Pattern-based Topics for Document Modelling in Information Filtering

Yang Gao,Yue Xu and Yuefeng Li

Abstract—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, and this has been widely utilized in the fields of machine learning and information retrieval, etc. But its effectiveness in information filtering has not been so well explored. Patterns are always thought to be more discriminative than single terms for describing documents. 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 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. Extensive experiments are conducted to evaluate the effectiveness of the proposed model by using the TREC data collection Reuters Corpus Volume 1. The results show that the proposed model significantly outperforms both state-of-the-art term-based models and pattern-based models.

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

1 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]–[8].

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. 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]–[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 Latent Dirichlet Allocation (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 represent documents which have different semantic content since many words in the topic representation are frequent general words.

In order to alleviate the ambiguity of the topic representations in LDA, in [13], we proposed a promising way to meaningfully represent topics by patterns rather than single words through combining topic models with pattern mining techniques. Specifically, the patterns are

• *Authors are with the Faculty of Science and Engineering, Queensland University of Technology, Brisbane, QLD, 4000.*
E-mail: y21.gao@qut.edu.au, yue.xu@qut.edu.au, y2.li@qut.edu.au

Manuscript received ; revised .

generated from the words in the word-based topic representations of a traditional topic model such as the LDA model. This ensures that the patterns can well represent the topics because these 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 pattern-based topic model, which has been utilized in IF [14], can be considered as a "post-LDA" model in the sense that the patterns are generated from the topic representations of the LDA model. Because 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. However, 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 Maximum matched Pattern-Based Topic Model (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.

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 to the background works of LDA. Sections 4 and 5 present the details of our proposed model. Then, extensive 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 [14], 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 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 relevant documents. More precisely, denoting the space of incoming documents as D, the mapping $rank : D \rightarrow R$ such that $rank(d)$ corresponds to the relevance of a document d. The filtering track in 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. 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 number of returned patterns is huge because if a pattern is frequent, then each of its subpatterns is frequent too. Thus, selecting reliable patterns [8] is always very crucial. For example, a number of condensed representations of frequent itemsets have been

proposed such as closed itemsets [6], maximal itemsets [5], free itemsets [25], disjunction-free itemsets [26] etc. The primary purpose of these condensed representations is to enhance the efficiency of using the generated frequent itemsets without losing any information. Among these proposed itemsets, 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. So, in this paper, we intend to utilize the hierarchical structure of patterns based on equivalence class partitions to represent user profiles creatively.

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 individual 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.

3 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). Latent Dirichlet Allocation (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 $D = \{d_1, d_2, \cdots, d_M\}$ 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 a fixed vocabulary of 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 d_i is represented by topic distribution $\theta_{d_i} = (\vartheta_{d_i,1}, \vartheta_{d_i,2}, \cdots, \vartheta_{d_i,V})$, 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, ϕ_j for topic j. Overall, we have $\Phi = \{\phi_1, \phi_2, \cdots, \phi_V\}$ 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

TABLE 1
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}

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, $\theta_D = (\vartheta_{D,1}, \vartheta_{D,2}, \cdots, \vartheta_{D,V})$, where $\vartheta_{D,j}$ indicates the importance degree of the topic Z_j in the collection D.

The topical n-Gram model proposed in [34] automatically and simultaneously discovers topics and extracts topically relevant phrases. It has been seamlessly integrated into the language modelling based IR task [34]. Readers can refer to [34] for more details. Compared with word representation, phrases are more discriminative and carry more concrete semantics. 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.

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. Furthermore, for a new incoming document, various methods can be utilized to situate its content in terms of the trained topics. However, single word based topic representations contain ambiguous semantics. Thus, TNG improves the LDA model by expanding word-based topic representation to phrase-based, which enhances the explicit semantics of topics. However, TNG suffers from the low occurrence problem and fails 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.

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 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.

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

4.1 Construct Transactional Dataset

Let R_{d_i,Z_j} represent the word-topic assignment to topic Z_j in document d_i. R_{d_i,Z_j} 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, $R_{d_1,Z_1} = \langle w_1, w_2, w_3, w_2, w_1 \rangle$. We construct a set of words from each word-topic assignment R_{d_i,Z_j} instead of using the sequence of words in R_{d_i,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_{d_i,Z_j}, $I_{ij} = \{w | w \in R_{d_i,Z_j}\}$, 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_{d_i,Z_j} to Z_j, $i = 1, ..., M$, we can construct a transactional dataset Γ_j. Let $D = \{d_1, \cdots, d_M\}$ be the original document collection, the transactional dataset Γ_j for topic Z_j is defined as $\Gamma_j = \{I_{1j}, I_{2j}, \cdots, I_{Mj}\}$. For the topics in D, we can construct V transactional datasets $(\Gamma_1, \Gamma_2, \cdots, \Gamma_V)$. An example of transactional datasets is illustrated in TABLE 2, which is generated from the example in TABLE 1.

