it discovers a set of "topics" recurring themes that are discuss in the collection of documents. A topic modeling can connect words with similar meanings and distinguish between uses of words with multiple meanings[13]. So, the idea of topic models is that term which can be working with documents and these documents are mixtures of topics, where a topic is a probability distribution over words. In other word, topic model is a generative model for documents. It specifies a simple probabilistic procedure by which documents can be generated [14]. It creates a new document by choosing a distribution over topics. After that, each word in that document could choose a topic at random depends on the distribution.

User Interest Model: A User model represents a collection of personal data associated with a specific user. Therefore, it is the basis for any adaptive changes for any system behaviour which data is included in the model depends on the purpose of the application. It can include personal information such as user's name, id, password and e-mail id [15]. There are different design patterns for user model they are classified as static and dynamic user models. Static user models are the most basic kind of user models [16]. User interest model is based on the user interest in this user is to select the particular topic, according to this topic modelling there are different types of topics are available in the documents. By User Satisfaction the topic modelling was modelled from the documents and the values are identify and calculated.

Relevance Ranking: Relevance Ranking is based on the ranking based method, it describes about the probability ratio. Relevance is denoted as how well a retrieved document or set of documents meets the information needs. Ranking criteria are phrased in terms of relevance of documents with respect to an information needs. It is often reduced to the computation of numeric query [4].

Maximum Matched Patterns: In the Maximum Matched Patterns which represent user interests are not only grouped in terms of topics, but also partitioned based on equivalence class in each topic group. The patterns in different group or different equivalence classes have different meanings and distinct properties. Thus, user information needs are clearly represented according to

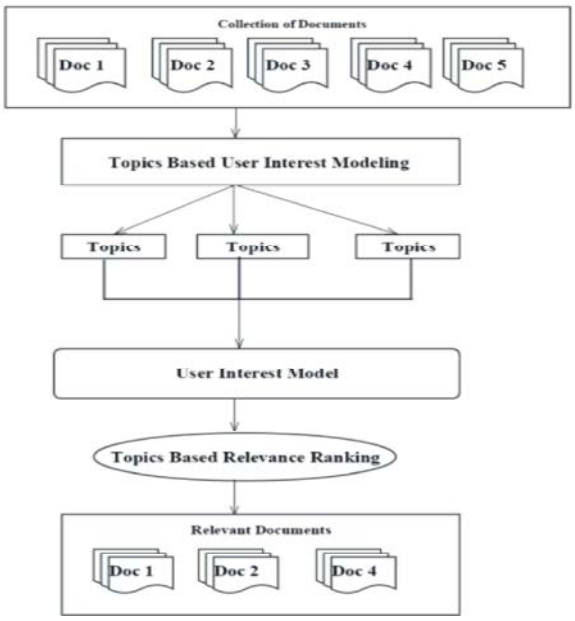

Fig. 2: Architecture Diagram

various semantic meanings as well as distinct properties of the specific patterns in different topic group and equivalence classes.

In Collection of documents there are different numbers of documents are available it is used to list out the more than hundred number of documents. These documents are classified and analyzed with the user interest model [10]. User interest model are based on the selection process and it is used to based on the more number of topics that are carry out the topics based representation. These topics are spitted with the frequent patterns. Patterns are enhanced with the more number of probability ratio [1]. Relevant documents are based on the values of identifying the probability ratio, term weight. Irrelevant documents are identified based on the number of documents and the hidden topics are identified and filter out the unwanted information as mentioned in the Figure 2.

Experimental Results: The Statistical topic modeling technique has attracted great attention due to its robust and interpretable topic representations. The most popular used topic modeling method is LDA and its various extensions are each document is a mixture of topics. Each topic is represented by distribution of words.

In the Table I represents the common words in different topics procedure ambiguous meaning across of topics. Single words are not discriminative enough to represent the meaning of topics.

