International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 21 Issue 1 – APRIL 2016.

TOPICAL PATTERN BASED DOCUMENT MODELLING IN INFORMATION FILTERING

J. Parveenraj[#1] & Dr. P. Indirapriya[*2]

[#1] *PG Student, Department of Computer Science and Engineering, Tagore Engineering College, Chennai*
[*2] *Professor, Department of Computer Science and Engineering, Tagore Engineering College, Chennai*

***Abstract*—This paper presents the way how to implement multi document summarization in information filtering. Many mature term-based or pattern-based approaches have been used in the field of information filtering to generate users' information needs from a collection of documents. A fundamental assumption for these approaches is that the documents in the collection are all about one topic. However, in reality users' interests can be diverse and the documents in the collection often involve multiple topics. Topic modelling, such as Latent Dirichlet Allocation (LDA), was proposed to generate statistical models to represent multiple topics in a collection of documents. However, the enormous amount of discovered patterns hinders 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 novel information filtering model, Maximum matched Pattern-based Topic Model (MPBTM), is proposed. The main distinctive features of the proposed model include: (1) user information needs are generated in terms of multiple topics; (2) each topic is represented by patterns; (3) patterns are generated from topic models and are organized in terms of their statistical and taxonomic features; and (4) the most discriminative and representative patterns, called Maximum Matched Patterns, are proposed to estimate the document relevance to the user's information needs in order to filter out irrelevant documents.**

Index Terms—Topic model, information filtering, pattern mining, relevance ranking, user interest model.

I. INTRODUCTION

Information filtering (IF) is a system to remove redundant or unwanted information from an information or document stream based on document representations which represent users' interest. Traditional IF models were developed using a term-based approach. The advantage of the term-based approach is its efficient computational performance, as well as mature theories for term weighting, such as Rocchio, BM25, etc. [1], [2]. But term-based document representation suffers from the problems of polysemy and synonymy. To overcome the limitations of term-based approaches, pattern mining based techniques have been used to utilize patterns to represent users' interest and have achieved some improvements in effectiveness [3], [4], since patterns carry more semantic meaning than terms. 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 [5], [6], [7], [8].

For example, one news article talking about a" car" is possibly related to price, policy, market and so on. At any time, new topics may be introduced in the document stream, which means the user's interest can be diverse and changeable. 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 modelling [9], [10], [11] has become one of the most popular probabilistic text modelling techniques and has been quickly accepted by machine learning and text mining communities. It can automatically classify documents in a collection by a number of topics and represents every document with multiple topics and their corresponding distribution. Two representative approaches are Probabilistic Latent Semantic Analysis (PLSA) [12] and LDA [11]. However, there are two problems in directly applying topic models for information filtering. The first problem is that the topic distribution itself is insufficient to represent documents due to its limited number of dimensions (i.e. a pre-specified number of topics). The second problem is that the word- based topic representation (i.e. each topic in a topic model is represented by a set of words) is limited to distinctively

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.

The original contributions of the proposed MPBTM to the field of IF can be described as follows: 1) We propose

to model users' interest with multiple topics rather than a single topic under the assumption that users' information interests can be diverse. 2) We propose to integrate data mining techniques with statistical topic modelling techniques to generate a pattern-based topic model to represent documents and document collections. The proposed model MPBTM consists of topic distributions describing topic preferences of each document or the document collection and pattern-based topic representations representing the semantic meaning of each topic.
3) We propose a structured pattern-based topic representation in which patterns are organized into groups, called equivalence classes, based on their taxonomic and statistical features. Patterns in each equivalence class have the same frequency and represent similar semantic meaning. With this structured representation, the most representative patterns can be identified which will benefit the filtering of relevant documents. 4) We propose a new ranking method to determine the relevance of new documents based on the proposed model and, especially, the structured pattern-based topic representations. 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. The maximum matched patterns are the most representative and discriminative patterns to determine the relevance of incoming documents.

II. RELATED WORK

IF systems obtain user information needs from user profiles. IF systems are commonly personalized to support the long-term information needs of a particular user or a group of users with similar needs [15]. In an IF process, the primary objective is to perform a mapping from a space of incoming documents to a space of user more precisely, 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 n the TREC data collection [16] was to measure the ability of IF systems to separate relevant from irrelevant documents. The document filtering can be regarded as a classification task or a ranking task. Methods [17], 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 modelled by various approaches that primarily include a term-based model [2], a pattern-based model [18], [19], a probabilistic model [20] and a language model [21]. The popular term-based models include tf*idf, Okapi BM25 and various weighting schemes for the bag of words representation [1], [17], [22]. Term-based models have an unavoidable 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 [23], [24]. But the performance of n-Gram is restricted due to the low frequency of phrases.