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 Γ_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 σ, an itemset X in Γ_j is frequent if $supp(X) >= \sigma$, where $supp(X)$ is the support of X which is the number of transactions in Γ_j that contain X. The frequency of the itemset X is defined as $\frac{supp(X)}{|\Gamma_j|}$. Topic Z_j can be represented by a set of all frequent patterns, denoted as $\mathbf{X}_{Z_i} = \{X_{i1}, X_{i2}, \cdots, X_{im_i}\}$, where m_i is the total number of patterns in $\mathbf{X}_{Z_i}$ and V is the total number of topics. Take Γ_2 in TABLE 2 as an example, which is the transactional dataset for Z_2. For a minimal support threshold $\sigma = 2$, all frequent patterns generated from Γ_2 are given in TABLE 3 ('itemset' and 'pattern' are interchangeable in this paper).

TABLE 3
The frequent patterns for Z_2, $\sigma = 2$

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

TABLE 4
The equivalence classes in Z_2

EC_{21} ($f_{21} = 0.75$)	EC_{22} ($f_{22} = 0.5$)	EC_{23} ($f_{23} = 0.5$)
$\{w_1, w_8\}$	$\{w_1, w_8, w_9\}$	$\{w_7\}$
$\{w_1\}$	$\{w_1, w_9\}$	
$\{w_8\}$	$\{w_8, w_9\}$	
	$\{w_9\}$	

5 Information Filtering Model based on Pattern Enhanced LDA

The representations generated by the pattern enhanced LDA model, discussed in Section 4, carry more concrete and identifiable meaning than the word-based representations generated using the original LDA model. However, 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. 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, Maximum matched Pattern-based Topic 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. Moreover, the proposed model estimates the relevance of incoming documents based on Maximum Matched Patterns, which are the most distinctive and representative patterns, as proposed in this paper. The details are described in the following subsections.

5.1 Pattern Equivalence Class

Normally, the number of frequent patterns is considerably large and many of them are not necessarily useful. Several concise patterns have been proposed to represent useful patterns generated from a large dataset instead of frequent patterns such as maximal patterns [5] and closed patterns [6]. The number of these concise 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 [7], [8].

Definition 1. *Closed Itemset*: for a transactional dataset, an itemset X is a closed itemset if there exists no itemset X' such that (1) $X \subset X'$, (2) $supp(X) = supp(X')$.

Definition 2. *Generator*: for a transactional dataset Γ, let X be a closed itemset and $T(X)$ consists of all transactions in Γ that contain X, then an itemset g is said to be a generator of X iff $g \subset X, T(g) = T(X)$ and $supp(X) = supp(g)$.

Definition 3. *Equivalence Class*: for a transactional dataset Γ, let X be a closed itemset and $G(X)$ consist of all generators of X, then the equivalence class of X in Γ, denoted as $EC(X)$, is defined as $EC(X) = G(X) \cup \{X\}$.

Let EC_1 and EC_2 be two different equivalence classes of the same transactional dataset. Then $EC_1 \cap EC_2 = \emptyset$, which means that the equivalence classes are exclusive of each other.

All the patterns in an equivalence class have the same frequency. The frequency of a pattern indicates the statistical significance of the pattern. The frequency of the patterns in an equivalence class is used to represent the statistical significance of the equivalence class. TABLE 4 shows the three equivalence classes within the patterns for topic Z_2 in TABLE 3, where f indicates the statistical significance of each class.

There are two parts in the proposed model MPBTM: the training part to generate user interest information from a collection of training documents (i.e. user interest modelling introduced in Section 5.2) and the filtering part to determine the relevance of incoming documents based on the user's interests (i.e. document relevance ranking introduced in Section 5.3).

5.2 Topic-based User Interest Modelling

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, θ_D represents the topic distribution of D and can be used to represent the user's topic interest distribution, $\theta_D = (\vartheta_{D,1}, \vartheta_{D,2}, \cdots, \vartheta_{D,V})$,

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, i.e. $\vartheta_{D,j} = \frac{1}{M}\sum_{i=1}^{M}\theta_{d_i,j}$. The probability distribution of topics in θ_D represents the degree of interest that the user has in these topics.

