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

A Clustering Based Collaborative and Pattern based Filtering approach for Big Data Application

Prachi Pardeshi[1], Komal Patil[2], Priyanka Patil[3],Komal Chavan[4]

[1] *Student, Computer Department, KKWIEER, Maharashtra, India*
[2] *Student, Computer Department, KKWIEER, Maharashtra, India*
[3] *Student, Computer Department, KKWIEER, Maharashtra, India*
[4] *Student, Computer Department, KKWIEER, Maharashtra, India*

Abstract - *Number of services is emerging on the Internet because of service and cloud computing. Result of this is, service-relevant data become too big to be effectively processed by traditional approaches. The most fundamental challenge for the Big Data application is to explore the large volume of data and extract useful information or knowledge for future action. Many mature term-based or pattern-based approaches have been used in the field of information filtering to generate user's information needs from a collection of documents. In this approaches they refer only single topic to finding user needs. But, in reality users interests can be diverse and the documents in the collection often involve multiple topics. Topic modeling was proposed to generate statistical models to represent multiple topics in a collection of documents and this has been mostly utilized in the fields of machine learning and information retrieval. But it is not gives effectiveness result. In general pattern refers more discriminative than single word. However, the enormous amount of discovered patterns hinder them from being effectively and efficiently used in real applications, therefore, selection of the most discriminative and representative patterns from the huge Amount of discovered patterns becomes crucial. To deal with the above mentioned limitations and problems, in this paper, a useful information filtering model, Maximum matched Pattern-based Topic Model (MPBTM), is proposed. The result shows that the proposed model significantly outperforms both state-of-the-art term-based models and pattern-based models.*

Key Words:
Topic modeling, Information filtering, pattern mining, relevance ranking, user interest model, clustering.

1. INTRODUCTION

Big data has gain large popularity and attracting attentions from government, industry and academic. Big Data concerns large-volume, complex, growing data sets with multiple autonomous sources. Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within an efficient elapsed time. The most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information from that dataset for future references. INFORMATION filtering (IF) is a system which is used to remove unwanted data and redundant data from given document based stream which represent user interest. Traditional IF models were used term based approaches for filtering the data. The advantage of these approaches is its efficient computational performance, But it is suffers from the problems of polysemy and synonymy. To overcome the limitations of term-based approaches, pattern mining based techniques are used. Pattern based technique have been used to utilize patterns to represent users' interest. It is also achieved some improvement in terms of patterns which is represent user interest more efficiently than single word. Also, some data mining techniques have been developed to improve the quality of patterns (i.e. maximal patterns, closed patterns and master patterns) for removing the redundant and noisy patterns. But, all these data mining and text mining techniques hold the assumption that the user's interest is only related to a single topic. However, in reality this is not necessarily the case. The user interest can be change or diverse continuously according to situation and demand of services. Therefore, in this paper, we propose to model users' interest in multiple topics rather than a single topic, which reflects the dynamic nature of user information needs.

Topic modeling has become one of the most popular 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. Two representative approaches are Probabilistic Latent Semantic Analysis (PLSA) and LDA. In this, there are two problems in directly applying topic models for information filtering. The first problem is that limited number of dimension i.e., a pre-specified number of topics and second problem is that each topic in topic model is represented by set of words. The patterns can well represent the topics because the patterns are comprised of the words which are extracted by LDA based on sample

occurrence and co-occurrence of the words in the documents. The patterns can represent more specific meanings than single words, the pattern-based topic models can be used to represent the semantic content of the user's. Documents more accurately compared with the word-based topic models. In this, 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. 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 effectively and efficiently selected and used to represent and rank documents.

The main feature or importance of MPBTM in IF system:
1) We use MPBTM model for information search, in this we are performing search on the basis of multiple topics because in real time system user information interest can be diverse continuously.
2) In this we are integrating data mining technique with topic modeling technique to generate pattern based topic model to represent document collection. In this topic representation, it is gives information of the semantic meaning of each topic.
3) After pattern mining we are formed equivalence classes of that pattern based on their taxonomy statistical feature. In this each pattern have same frequency and represents similar semantic meaning. So we are taking the benefits of this for filtering of relevant document.
4) In this we use a new ranking method to find the relevance of new documents. The maximum matched patterns, which are the largest patterns in each equivalence class that exist in the incoming documents, are used to calculate the relevance of the incoming documents to the user's interest

In Section 2, we discuss the related work about some state-of-the-art IF models and related techniques. Section 3 provides a brief introduction of LDA. Sections 4 and 5 present the details of our proposed model. Then, experiments on the proposed model and baseline models have been conducted on a popular benchmark data collection in Section 6. According to the experimental results, we discuss the strengths of the proposed model from different perspectives in Section 7. Specifically, compared with, we conduct more baseline models and discuss further benefits of our proposed model. Finally, Section 8 concludes the whole work and presents ideas for future work.