Thus, selecting reliable patterns [8] is always very crucial. For example, a number of condensed representations of frequent item sets have been proposed such as closed item sets [6], maximal item sets [5], free item- sets [25], disjunction-free item sets [26] etc. The primary purpose of these condensed 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. Wang et al. [27] proposed the TFP algorithm to extract the top-k most representative closed patterns by pattern length that no less than min l instead of traditional support confidence criteria. In addition, closed patterns stand on the top of the hierarchy induced by each equivalence class, allowing the algorithm to informatively infer the supports of frequent patterns. Topic models techniques have been incorporated in the frame of language model and have achieved successful retrieval results [9], [21], [28], which has opened up a new channel to model the relevance of a document.

The LDA- based document models are state-of-the-art topic modelling approaches. Information retrieval systems based on these models have achieved good performance. The authors claimed the retrieval performance achieved by [9] 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. In these document models, smoothing techniques [29] utilize the word probability across the whole collection to smooth the maximum likelihood (ML) estimate of observing a word in a particular document, which has the same effect as IDF in a term weighting model.

Probabilistic topic modelling [10] can also extract long-term user interests by analysing 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 [30]. 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 [31] 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. Especially when information needs are sensitive to some parameters, both the topic model and the language models are very limited in representing the specificities. In order to overcome the weakness of topic models to interpret specificity, labelling topic techniques [32] are developed for interpreting the semantics of topics by phrases instead of the word-based representations. N-gram statistics can be incorporated with latent topic variables forming a generative probabilistic model to automatically generate topically relevant phrases, such as bigram topic model [33]. The topical n-Gram (TNG) in [34] is seamlessly integrated into the language modelling based IR task, but the improvement this provides is not that significant. In our proposed model, patterns are used to represent corpus and documents, which not only can solve the synonymy problem, but also can deal with the low frequency problem of phrases. In [35], frequent patterns are pre-generated from the original documents and then inserted 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. It can be considered a partial pattern-based topic model since both individuals. words and patterns are used to represent topics. It was applied to classification rather than information filtering. Our proposed model MPBTM is different from the model in [35] in the sense that the topics in the MPBTM model are represented by patterns only. Most importantly, the patterns in the model are well structured so that only the maximum matched patterns are identified and used to estimate document relevance.

III. LATENT DIRICHLET ALLOCATION

Topic modelling algorithms are used to discover a set of hidden topics from collections of documents, where a topic is represented as a distribution over words. Topic models provide an interpretable low-dimensional representation of documents (i.e. with a limited and manageable number of topics). LDA [11] is a typical statistical topic modelling technique and the most common topic modelling tool currently in use. It can discover the hidden topics in collections of documents using the words that appear in the documents. Let D= {$d_1, d_2 \ldots, dM$} be a collection of documents. The total number of documents in the collection is M. The resulting representations of the LDA model are at two levels, document level and collection level. At document level, each document d_i is represented by topic distribution $0_{di}=\{vdi_{,1}, vd_{i,2},..vd_{i,v)}$ v is the number of topics. At collection level, D is represented by a set of topics each of which is represented by a probability distribution over words, $\mathbf{f_j}$ for topic **j**. Overall, we have g for all 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. Take a simple example and let D = $\{d_1, d_2, d_3, d_4\}$ be a small collection of four documents with 12 words appearing in the documents and assume the documents in D involve 3 topics, Z_1, Z_2 and Z_3. Table 1 illustrates the topic distribution over the documents and the word-topic assignments in this small collection. From the outcomes of the LDA model, the topic distribution over the whole collection D can be calculated, $0_D = v_{D,1}, v_{D2,} \ldots v_{D,v}$), where $v_{D,j}$ indicates the importance degree of the topic Z_j in the collection D. Since 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 significant 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 to significantly improve the LDA model. In this paper, we propose a new approach for generating a pattern-based topic model to represent documents and also a new ranking Method to determine relevant documents based on the topic model

IV. 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 reveal the association between words. In order to discover semantically 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 $R_{di,Zj}$ represent the word-topic assignment to topic Z_j in document d_i. $R_{di,Zj}$ is a sequence of words assigned to topic Z_j. For the example illustrated in Table 1, for topic Z_1 in document d_1, $R_{d1,Z1} = \{w_1, w_2, w_3, w_2, w_1\}$.