By using the methods 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, $\Gamma_1, \cdots, \Gamma_V$ can be generated from which the pattern-based topic representations for the collection, $U = \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$, can be generated, where each $\mathbf{X}_{Z_i} = \{X_{i1}, X_{i2}, \cdots, X_{im_i}\}$ is a set of frequent patterns generated from transactional dataset Γ_i. U is considered the user interest model, the patterns in each $\mathbf{X}_{Z_i}$ represent what the user is interested in terms of topic Z_i.

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 effectiveness of document filtering 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. As discussed in Section 5.1, equivalence class is a useful structure which collects the frequent patterns with the same frequency into one group. The statistical significance of the patterns in one equivalence class is the same. This distinctive feature of equivalence classes can make the patterns more effectively used in document filtering. In this paper, we propose to use equivalence classes to represent topics instead of using frequent patterns or closed patterns.

Assume that there are n_i frequent closed patterns in $\mathbf{X}_{Z_i}$, which are $c_{i1}, \cdots, c_{in_i}$, and that $\mathbf{X}_{Z_i}$ can be partitioned into n_i equivalence classes, $EC(c_{i1}), \cdots, EC(c_{in_i})$. For simplicity, the equivalence classes are denoted as $EC_{i1}, \cdots, EC_{in_i}$ for $\mathbf{X}_{Z_i}$, or simply for topic Z_i. Let $\mathbb{E}(Z_i)$ denote the set of equivalence classes for topic Z_i, i.e. $\mathbb{E}(Z_i) = \{EC_{i1}, \cdots, EC_{in_i}\}$. In the model MPBTM, the equivalence classes $\mathbb{E}(Z_i)$ are used to represent user interests which are denoted as $\mathbb{U}_E = \{\mathbb{E}(Z_1), \cdots, \mathbb{E}(Z_V)\}$.

5.3 Topic-based Document Relevance Ranking

In terms of the statistical significance, all the patterns in one equivalence class are the same. The differences among them are their size. If a longer pattern and a shorter pattern from the same equivalence class appear in a document simultaneously, the shorter one becomes insignificant since it is covered by the longer one and it has the same statistical significance as the longer one.

In the filtering stage, document relevance is estimated to filter out irrelevant documents based on the user's information needs. In this paper, for a new incoming document d, the basic way to determine the relevance of d to the user interests is firstly to identify maximum patterns in d which match some patterns in the topic-based user interest model and then estimate the relevance of d based on the user's topic interest distributions and the significance of the matched patterns.

The significance of one pattern is determined not only by its statistical significance, but also by its size since the size of the pattern indicates the specificity level. Among a set of patterns, usually a pattern taxonomy exists. For example, Fig. 1 depicts the taxonomy constructed for $\mathbf{X}_{Z_2}$ in TABLE 3. This tree-like structure demonstrates the subsumption relationship between the discovered patterns in Z_2. The longest pattern in a pattern taxonomy, such as $\{w_1, w_8, w_9\}$ in Fig. 1, is the most specific pattern that describes a user's interests since longer patterns have more specific meanings, while single words, such as w_1 in Fig. 1, are the most general patterns which are less capable of discriminating the meaning of the topic from other topics as compared to longer patterns such as $\{w_1, w_8, w_9\}$. The pattern taxonomy presents different specificities of patterns according to the level in the taxonomy structure and thus the size of the pattern.

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. Therefore, we define the pattern specificity below.

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 $0 < m < 1$.

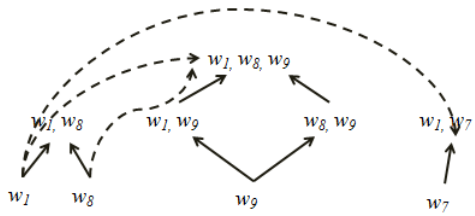


Fig. 1. Pattern Taxonomy in Z_2

Definition 5. *Topic Significance*: Let d be a document, Z_j be a topic in the user interest model, PA^d_{jk} be a set of matched patterns for topic Z_j in document d,

$k = 1, \cdots, n_j$, and $f_{j1}, \cdots, f_{jn_j}$ be the corresponding supports of the matched patterns, then the topic significance of Z_j to d is defined as:

$$sig\left(Z_j, d\right) = \sum_{k=1}^{n_j} spe\left(PA_{jk}^d\right) \times f_{jk} = \sum_{k=1}^{n_j} a|PA_{jk}^d|^m \times f_{jk} \tag{1}$$

where m is the scale of pattern specificity (we set $m = 0.5$), and a is a constant real number (in this paper, we set $a = 1$).