2. RELATED WORKS :

User profiles are used to extract user need by IF system. IF system is support the long term information need of a particular user or group of user. Main objective of IF system is that provide accurate and efficient mapping of incoming document to a space of user relevant documents. In each dataset, denoting the space of incoming documents as D, the mapping _rank: D -> R such that rank (d) corresponds to the relevance of a document d. The filtering track in the TREC data collection was to measure the ability of IF systems to separate relevant from irrelevant documents. The main task of document filtering can be regarded as a classification task or a ranking task. Methods such as Naive Bayes, kNN and SVM, assign binary decisions to documents (relevant or irrelevant) as a special type of classification. The relevance of a document can be modeled by various approaches that primarily include a term-based model, a pattern-based model a probabilistic model and a language model.

Term-based models have a many limitation on expressing semantics and problems of polysemy and synonymy. Therefore, people tend to extract more semantic features (such as phrases and patterns) to represent a document in many applications. Data mining techniques were applied to text mining and classification by using word sequences as descriptive phrases (n-Gram) from document collections. But the performance of n-Gram is restricted due to the low frequency of phrases. Pattern mining has been extensively studied for many years. A variety of efficient algorithms such as Apriori, PrefixSpan, and FP-tree have been proposed and extensively developed for mining frequent patterns more efficiently. But normally, the result of this approaches is number of returned patterns is huge because if a pattern is frequent, then each of its sub patterns is frequent too. Thus, selecting reliable patterns from given number of patterns is always very crucial. For example, a number of frequent item sets have been proposed such as closed item sets, maximal item sets, and free item sets, disjunction-free item sets etc. The primary purpose of these representations is to enhance the efficiency of using the generated frequent item sets without losing any information. Among these proposed item sets, frequent closed patterns show great potential for representing user profiles and documents. That is mainly because for a given support threshold, all closed patterns contain sufficient information about all that is involved in all corresponding frequent patterns.

Topic models techniques have been incorporated in the frame of language model and have achieved successful retrieval results which have opened up a new channel to model the relevance of a document. The LDA based document models are state-of-the-art topic modeling approaches. Information retrieval systems based on these models have achieved good performance. So we are

achieved retrieval performance , was not only because of the multiple topic document model, but also because each topic in the topic model is represented by a group of semantically similar words, which solves the synonymy problem of term based document models.

Probabilistic topic modeling can also extract long-term user interests by analyzing content and representing it in terms of latent topics discovered from user profiles. The relevant documents are determined by a user-specific topic model that has been extracted from the user's information needs. These topic model based applications are all related to long-term user needs extraction and related to the task of this paper. But, there is a lack of explicit discrimination in most of the language model based approaches and probabilistic topic models. This weakness indicates that there are still some gaps between the current models and what we need to accurately model the relevance of a document.

3. LATENT DIRICHLET ALLOCATION

Topic modeling algorithms are used to determine a set of hidden topics from collections of documents, where a topic is represented as a set of the words. Topic models provide a low-dimensional representation of documents (i.e. with a limited and manageable number of topics). LDA is a statistical topic modeling technique and the most common topic modeling tool currently in use. It can search the hidden topics in collections of documents using the words that appear in the documents. Let D = {d1, d2, d3, dm} be a collection of documents. The total number of documents in the collection is M. The idea behind LDA is that each document is considered to contain multiple topics and each topic can be defined as a distribution over fixed words that appear in the documents. The resulting representations of the LDA model are at two levels, document level and collection level. At document level, each document di is represented by topic distribution Pdi= (Vd1,Vd2,.....Vdn} where,V is the number of topics. Apart from these two levels of representations, the LDA model also generates word-topic assignments, that is, the word occurrence is considered related to the topics by LDA.

Phrases are less ambiguous than words, they have been widely explored as text representation for text retrieval, but few studies in this area have shown improvements in effectiveness. The likely reasons for the discouraging performances include: (1) low occurrences of phrases in relevant documents; and (2) lack of a flexible number of words for a set of discovered phrases, which restricts the semantic expression.