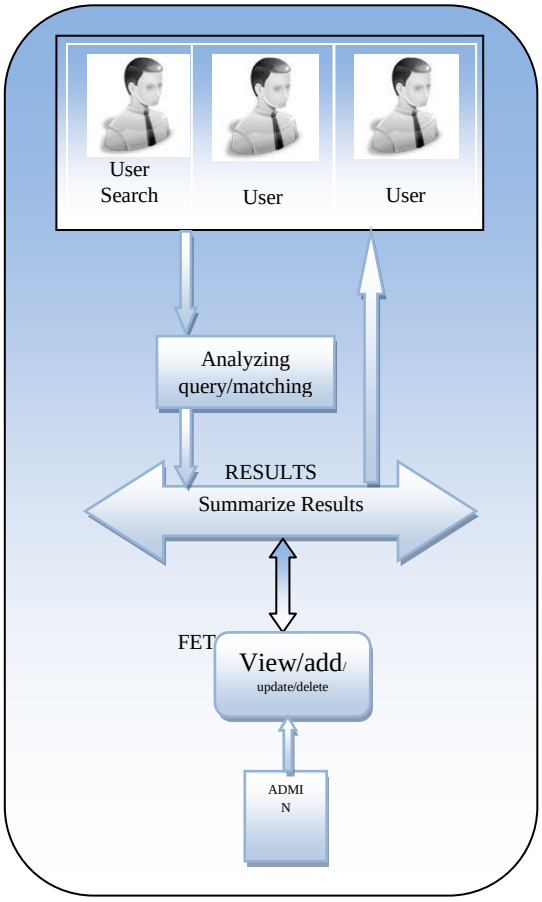
We construct a set of words from each word-topic assignment $R_{di,Zj}$ instead of using the sequence of words in R_{di} ;Z_j , because for pattern mining, the frequency of a word within a transaction is insignificant. Let I_{ij} be a set of words which occur in $R_{di,Zj}$, $I_{ij}=\{w|w \in R_{di,Zj}\}$ i.e. I_{ij} contains the words which are in document d_i and assigned to topic Z_j by LDA. I_{ij} , called a topical document transaction, is a set of words without any duplicates. From all the word-topic assignments $R_{di,Zj}$ to Z_j , i=1, …,M , we can construct a transactional dataset G_j . Let D= $\{d_1, \ldots, dm\}$ be the original document collection, the transactional dataset G_j for topic Z_j is defined as $G_j = \{I_{1j}, I_{2j}, \ldots, I_{Mj}\}$. For the topics in D, we can construct V transactional datasets ($G_1, G_2, \ldots, G_V$). An example of transactional datasets is illustrated in Table 2, which is generated from the example in Table.

4.2 Generate Pattern Enhanced Representation

The basic idea of the proposed pattern-based method is to use frequent patterns generated from each transactional dataset G_j to represent Z_j . In the two-stage topic model [13], frequent patterns are generated in this step. For a given minimal support threshold s, an item set X in G_j is frequent if sup(X)>=α, where sup(X) is the support of X which is the number of transactions in G_j that contain X. The fre-quency of the item set X is defined as $\frac{sup(X)}{}$. Topic Z_j can be represented by a set of all frequent patterns, denoted as $X_{zi}=\{X_{i,1},x_{i,2},..x_{i,mi}\}$ where m_i is the total number of patterns in X_{Zi} and V is the total number of topics. Take G_2 in Table 2 as an example, which is the transactional dataset for Z_2 . For a minimal support threshold s ¼ 2, all frequent patterns generated from G_2 are given in 'item set' and 'pattern' are interchangeable.

V. NOVEL INFORMATION FILETERING

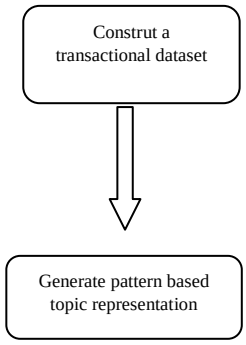
The system architecture consists of admin, user in which both are having login process. Here admin add and update the materials to the database for user usage purpose. In this architecture mainly focus on searching on user, if user enter searching query these query is checking whether it having any mistake or not. It having any mistake, that mistake is corrected by using incremental query construction and neighborhood method. Finally, it shows only relevant information.



transaction	topic document transaction
1	$\{w_1, w_8, w_9\}$
2	$\{w_1, w_7, w_8\}$
3	$\{w_2, w_3, w_7\}$
4	$\{w_1, w_8, w_9\}$

Γ_2

Patterns	supp
$\{w_1\}, \{w_8\}, \{w_1, w_8\}$	3
$\{w_9\}, \{w_7\}\{w_8, w_9\}, \{w_1, w_9\}, \{w_1, w_8, w_9\}$	2



The basic idea of the proposed 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.