In the MPBTM model, the topic significance is determined by maximum matched pattern, which is defined below.

Definition 6. *Maximum Matched Pattern*: Let d be a document, Z_j be a topic in the user interest model, $EC_{j1}, \cdots, EC_{jn_j}$ be the pattern equivalence classes of Z_j, then a pattern in d is considered a maximum matched pattern to equivalence class EC_{jk}, denoted as MC_{jk}^d, if the following conditions are satisfied:

1) $MC_{jk}^d \subseteq d$ and $MC_{jk}^d \in EC_{jk}$;
2) $\nexists X$ such that $X \in EC_{jk}, X \subseteq d$ and $MC_{jk}^d \subset X$.

The maximum matched pattern MC_{jk}^d to equivalence class EC_{jk} must be the largest pattern in EC_{jk} which is contained in d and all the patterns in EC_{jk} that are contained in d must be covered by MC_{jk}^d. Therefore, the maximum matched patterns MC_{jk}^d, where $k = 1, \cdots, n_j$ are considered the most significant patterns in d which can represent the topic Z_j. Take the equivalence class EC_{22} in Z_2 shown in TABLE 4 as an example, for a document $d' = \{w_1, w_2, w_9, w_{10}, w_{12}\}$, the maximum matched patterns would be $MC_{22}^{d'} = \{w_1, w_9\}$.

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:

$$Rank(d) = \sum_{j=1}^{V} sig(Z_j, d) \times \vartheta_{D,j} \tag{2}$$

For the MPBTM, the patterns PA_{jk}^d in the topic significance $sig\left(Z_j, d\right)$ are maximum matched patterns in $\mathbb{U}_E$. By incorporating Equation (1) into Equation (2), the relevance ranking of d, denoted as $Rank_E(d)$, is estimated by the following equation:

$$Rank_E(d) = \sum_{j=1}^{V} \sum_{k=1}^{n_j} |MC_{jk}^d|^{0.5} \times \delta(MC_{jk}^d, d) \times f_{jk} \times \vartheta_{D,j} \tag{3}$$

where V is the total number of topics, MC_{jk}^d is the maximum matched patterns to equivalence class $EC_{jk}, k = 1, \cdots, n_j$ and $f_{j1}, \cdots, f_{jn_j}$ is the corresponding statistical significance of the equivalence classes, $\vartheta_{D,j}$ is the topic distribution, and

$$\delta(X, d) = \begin{cases} 1 & \text{if } X \in d \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The higher the $Rank_E(d)$, the more likely the document is relevant to the user's interest.

5.4 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 σ_j as threshold for topic Z_j; number of topics V
Output: $\mathbb{U}_E = \{\mathbb{E}(Z_1), \cdots, \mathbb{E}(Z_V)\}$

```
Generate topic representation φ and word-topic assignment z_{d,i} by applying LDA to D
U_E := ∅
for each topic Z_j ∈ [Z_1, Z_V] do
    Construct transactional dataset Γ_j based on φ and z_{d,i}
    Construct user interest model X_{Z_j} for topic Z_j using a pattern mining technique so that for each pattern X in X_{Z_j}, supp(X) > σ_j
    Construct equivalence class E(Z_j) from X_{Z_j}
    U_E := U_E ∪ {E(Z_j)}
end for
```