The topic representation using word distribution and the document representation using topic distribution are the most important contributions provided by the LDA model. The topic representation indicates which words are important to which topic and the document representation indicates which topics are important for a particular document. Given a collection of documents, the LDA can learn topics and decompose the documents according to the topics. However, single word based topic representations contain ambiguous semantics. Thus, MPBTM improves the LDA model by expanding word-based topic representation to phrase-based, which enhances the explicit semantics of topics.

TABLE 1
Example Results of LDA: Word-Topic Assignments

Topic	Z_1		Z_2		Z_3	
Document	$\theta_{d,1}$	words	$\theta_{d,2}$	words	$\theta_{d,3}$	words
d_1	0.6	w_1, w_2, w_3, w_5, w_7	0.2	w_1, w_3, w_8	0.2	w_7, w_{10}, w_{10}
d_2	0.2	w_2, w_4, w_4	0.5	w_7, w_8, w_2, w_8, w_8	0.3	w_1, w_{11}, w_{12}
d_3	0.3	w_2, w_3, w_7, w_5	0.3	w_7, w_3, w_2, w_2	0.4	w_4, w_7, w_{13}, w_{11}
d_4	0.3	w_2, w_7, w_8	0.4	w_8, w_8, w_2	0.3	w_1, w_{11}, w_{10}

4. PATTERN ENHANCED LDA

Pattern-based representations are considered more meaningful and more accurate to represent topics than word based representations. Moreover, pattern-based representations contain structural information which can give the association between words. In order to discover meaningful patterns to represent topics and documents, two steps are proposed: firstly, construct a new transactional dataset from the LDA model results of the document collection D; secondly, generate pattern-based representations from the transactional dataset to represent user needs of the collection D.

4.1 Construct Transactional Dataset

Let Rdi,Zj represent the word-topic assignment to topic Zj in document di. Rdi,Zj is a sequence of words assigned to topic Zj. For the example illustrated in Table 1, for topic Z1 in document d1, Rdi, Zj = {w1, w2, w3……, wn}. We construct a set of words from each word-topic assignment Rdi;Zj instead of using the sequence of words in Rdi,Zj, because for pattern
mining, the frequency of a word within a transaction is insignificant. Let Iij be a set of words which occur in Rdi,Zj, in this Iij contains the words which are in document di and assigned to topic Zj by LDA. Iij, called a topical document transaction, is a set of words without any duplicates.

TABLE 2
Transactional Datasets Generated from Table 1
(Topical Document Transaction (TDT))

T	TDT	TDT	TDT
1	$\{w_1, w_2, w_3\}$	$\{w_1, w_8, w_9\}$	$\{w_7, w_{10}\}$
2	$\{w_2, w_4\}$	$\{w_1, w_7, w_8\}$	$\{w_1, w_{11}, w_{12}\}$
3	$\{w_1, w_2, w_5, w_7\}$	$\{w_2, w_3, w_7\}$	$\{w_4, w_7, w_{10}, w_{11}\}$
4	$\{w_2, w_6, w_7\}$	$\{w_1, w_8, w_9\}$	$\{w_1, w_{11}, w_{10}\}$
	Γ_1	Γ_2	Γ_3

4.2 Generate Pattern Enhanced Representation

The basic idea behind the pattern-based method is to use frequent patterns generated from each transactional dataset Gj to represent Zj. In the two-stage topic model, frequent patterns are generated in this step. For a given minimal support threshold s, an item set X in Gj is frequent if supp(X)>=∂, where supp(X) is the support of X which is the number of transactions in Gj that contain X. The frequency of the itemset X is defined as supp(X)/Gj. Topic Zj can be represented by a set of all frequent patterns, denoted as XZi = {Xi1,Xi2,….,Ximi} , where mi is the total number of patterns in XZi and V is the total number of topics. Take G2 in Table 2 as an example, which is the transactional dataset for Z2. For a minimal support threshold s ∂= 2, all frequent patterns generated from G2 are given in Table 3.

5. INFORMATION FILTERING MODEL BASED ON PATTERN ENHANCED LDA

Number of patterns in some of the topics can be huge and many of the patterns are not discriminative enough to represent specific topics. As a result, documents cannot be accurately represented by these topic representations. That means, these pattern-based topic representations which represent user interests may not be sufficient or accurate enough to be directly used to determine the relevance of new documents to the user interests. In this section, one novel IF model, MPBTM, is proposed based on the pattern enhanced topic representations. The proposed model consists of topic distributions describing topic preferences of documents or a document collection and structured pattern-based topic representations representing the semantic meaning of topics in a document.