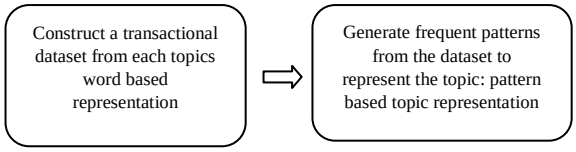


Fig 5.1 Architecture Diagram

5.1 PROCESS OF PROPOSED MODEL

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, Rd1,Z1 = {w1,w2,w3,w2,w1}. Then construct a set of words from each word-topic assignment Rdi,Zj instead of using the sequence of words in Rdi,Zj.

Topic based user Interest Model

1. For a document collection *D* and *V* pre-specified latent topics, from the results of LDA to *D*, generate *V* transactional datasets $\Gamma_{1...}\ \Gamma_v$.

2. Generate user interest model, $U=\{X_{Z1},X_{Z2},\ldots X_{ZV}\}$ $X_{Zi}=\{X_{i1},X_{i2}.,,X_{imi}\}$ is a set of frequent patterns generated from Γi. The patterns in xzi represent what the user is interested in terms of topic *Zi*.

5.2 ALGORITHMS

The proposed IF model can be formally described in two algorithms: User Profiling (i.e. generating user interest models) Algorithm and Document Filtering (i.e. relevance ranking of incoming documents) Algorithm. The former generates pattern-based topic representations to represent the user's information needs. The latter ranks the incoming documents based on the relevance of the documents to the user's needs.

Algorithm 1. User Profiling

Input: a collection of positive training documents D;
minimum support sj as threshold for topic Zj;
number of topics V
Output: U_E $\{E(Z_1), \ldots ;E(Z_V)\}$
1: Generate topic representation f and word-topic assignment
$Z_{d,i}$ by applying LDA to D
2: $U_E := Ø$
3: for each topic Zj ℂ[Z1, ZV] do
4: Construct transactional dataset Gj based on Ø and$z_{d,i}$

EXAMPLE RESULTS OF LDA: WORD-TOPIC ASSIGNMENTS

Topic		Z_1		Z_2		Z_3
Document	$\vartheta_{d,1}$	words	$\vartheta_{d,2}$	words	$\vartheta_{d,3}$	words
d_1	0.6	w_1, w_2, w_3, w_2, w_1	0.2	w_1, w_9, 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_1, w_8, w_8	0.3	w_1, w_{11}, w_{12}
d_3	0.3	w_2, w_1, w_7, w_5	0.3	w_7, w_3, w_3, w_2	0.4	w_4, w_7, w_{10}, w_{11}
d_4	0.3	w_2, w_7, w_6	0.4	w_9, w_8, w_1	0.3	w_1, w_{11}, w_{10}

Table 5.1 TOPIC ASSIGNMENTS

5: Construct user interest model X_{Zj} for topic Z_j using a pattern mining technique so that for each pattern X in XZj , sup(X) > Ø $_j$
6: Construct equivalence class E(Zj) from X_{Zj}
7: $U_E := U_E \cup \{ E(Zj)\}$
8: end for

Algorithm 2. Document Filtering

Input: user interest model $U_E = \{E(Z_1) \ldots E(Z_V)\}$, a list of incoming document D_{in}
Output: $rank_E(d)$, d Є Din
1: rank(d) := 0
2: for each d Є D_{in} do
3: for each topic Z_j Є $[Z_1, Z_V]$ do
4: for each equivalence class EC_{jk} Є $E(Z_j)$ do
5: Scan EC_{jk} and find maximum matched pattern MC^d_{jk}
Which exists in d
6: update $rank_E(d)$ using Equation (3):
7: rank (d) :=rank(d) + $[MC^d_{jk}]^{0.5}$ x f_{jk} x $V_{D,j}$
8: end for
9: end for
10: end for

VI. RESULTS

For different collections can be different. Therefore, selecting an appropriate number of topics is important. As Table 5 shows, the result of the MPBTM with 5 or 10 topics achieves relatively the best performance for this particular dataset. When the topic number rises or reduces, the performance drops. Especially when the topic number rises to 15, the performance drops dramatically, although still outperforms most of the baseline models in Table 6. The proposed model MPBTM with 10 topics, is com-pared 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 Section 6.2. 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.