Algorithm 2 *Document Filtering*

Input: user interest model $\mathbb{U}_E = \{\mathbb{E}(Z_1), \cdots, \mathbb{E}(Z_V)\}$, a list of incoming document D_{in}
Output: $rank_E(d), d \in D_{in}$

```
rank(d) := 0
for each d ∈ D_in do
    for each topic Z_j ∈ [Z_1, Z_V] do
        for each equivalence class EC_jk ∈ E(Z_j) do
            Scan EC_{k,j} and find maximum matched pattern MC_jk^d which exists in d
            update rank_E(d) using Equation 3:
            rank(d) := rank(d) + |MC_jk^d|^0.5 × f_jk × ϑ_{D,j}
        end for
    end for
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. To verify the hypotheses, experiments and evaluation have been conducted. This section discusses the experiments and evaluation in terms of data collection, baseline models, measures and results. The results show that the proposed topic-based model significantly outperforms the state-of-the-art models in terms of effectiveness.

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 [36], 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. In this paper, only the first 50 collections are used for experiments. The 'title' and 'text' of the documents are used by all the models in the experiments.

6.2 Measures

The effectiveness is assessed by five different measures: average precision of the top K ($K = 20$) documents, $F_\beta(\beta = 1)$ measure, Mean Average Precision (MAP), break-even point (b/p) and Interpolated Average Precision (IAP) on 11-points. F_1 is a criterion that assesses the effect involving both precision (p) and recall (r),which is defined as $F_1 = \frac{2pr}{p+r}$. The larger the *top20*, MAP, b/p or F_1 score, the better the system performs. The 11 points measure is the precision at 11 standard recall levels (i.e. recall = 0, 0.1,$\cdots$, 1).

The experiments tested across the 50 collections of independent datasets, which satisfy the generalized cross-validation for the statistical estimation model.

The statistical method, T-Test, was also used to verify the significance of the experimental results. If the p-value associated with t is significantly low (< 0.05), there is evidence to verify that the difference in means across the paired observations is significant.

6.3 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. For each category, some state-of-the-art methods were chosen as the baseline models. For the topic modelling category, three topic modelling methods are chosen as baseline models, two of them are PLSA_word and LDA_word which represent topics with single terms, the other is TNG which uses phrases to represent topics. We have proposed two topic-based models [14], Pattern-based Topic Models PBTM_FP and PBTM_FCP which use frequent patterns and closed patterns to represent topics, respectively. These two models were also chosen as topic-based baseline models. For the pattern mining category, the baseline models include frequent closed patterns (FCP), frequent sequential closed patterns (SCP) and phrases (n-Gram). The third category includes the classical term-based methods BM25 and SVM. An important difference between the topic modelling methods and other methods is worth mentioning, that is, the topic modelling methods consider multiple topics in each document collection and use patterns (e.g. PBTM and MPBTM) or words (e.g. LDA_word) to represent the topics, whereas the pattern mining and term-based methods assume that the documents within one collection are about one topic and use patterns or terms/words to represent documents directly. More details about these baseline models are given below.

(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 [34] introduced in Section 3 are used to represent user interest needs and phrase frequency is used to represent topic relevance.

- PBTM

In [14], two PBTM models have been proposed and they use frequent patterns (FP) and frequent closed patterns (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. The following equation is used to calculate the relevance of a document d:

$$Rank(d) = \sum_{j=1}^{V} \sum_{k=1}^{m_j} |X_{jk}^d|^{0.5} \times \delta(X_{jk}^d, d) \times f_{jk} \times \vartheta_{D,j} \quad (5)$$

where $|X_{jk}^d|$ is a frequent pattern in PBTM_FP and a closed pattern in the PBTM_FCP.

We implement PLSA model with Lemur toolkit [1] with 1000 iterations as default setting. And the LDA model is implemented by MALLET toolkit [2]. The parameters for all LDA-based topic models are set as follows: the number of iterations of Gibbs sampling is 1000, the hyper-parameters of the LDA model are $\alpha = 50/V$ and

1. http://www.lemurproject.org/
2. http://mallet.cs.umass.edu/topics.php

$\beta = 0.01$, which were used and justified in [37]. Our experience shows that the performance of the proposed MPBTM model is not very sensitive to the settings of these parameters. But the number of topics V affects the results depending on the size of various data collections. The results are shown in TABLE 5.

In the process of generating pattern enhanced topic representations, the minimum support σ_{rel} for every topic in each collection is different, because the number of positive documents in different collections of the RCV1 is very different. In order to ensure enough transactions from positive documents to generate accurate patterns for representing user needs, the minimum support σ_{rel} is set as follows :

$$\sigma_{rel} = \begin{cases} 1 & n \leq 2 \\ max(2/n, 0.3) & 2 < n \leq 10 \\ max(3/n, 0.3) & 10 < n \leq 13 \\ max(4/n, 0.3) & 13 < n \leq 20 \\ 0.3 & otherwise. \end{cases} \quad (6)$$

where n is the number of transactions from relevant documents in each transactional database.