5.1 Pattern Equivalence Class

The number of frequent patterns is considerably large and many of them are not necessarily useful. The number of these patterns is significantly smaller than the number of frequent patterns for a dataset. In particular, the closed pattern has drawn great attention due to its attractive features.

Definition 1 (Closed Item set). For a transactional dataset, an item set X is a closed item set if there exists no item set X' such that X is belong X',

Definition 2 (Generator). For a transactional dataset G, let X be a closed item set and T(X) consists of all transactions in G that contain X, then an item set g is said to be a generator of X if g belong X,T(g)=T(X) and supp(X)=SUPP(g).

Definition 3 (Equivalence Class). For a transactional dataset G, let X be a closed item set and G(X) consist of all generators of X, then the equivalence class of X in G, denoted as E(X).

Is defined as EC(X) =G(X) Ù{X}

5.2 Topic-based User Interest Modeling

For a collection of documents D, the user's interests can be represented by the patterns in the topics of D. As discussed in Section 3, us represents the topic distribution of D and can be used to represent the user's topic interest distribution, Pedi= (Vd1,Vd2,…..Vdn} V Þ, and V is the number of topics. In this paper, the topic distribution in the collection D is defined as the average of the topic distributions of the documents in D.

By using the method described in section 4, for a document collection D and V pre-specified latent topics, from the results of LDA to D, V transactional datasets, G1; . . . ; GV can be generated from which the pattern-based topic representations for the collection can be generated, where each topic is a set of frequent patterns generated from transactional dataset Gi. U is considered the user interest model, the patterns in each XZi represent what the user is interested in terms of topic Zi. As mentioned before, normally, the number of frequent patterns generated from a dataset can be huge and many of them may be not useful A closed pattern reveals the largest range of the associated terms. It covers all the information that its subsets describe. Closed patterns are more effective and efficient to represent topics than frequent patterns. However, only using closed patterns to represent topics may impact the effective since closed patterns often may not exist in new incoming documents. On the other hand, frequent patterns can be well organized into groups based on their statistics and coverage. effectiveness of document filtering .

5.3 Topic-based Document Relevance Ranking

All the pattern which are in the equivalence class are same. But there is difference in their size. So, the longer size pattern and shorter size pattern in the same equivalence class then smaller size is covered by longer one so it is given less significant. In filtering stage, relevance document is obtained by filter outing irrelevance data. We first find out maximum match pattern based on user interest model and estimate relevance of document. Thus, significance of pattern is based upon its size which is its specificity level and thus pattern taxonomy exists .

In a pattern taxonomy, the longer a pattern is, the more specific it is. As a result, the specificity of a pattern can be estimated as a function of pattern length. For example, a single word 'mining' usually represents the '-ing' form of 'mine' and it has a general meaning indicating any kind of 'prospecting', whereas 'pattern mining' represents a specific technique in data mining. 'Closed pattern mining' is even more specific but still in the same technique area. Generally, the specificity is not necessarily linearly increasing as the pattern size increases. Based on our experimental results, the increase in the specificity of a pattern should be slower than the increase in the pattern size.

Definition 4 (Pattern specificity). The specificity of a pattern X is defined as a power function of the pattern length with the exponent less than 1, denoted as spe(X), spe(X)=a|X|^m, where a and m are constant real numbers and 0 < m < 1.

Definition 5 (Topic Significance). Let d be a document, Zj be a topic in the user interest model, PAdjk be a set of matched patterns for topic Zj in document d, k=1. . . ; nj, and fj1; . . . ; fjnj be the corresponding supports of the matched patterns, then the topic significance of Zj to d is defined as:

Definition 6 (Maximum Matched Pattern). Let d be a document, Zj be a topic in the user interest model, ECj1; . . . ; ECjnj be the pattern equivalence classes of Zj, then a pattern in d is considered a maximum matched pattern to equivalence class ECjk, denoted as MCdjk, if the following conditions are satisfied:
[$MCjkd \subseteq d$ and $MCjkd \in ECjk$] here {$\nexists X$ *such that* $X \in ECjk, X \subseteq d$ *and* $MCjkd \subseteq X$}
The maximized matched pattern MCjk to equivalence class ECjk must be the largest pattern in ECjk which is contained in d and all the patterns in ECjk that are contained in d must be covered by MCdjk. Therefore, the maximum matched patterns MCd jk, where k=1; . . . ; nj are considered the most significant patterns in d which can represent the topic Zj. For an incoming document d, we propose to estimate the relevance of d to the user interest based on the topic significance and topic distribution. The document relevance is estimated using the following equation:
Rank(d)=Σ$sig(zj,d)Vj$=1×VD,j
Higher Rank$\delta(x,d)$={1,$if\ x$∈00,$otherwise$