6..1 Comparisons with Topic-based Models

From the top part of Table 6, we can see that, the MPBTM outperforms all other topic-based models for all the four measures. The PBTM_FCP is the second best model for measures top 20 and b=p, and is in a tie with the PBTM_FP as the second best model for measure F_1 . The PBTM_FP is the second best model for measure MAP.

This result demonstrates that using closed patterns (PBTM_FCP) and, especially, using the proposed maximum matched patterns (MPBTM) to represent topics achieved better results than using frequent patterns (PBTM_FP) for most measures and better than using phrases (TNG) or words (PLSA_word and LDA_word) for all measures. The improvement% line in the top part of Table 5 shows that, the MPBTM which uses the maximum matched patterns consistently achieves the best performance with the improvement percentage against the second best model from a minimum of 8.5 percent to a maximum of 11.7 percent. The comparison results clearly support the second hypothesis.

Methods	top20	b/p	MAP
PBTM FCP	0.00218	0.02990	0.00048
0.00020			
PBTM FP	0.00093	0.00204	0.00223
0.00360			
LDA word	0.00051	0.02210	0.00117
0.00951			
PLSA word	5:05 10 5	0.00594	0.00022
0.00016			
TNG	0.00052	0.00054	0.00026
0.00017			
SCP	1:22 10 5	6:26 10 5	4:44
10 5 0.00019			
n-Gram	0.00034	0.00011	0.00013
0.00026			
FCP	0.00031	3:94 10 5	2:54 10
5 0.00013			
BM25	0.00227	0.03414	0.00249
0.00539			
SVM	0.00051	0.04504	0.00307
0.01714			

TABLE 6 T-Test p-values for All Models Compared with the MPBTM

than FCP simply because it takes multiple topics into consideration when generating user interests. The same reason applies for the better performance of LDA_word over BM25 and SVM; all of these use words to represent user interest, but LDA_word is a topic modelling method while BM25 and SVM are not. These comparisons can strongly validate the first hypothesis, i.e. taking multiple topics into consideration can generate more accurate user information needs. However, the performance of the PLSA_word model is not better than BM25 or SVM. The poor performance of the PLSA_word model indicates its weakness on topic classification, especially lack of discriminative topic representation.

6.2 Comparisons with Pattern-based Models

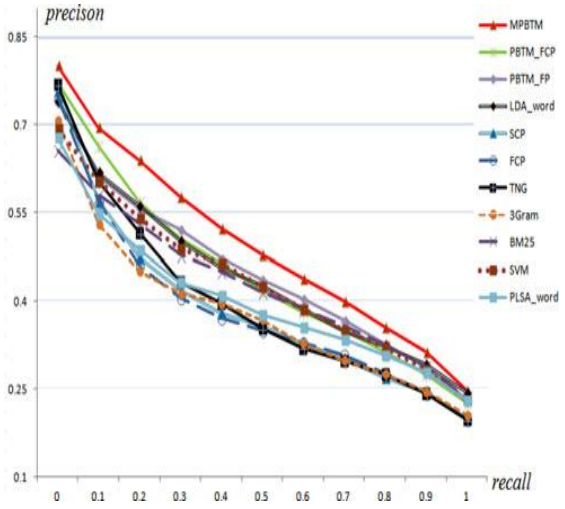
The comparison results among the proposed model and pattern-based baseline models are in the middle part of Table 6. We can see that all the three pattern-based topic modelling models, i.e. MPBTM, PBTM_FCP and PBTM_FP, outperform the three pattern-based baseline models, i.e. SCP, n-Gram, and FCP, which clearly shows the strength obtained by combining topic modelling with pattern-based models. Among the three baseline models, the SCP out per- forms the other two models for b=p, MAP and F_1 , while the FCP model performs the best for top20. The bottom line of the pattern-based part in the table provides the percentage of improvement achieved by the MPBTM against the SCP for b/p, MAP and F_1 , and against the FCP model for top20. The MPBTM achieves excellent performance in improvement percentage with a maximum of 32.3 percent and a minimum of 17.9 percent.