(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 [19]. In this model, every document in the training dataset (D) is split in paragraphs which are the transactions for pattern mining. Readers who are interested in the details can refer to [38] and [19].

- 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:

$$W(t) = \frac{tf \times (k+1)}{k \times ((1-b) + b\frac{DL}{AVDL}) + tf} \times \log(\frac{N - n + 0.5}{n + 0.5}) \quad (7)$$

where N is total number of documents in the collection; n is the number of documents that contain term t; tf is the term frequency; DL and $AVDL$ are the document length and average document length, receptively; and k and b are the parameters, which are set as 1.2 and 0.75 as used and explained in [39].

- Support Vector Machine (SVM)

The linear SVM has been proven very effective for text categorization and filtering [40]. We would compare it with other baseline models, however, most existing SVMs are designed for making a binary decision rather than ranking documents. In this paper, we adopted the ranked-based SVM (see *http://svmlight.joachims.org*).

The SVM only uses term-based features extracted from training documents. There are two classes: $y_i \in \{-1, 1\}$ where $+1$ is assigned to a document if it is relevant; otherwise it is assigned with -1 and there are N labelled training examples: $(d_1, y_1), \ldots, (d_i, y_i), \ldots, (d_N, y_N)$, $d_i \in \mathbb{R}^n$ where n is the dimensionality of the vector. Given a function $h(d) =< w \cdot d > +b$ where b is the bias, $h(d) = +1$ if $< w.d > +b \geq 0$; otherwise $h(d) = -1$, and $< w \cdot d >$ is the dot product of an optimal weight vector w and the document vector d. To find the optimal weight vector w for the training set, we perform the following function: $w = \sum_{i=1}^{N} y_i \alpha_i d_i$ subject to $\sum_{i=1}^{l} \alpha_i y_i = 0$ and $\alpha_i \geq 0$, where α_i is the weight of the sample d_i. For the purpose of ranking, b can be ignored.

6.4 Results

For different document collections, the number of topics involved in the 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.

TABLE 5
The MPBTM with different topic number

Number of Topics	*top20*	*b/p*	*MAP*	F_1
3	0.517	0.428	0.449	0.436
5	0.551	0.464	0.481	0.457
10	0.552	0.467	0.478	0.460
15	0.473	0.409	0.430	0.433

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 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

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

$improvement\%$ 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.

6.4.1 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 $top20$ 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% to a maximum of 11.7%. The comparison results clearly support the second hypothesis.

By observing the results in TABLE 5 accross all three parts, we can see that the topic-based models in the top part (except for TNG and PLSA_word) achieved better performance than the models in the second and third parts. Even though the TNG model did not always perform better than the non topic modelling based models, it performs considerably better than the n-Gram models. Both TNG and n-Gram use phrases, the difference is that TNG uses phrases to represent the semantic meaning of topics and also uses topic distributions to represent the user's topic preference, whereas n-Gram directly uses phrases to represent the user's information needs. Similar comparisons can be made between PBTM_FCP and FCP, LDA_word and BM25, and between LDA_word and SVM. Both PBTM_FCP and FCP use frequent closed patterns to represent user interests. However, PBTM_FCP achieved clearly better performance 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.

6.4.2 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 outperforms 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% and a minimum of 17.9%.

6.4.3 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 is 23.5% and 9.3%, respectively.

We also conducted the T-test to 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

TABLE 7
T-Test p-values for all models compared with the MPBTM

Methods	*top20*	*b/p*	*MAP*	F_1
$PBTM\_FCP$	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

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

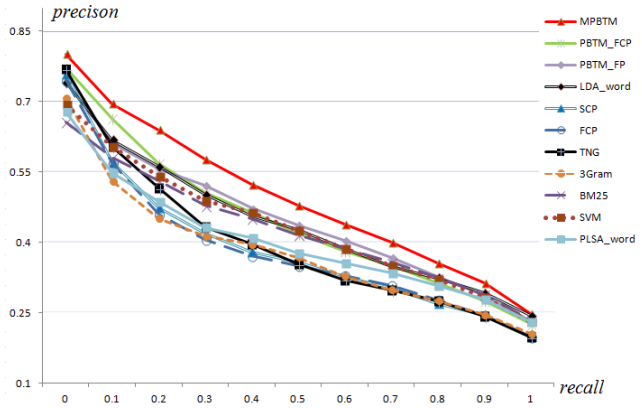


Fig. 2. 11 point results of comparison between the proposed MPBTM and baseline models

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.

7 Discussion

As we can see from the experiment results, taking multiple topics and topic distribution into consideration in generating user interest models and also in document relevance ranking can greatly improve the performance of information filtering. The reason behind the MPBTM and the PBTM achieving the excellent performance is mainly because we inventively incorporated pattern mining techniques into topic modelling to generate pattern-based topic models which can represent user interest needs in terms of multiple topics. Most importantly, the topics are represented by patterns which bring concrete and precise semantics to the user interest models. Moreover, the outstanding performance of the MPBTM over the PBTM_FP and the PBTM_FCP indicates the significant benefit of using the proposed maximum matched patterns in estimating document relevance over using frequent patterns and frequent closed patterns. Clustering is a classical technique for classifying objects (e.g., documents) into multiple clusters and has been used for user interest modelling and document modelling [9]. It could be an alternative multi-topic model for generating pattern-based topic models. However, the performance of the clustering-based technique has been proved worse than the LDA model in document modelling for IR [9] and document clustering [41]. Moreover, since the clustering-based technique assigns documents to clusters without identifying topic related words for each cluster, all the words appearing in the documents would be used to generate patterns. In contrast, the topic representations of the LDA model specify topic related words from which more accurate patterns can be generated. This advantage makes the LDA model superior to the clustering based technique to generate pattern-based topic models. This section will provide more discussions on the performance of the proposed model.