Table I: Topics in Word Assignment

Topic 0		Topic 10		Topic 11	
String	Time	String	Time	String	Time
Method	0.043	Data	0.437	Method	0.072
Sample	0.040	Mine	0.062	Weight	0.028
High	0.024	Real	0.039	Salary	0.025
Gene	0.023	Value	0.030	Variety	0.025
District	0.031	Word	0.09	Recent	0.023

Table II Example Results of LDA

Topic	Z1		Z2	
Document	*value	Words	*Value	Words
D1	0.6	w1,w2,w3,w2,w1	0.2	w1,w9,w8
D2	0.2	w2,w4,w3	0.5	w7,w8,w2

Table III: Topic Document Transactions

Transaction	TopicDocumentTransaction
1	{w1,w8,w9}
2	{w1,w7,w8}
3	{w2,w3,w7}

Table IV: Pattern Enhancements

Patterns	Support
{w1},{w8},{w1,w8}	3
{w9},{w7},{w8,w9},{w1,w9}	2

The results of LDA in the Table II is based on the probability value and the word topic assignment statement are assigned due to the value of words are into different patterns..

In the Table III is used to construct the transactional data set and it converts generates pattern based topic representation.

In the Table IV represents the patterns are enhanced with the frequent number of support value and confidence and it is used to calculate the values to determine the sequences.

Dataset: The Reuters Corpus volume1 (RCV1) dataset was collected by Reuter's journals between August 20 1996 and August 9,1997, a total of 806,791 documents that cover a variety of topics and a large amount of information [1]. Each collection is divided into a training set and a testing set. In TREC track, a collection is referred to as a 'topic'. In this Section, to differentiate from the 'topic' in LDA model, ' collection' is used to refer to a collection of documents in the TREC dataset.

Measures: The effectiveness is assessed by five different measures: average precision(K=20) documents, F(B=1) measure, Mean Average Precision (MAP), break-even

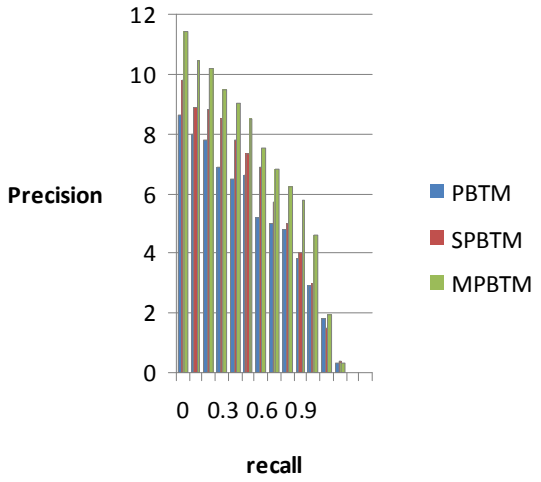

Fig. 3: Comparison between the proposed pattern basedmethod..

Table V: Comparison between three Methods

Methods	Top 20	b/p	MAP	F1
MPBTM [17]	0.494	0.420	0.426	0.424
PBTM [16]	0.447	0.410	0.235	0.432
SPBTM [16]	0.332	0.208	0.212	0.348

point(b/p)and Interpolated Average Precision (IAP) on 11 points.F1 is a criterion that assess the effect involving both precision and recall levels(i.e.,recall=0,0.1.....,1).

Result of MPBTM: The results of the MPBTM, which includes the maximum matched patterns in the topics it describe about the words topic assignment results of the baseline model of the three categories i.e.,topic-based, pattern-based and term-based. It filter out the irrelevant documents and gives the accurate results.

The Pattern Based method is describes the methods of MPBTM,PBTM,SPBTM are Compared with the three methods based on top 20, b/p, MAP,F1,Which gives the MPBTM method is the highest priority value Compared to the others.

The maximum matched pattern model is based on the higher probability value based on the support count. It is used to describe the maximum matched patterns only. The others Pattern based topic model gives the high priority.