6.3 Comparisons with Term-based Models

From the bottom section of Table 6, we can see that the SVM achieved better performance than the BM25, while the MPBTM and the PBTM_FCP and the PBTM_FP consistently outperform the SVM. The maximum and minimum improvement achieved by the MPBTM against the SVM is23.5 and 9.3 percent, respectively. We also conducted the T-test o compare the MPBTM with all other PBTM models and baseline models. The results are listed in Table 7. The statistical results indicate that the proposed MPBTM significantly outperforms all the other models (all values in Table 7 are less than 0.05) and the improvements are consistent on all four measures. Therefore, we conclude that the MPBTM is an exciting achievement in discovering high-quality features in text documents mainly because it represents the text documents not only using the topic distributions at a general level but also using hierarchical pattern

representations at a detailed specific level, both of which contribute to the accurate document relevance ranking.

The 11-points results of all methods are shown in Fig. 2. The results indicate that the MPBTM has achieved the best performance compared with all the other baseline models.



VII. CONCLUSION

This paper presents an innovative pattern enhanced topic model for information filtering including user interest modelling and document relevance ranking. 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 col- lections 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.The proposed model automatically generates discriminative and semantic rich representations for modelling topics a documents by combining statistical topic modelling techniques and data mining techniques. The technique not only can be used for information filtering, but also can be applied to many content-based feature extraction and modelling tasks, such as information retrieval and recommendations.

REFERENCES

[1] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in Proc. 13th ACM Int. Conf. Inform. Knowl. Manag., 2004, pp. 42–49.

[2] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2002, pp. 436–442.

[3] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, "Mining frequent patterns with counting inference," ACM SIGKDD Explorations Newslett., vol. 2, no. 2, pp. 66–75, 2000.

[4] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective classification," in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 716–725.

[5] R. J. Bayardo Jr, "Efficiently mining long patterns from databases," in Proc. ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 85–93.

[6] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," Data Min. Knowl. Discov., vol. 15, no. 1, pp. 55–86, 2007.

[7] M. J. Zaki and C.-J. Hsiao, "CHARM: An efficient algorithm for closed itemset mining." in Proc. SDM, vol. 2, 2002, pp. 457–473.

[8] Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules," Data Knowl. Eng., vol. 70, no. 6, pp. 555–572

[9] 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. 178–185.

[10] 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. 448–456.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[12] T. Hofmann, "Probabilistic latent semantic indexing," in Proc. 22nd Annu. Int. ACM SIGIR Conf. on Res. Develop. Inform. Retrieval, 1999, pp. 50–57.

[13] Y. Gao, Y. Xu, Y. Li, and B. Liu, "A two-stage approach for generating topic models," in Advances in Knowledge Discovery and Data Mining, PADKDD'13. New York, NY, USA: Springer, 2013, pp. 221–232.

[14] Y. Gao, Y. Xu, and Y. Li, "Pattern-based topic models for information filtering," in Proc. Int. Conf. Data Min. Workshop SENTIRE, 2013, pp. 921–928.

[15] J. Mostafa, S. Mukhopadhyay, M. Palakal, and W. Lam, "A multilevel approach to intelligent information filtering: Model, system, and evaluation," ACM Trans. Inform. Syst., vol. 15, no. 4, pp. 368– 399, 1997.

[16] S. E. Robertson and I. Soboroff, "The TREC 2002 filtering track report," in Proc. TREC, 2002, vol. 2002, no. 3, p. 5.

[17] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, "Adapting ranking svm to document retrieval," in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 186–193.

[18] S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in Proc. 6th Int. Conf. Data Min., 2006, pp. 1157–1161.

[19] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 30–44, Jan.

2012.

[20] J. Lafferty and C. Zhai, "Probabilistic relevance models based on document and query generation," in Language Modeling for Information Retrieval. New York, NY, USA: Springer, 2003, pp. 1–10.

[21] L. Azzopardi, M. Girolami, and C. Van Rijsbergen, "Topic based
language models for ad hoc information retrieval," in Proc. Neural Netw. IEEE Int. Joint Conf., 2004, vol. 4, pp. 3281–3286.

[22] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data," in Proc. Int. Joint Conf. Artif. Intell., 2003, vol. 3, pp. 587–592.

[23] J. Fürnkranz, "A study using n-gram features for text catego-rization," Austrian Res. Inst. Artif. Intell., vol. 3, no. 1998, pp. 1–10,1998.

[24] W. B. Cavnar and J. M. Trenkle, "N-gram-based text catego- rization," Ann Arbor MI, vol. 48113, no. 2, pp. 161–175, 1994.