7.1 Topic based Relevance Estimation

TABLE 6 shows that all the topic based models (except for PLSA_word) outperform all the other baseline models including the pattern-based, phrase-based, and term-based models. As we have mentioned above, this is mainly because the topic-based models represent the documents not only using patterns, phrases, or words, but also using topic distributions. Most importantly, the patterns, phrases or words used by the topic-based models are topic related, which is a key difference from the pattern-based or word-based baseline models.

All the topic-based models estimate the relevance of a new document based on topic distribution as well as topic significance represented by patterns (i.e. MPBTM, PBTM_FCP), phrases (i.e. TNG), or words (LDA_word). However, a key difference between the MPBTM and the other topic-based models is that the MPBTM estimates the relevance only utilizing the most representative patterns, the maximum matched patterns, instead of using all the patterns, whereas the other models use all matched patterns, phrases, or words. It is because of only using the maximum matched patterns which are sensitive to user specific interests, the MPBTM significantly outperforms all the other topic-based models.

7.2 Topical Transactions

As mentioned in pattern-based baseline models, the transactional datasets for generating patterns usually use sentences or paragraphs as transactions. Classifying an itemset as frequent means it is contained in many sentences or paragraphs. It makes sense to some extent when the collection of documents focuses only on one topic. In the case that multiple topics are involved in the collection, the frequent patterns generated from the whole collection may not be able to represent any of the topics and thus unlikely to be able to represent the collection correctly.

To emphasize the semantic structure of the user's interests which involve multiple topics, the MPBTM and the PBTM construct transactional databases in terms of different topics. As result, transactions in the same topical transactional database share relatively common interests. The discovered patterns from one topical transactional dataset are more likely to represent one aspect of the user's interests and be more sensitive to effecting accurate representations of this aspect.

For example, collection 101 in RCV1, the traditional method, like SCP, can only find patterns "vw, piech", "piech, carmake", "piech, vw" and other single words. But the MPBTM can discover topical patterns, such as "germany, vw, volkswagen, carmake, car, chairman" and "ag, manage, piech, largest, vw, develop, " in topic 2 and topic 4, respectively. This example shows that the patterns generated by the MPBTM are much longer than that generated by the SCP, not to mention the single words in LDA. Therefore, the MPBTM patterns are more specific and meaningful than the SCP patterns and single words in LDA.

7.3 Maximum Matched Patterns

In the MPBTM, the patterns which represent user interests are not only grouped in terms of topics, but also partitioned based on equivalence classes in each topic group. The patterns in different groups or different equivalence classes have different meanings and distinct properties. Thus, user information needs are clearly represented according to various semantic meanings as well as distinct properties of the specific patterns in different topic groups and equivalence classes.

7.3.1 Pattern Quality

TABLE 8
Comparison of the number of patterns or terms used for filtering by each method on all collections

	BM25 (Terms)	SCP (Patterns)	MPBTM (Equivalence classes)
Avg. Number	623	157	33

Closed patterns have been widely recognized as quality patterns to concisely represent the data in a given dataset. The model PBTM_FCP utilizes closed patterns to represent user interests. In this experiment, the effectiveness of the proposed MPBTM is verified by comparing the MPBTM with the PBTM_FCP since both models utilize closed patterns. The PBTM_FCP directly uses all closed patterns to represent user interests and also to estimate the relevance of a new document, whereas the MPBTM uses equivalence classes based on frequent closed patterns to represent user interests and estimates document relevance using maximum matched patterns only. From TABLE 6, we can see that the PBTM_FCP achieved better performance than all the other models but the MPBTM. This result is an excellent example to show the quality of closed patterns.