5.4 Algorithms

Algorithm 1: Clustering
Description: In this algorithm we are taking large amount of data and then formed cluster of that data which is based on their services.
Input: Big data
Output: clusters (c1, c2, .cn)
1. Stem the words in and using Porter Stemmer. The stemmed words in are put into and the stemmed words in are put into.
2. Compute D_sim(st,sj) and F_sim(st,sj) using Jaccard similarity coefficient respectively.
3. Compute c_sim(st,sj) by weighted sum of D_sim(st,sj) and F_sim(st,sj). Create a matrix D each entry of which is a characteristic similarity.
4. Cluster services according to their characteristics similarities in D using an agglomerative hierarchical clustering algorithm.

Algorithm 2: User Profiling
Description: In user profiling we are take transactional dataset as input and then making pattern and equivalence classes of that pattern which is based on multiple topics.
Input: A collection of positive training documents D; Minimum support as threshold for topic; Number of topics V
Output: User Profile
1. Generate topic representation and word-topic assignment by applying LDA to D.
2. **for** each topic **do**
3. Construct transactional dataset based on and
4. Construct user interest model for topic using a pattern mining techniques so that for each pattern X in
5. Construct equivalence class from

Algorithm 3: Document filtering and collaborative filtering
Description: In document filtering us are finding maximized matched pattern in the equivalence classes and ten ranks to that pattern according their frequency.
Input: user interest model, a list of incoming document
Output: filtered document
1: for each do
2: for each topic do
3: for each equivalence class do
4: Scan and find maximum matched pattern which exists in update
5: end for
6: end for

7: end for

6. Evaluation

Two hypotheses are designed for verifying the IF model proposed in this paper. The first hypothesis is given that user information needs involve multiple topics, then document modelling by taking multiple topics into consideration can generate more accurate user models to represent user information needs. The second hypothesis is that the proposed maximum matched patterns are more effective than other patterns to be used in determining relevant documents.

6.1 Data

The Reuters Corpus Volume 1 (RCV1) dataset covers a variety of topics and a large amount of information. 100 collections of documents were developed for the TREC filtering track. Each collection is divided into a training set and a testing set. According to Buckley and others, the 100 collections are stable and sufficient enough for high quality experiments. In the TREC track, a collection is also referred to as a 'topic'. In this paper, to differentiate from the term 'topic' in the LDA model, the term 'collection' is used to refer to a collection of documents in the TREC dataset. The first 50 collections were composed by human assessors and another 50 collections were constructed artificially from intersections collections.

6.2 Baseline Models and Settings

The experiments were conducted extensively covering all major representations such as terms, phrases and patterns in order to evaluate the effectiveness of the proposed topic based IF model. The evaluations were conducted in terms of three technical categories: topic modelling methods, pattern mining methods and term-based methods.

(1) Topic modelling based category

- PLSA_word and LDA_word

In the word-based topic model [11], [12], words associated with different topics are used to represent user interest needs and word frequency is used to represent topic relevance.

- TNG

In the phrase-based topic model, n-gram phrases that are generated by using the TNG model introduced in are used to represent user interest needs and phrase frequency is used to represent topic relevance.

- PBTM

In, two PBTM models have been proposed and they use frequent patterns (FP) and FCP, respectively, denoted as PBTM_FP and PBTM_FCP. The frequent patterns and the frequent closed patterns associated with different topics are used to represent user interest needs and the pattern support is used to represent topic relevance.

(2) Pattern-based category

- FCP

Frequent closed patterns are generated from the documents in the training dataset and used to represent the user's interests. The minimum support in the pattern-based models, including the following two models for sequential closed patterns and phrases, is set to 0.2.

- Sequential Closed Pattern (SCP)

The Pattern Taxonomy Model is one of the state-of-the art pattern-based models. It was developed to discover sequential closed patterns from the training dataset and rank incoming documents in the filtering stage with the relative supports of the discovered patterns that appear in the documents. In this model, every document in the training dataset (D) is split in paragraphs which are the transactions for pattern mining.

- n-Gram

Most researches on phrases in modelling documents have employed an independent collocation discovery module. In this way, a phrase with independent statistics can be indexed exactly as a word-based representation. In our experiments, we use n-Gram phrases to represent a document collection (i.e. user information needs), where n is empirically set to 3.