CONCLUSION

The problem of filtering in pattern based have been studied and hence proposed a system will filter out irrelevant document and gives relevant document, accuracy to the time based information filtering. To enable this method the latent dirichlet algorithm is used and it is based on the levels of devices it is used for the time accuracy and easy to implement and find out the topics in easily manner. By using the document can be spitted into different number of topics and these topics are spitted into different types according to the user based interest model. It is used to finding the high values from probability ratio, it gives the term weight value and support and confidence based on mining method. This method can be applied to real time system to finding out the high relevant topics in whole of the documents and it is to be considered as a high document in this system. Since the pattern based filtering is divided into the pattern enhancement process and it is used to reduce the more number of irrelevant words in a particular topic and it is necessary to use the document in one or more patterns, filtering requires the future based method to improve the documents in high probability value. The above proposed method is used only for documents eg: notepad files,.etc. and it is based on the number of documents are available in the dataset level or filter due to the number of words available in the one text files.

REFERENCES

1. Gao Yang and Yuefeng Li, 2015. Pattern-based Topics For Document Modeling in Information Filtering, IEEE Transaction On Knowledge and Data Engineering, 27(6): 1629-1642.
2. Ning Zhong, *et al*., 2012. Effective Pattern Discovery for Text Mining, IEEE Transaction On Knowledge and Data Engineering, 24(1): 30-44.
3. Klaus Moessner, *et al*., 2014. Probabilistic Matchmaking Methods For Automated Service Discovery, IEEE Transaction On Automated Service Discovery, 7(4): 654-665.
4. Abdulmohsen Algarni, *et al*., 2015. Relevance Feature Discovery for Text Mining, IEEE Transaction On Knowledge and Data Engineering, 27(6): 1656-1669.
5. Laxmidhar Behera, *et al*., 2013. A Context-Based Word Indexing Model For Document Summarization, IEEE Transaction On Knowledge and Data Engineering, 25(8): 1693-1705.
6. Lan Yanyan, *et al*., 2014. BTM: Topic Modeling over Short Text, IEEE Transaction On Knowledge and Data Engineering, 26(12): 2928-2941.
7. Odysseas Papapetrou, *et al*., 2012. Decentralized Probabilistic Text Clustering, IEEE Transaction On Knowledge and Data Engineering, 24(10): 1848-1861.

8. Soma Sundaram, S.K. and V. Vishnupriya, XXXX. A Survey on Topic Modeling Methods Over Information Filtering 'International Journal Of Innovative Research in Computer and Communication Engineering, 3(10): 9915-9922.
9. Qing He, *et al*., 2012. Mining Distinction and Commonality across Multiple Domains Using Generative Model For Text Classification, 24(11): 2025-2039.
10. Newman David, *et al*., 2012. Understanding Errors in Approximate Distributed Latent Dirichl*et al*location, IEEE Transaction On Knowledge and Data Engineering, 24(5): 952-960.
11. Welly Naptali, *et al*., 2012. Topic-Dependent –Class Based n-Gram Language Model, IEEE Transaction On Audio and Language Processing, 20(5): 1513-1525.
12. Ruizhang Huang, *et al*., 2013. Dirichlet Process Mixture Model For Document Clustering with Feature Partition, 24(6): 1748-1759.
13. Haidong Gao, *et al*., 2015. Probabilistic Word Selection via Topic Modeling, IEEE Transaction On Knowledge and Data Engineering, 27(6): 1643-1655.
14. Ostendorf Mari, *et al*., 2014. Learning Phrase Patterns For Text Classification, IEEE Transactions On Audio, Speech and Language Processing, 21(6): 1180-1190.
15. Banchs Rafael, E., *et al*., 2015. Decoupling Word-Pair Distance and Co-occurrence Information For Effective Long History Content Language Modeling, IEEE/ACM Transaction On Audio, Speed and Language Processing, 23(7): 1221-1232.
16. Lee Changhyun, *et al*., 2013. UTOPIAN: User-Driven Topic Modeling Based On Interactive Nonnegative Matrix Factorization, IEEE Transactions On Visualization and Computer Graphics, 19(12): 1992-2001.
17. Hanqi Guo, *et al*., 2014. FLDA : Latent Dirichlet allocation Based Unsteady Flow Analysis',. IEEE Transactions On Visualization and Computer Graphics, 20(12): 2545-2554.