(3) Term-based category

- BM25

BM25 [1] is one of the state-of-the-art term-based document ranking approaches. In this paper, the term weights are estimated using the following equation:

$$W(t) = \frac{tf \times (k+1)}{k \times \left((1-b) + b\frac{DL}{AVDL}\right) + tf} \times \log\left(\frac{N-n+0.5}{n+0.5}\right),$$

TABLE 6
Comparison of All Models on All Measures Using the First 50 Document Collections of RCV1

Methods	*top20*	*b/p*	*MAP*	F_1
MPBTM	**0.552**	**0.467**	**0.478**	**0.460**
PBTM_FCP	0.494	0.420	0.424	0.424
PBTM_FP	0.470	0.402	0.428	0.424
LDA_word	0.458	0.417	0.421	0.426
PLSA_word	0.434	0.393	0.386	0.395
TNG	0.446	0.367	0.374	0.388
improvement%	11.7	11.2	11.7	8.5
SCP	0.406	0.353	0.364	0.390
n-Gram	0.401	0.342	0.361	0.386
FCP	0.428	0.346	0.361	0.385
improvement%	29.0	32.3	31.3	17.9
BM25	0.434	0.339	0.401	0.410
SVM	0.447	0.409	0.408	0.421
improvement%	23.5	14.2	17.2	9.3

6.3 Results

The proposed model MPBTM with 10 topics, is compared with all the baseline models mentioned above using the 50 human assessed collections. The results are depicted in Table 6 and evaluated using the measures in Table 6 consists of three parts. The top, middle, and bottom parts in Table 6 provide the results of the topic modelling methods, the pattern mining methods, and term-based methods, respectively. The improvement% line at the bottom of each part provides the percentage of improvement achieved by the MPBTM against the best model among all the other baseline models in that part for each measure. From Table 6, we can see that the MPBTM consistently performs the best among all models.

There are two algorithms in the proposed model, user profiling and document filtering. The complexity of the MPBTM is discussed below. For user profiling, the proposed pattern-based topic modelling methods consist of two parts, topic modeling and pattern mining. For the topic modelling part, the initial user interest models are generated using the LDA model, and the complexity of each iteration of Gibbs sampling for the LDA model is linear with the number of topics (V) and the number of documents (N), i.e. O*N For pattern mining, there is no specific quantitative measure for the complexity of pattern mining reported in relevant literature. But the efficiency of the FP-Tree algorithm for generating frequent patterns has been widely accepted in the field of data mining and text mining. The proposed MPBTM and PBTM have the same computational complexity as SCP or frequent closed pattern mining. On the other hand, the MPBTM and the PBTM generate patterns from very small transactional datasets compared with the datasets used in general data mining tasks, because the transactional datasets used in the MPBTM and the PBTM are generated from the topic representations produced by the LDA model rather than the original document collections.

7. CONCLUSION

The proposed MPBTM model generates pattern enhanced topic representations to model user's interests across multiple topics. In the filtering stage, the MPBTM selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents. The proposed approach incorporates the semantic structure from topic modelling and the specificity as well as the statistical significance from the most representative patterns. The proposed model has been evaluated by using the RCV1 and TREC collections for the task of information filtering. In comparison with the state-of-the-art models, the proposed model demonstrates excellent strength on document modelling and relevance ranking.

REFERENCES

[1] Yang Gao,Yeu Xu and Yeefeng Li,"Pattern based topics for document modeling in information filtering"
[2] Rong Hu, Wanchan Dou, Jianxun Liu, "Clubcf: A clustering based collaborative filtering approach for big data application"
[3] C. Zhai, "Statistical language models for information retrieval," Synthesis Lectures Human Lang. Technol., vol. 1, no. 1, pp. 1–141, 2008.
[4] Y. Gao, Y. Xu, and Y. Li, "Pattern-based topic models for information filtering," in Proc. Int. Conf. Data Min. Workshop SENTIRE,2013
[5] Y. Gao, Y. Xu, Y. Li, and B. Liu, "A two-stage approach for generating topic models," in Advances in Knowledge Discovery and DataMining, PADKDD'13. New York, NY, USA: Springer, 2013,
[6] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative Frequent pattern analysis for effective classification," in Proc. IEEE
[7] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal,"Mining frequent patterns with counting inference," ACM
[8] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," InProc. 8th ACM SIGKDD Int.