TABLE 8 shows the average number of patterns or terms extracted from the whole collections using three models, BM25, SCP, and MPBTM, which represent the three categories of models, i.e. term-based, pattern-based, and topic-based with equivalence classes. For all the collections, the BM25 generates the largest feature space (i.e. the set of terms), while the SCP has a relatively fewer set of patterns. The number of patterns found by the MPBTM is much smaller than either of the other two models, being only about 21% of the number in the SCP and 5.3% of the number in the BM25. However, the performance of the MPBTM is the best of the three models. It should be mentioned that the number of maximum matched pattens when using the MPBTM to determine the relevance of an incoming document is always equal to the number of equivalence classes in the MPBTM model because only one pattern is selected from each equivalence class to estimate the document relevance. The selected patterns won't be repeatedly used for different equivalence classes since the patterns belong to only one equivalence class, i.e. they are partitioned exclusively by equivalence class. Thus, we believe that the patterns selected to estimate the relevance of a document are high quality patterns with excellent characteristics because they are, (1) comprehensive (i.e. cover all topics and also all equivalence classes of each topic), (2) non-redundant (i.e. not be repeatedly used for different equivalence classes), (3) representative (i.e. the maximum matched among the patterns in the same equivalence class), and compact (i.e. small number of selected patterns). With these distinctive characteristics, the maximum matched patterns make the MPBTM model achieve the best performance.

7.3.2 Pattern Specificity

LDA supports a very strong foundation for generating semantics in terms of topic representation and topic distribution. But simply utilizing topic distribution to represent user interests is insufficient. The topical phrases (i.e. the phrases in the n-Gram model) are too strict to exactly match the phrases in documents, while the topical words (i.e. the words in the LDA_word model) are single words and are often too general to represent specific topics. The patterns in the MPBTM, on the contrary, are grouped based on their support and are structured based on their taxonomic relationship. The patterns in the MPBTM deliver specificity that is enhanced by using the association of words (rather than single words as in LDA_word) and the taxonomic levels of the patterns. These specificity enhanced patterns can more accurately represent specific topics and thus more accurately represent users' information needs.

7.4 Complexity

As discussed in Section 5.4, 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 modelling 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. $O(V * N)$ [9].

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. The patterns used to represent topics are generated from the words which are considered to represent the document topics by the LDA model. These words are part of the original documents, whereas other pattern mining models generate patterns from the whole collection of documents.

Moreover, the MPBTM and PBTM models combine the topic modelling and pattern mining linearly. Thus, in summary, the complexity of the MPBTM and PBTM models can be determined by topic modelling or pattern mining. In most cases, the complexity of the MPBTM and PBTM models would be the same as pattern mining since, in general, the complexity of pattern mining is greater than that of topic modelling.

It should be mentioned that the user profiling part can be conducted off-line which means that the complexity of the user profiling part will not affect the efficiency of the proposed IF model.

For information filtering, the set of patterns or terms used to represent the user's information needs is usually called a feature space. For an incoming document, the complexity to determine its relevance to the user needs is linear to the size of the feature space for the pattern-based methods (i.e. SCP, n-Gram, and FCP) and the term-based methods (i.e. BM25 and SVM), $O(S)$ where S is the size of the feature space. For the topic modelling based methods, due to the use of topics, the complexity of determining a document's relevance is $O(V*S)$ where V is the number of topics and S is the number of patterns or terms in each topic representation. For the MPBTM model, even though it has an extra loop (i.e. step 5 in Algorithm 2) to check equivalence classes for each topic, it has the same complexity as the other topic modelling based methods because the patterns are partitioned into equivalence classes (i.e. no patterns will be included in more than one equivalence class and thus will not be scanned more than once) and the number of patterns in the worst case. Theoretically the complexity of the three models is the same. But in practice, the complexity of the MPBTM is actually lower than the other two models because only part of the patterns will be scanned, while the other two models have to scan all patterns or terms.

8 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 accross 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.

The proposed model automatically generates discriminative and semantic rich representations for modelling topics and 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.

Acknowledgements

This paper was partially supported by Grant DP140103157 from the Australian Research Council.

References

[1] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in *Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management*. ACM, 2004, pp. 42–49.

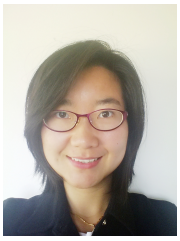
[2] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2002, pp. 436–442.

[3] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, "Mining frequent patterns with counting inference," *ACM SIGKDD Explorations Newsletter*, 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 *IEEE 23rd International Conference on Data Engineering, ICDE'2007*. IEEE, 2007, pp. 716–725.

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

[6] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery*, 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 *SDM*, vol. 2, 2002, pp. 457–473.
[8] Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules," *Data & Knowledge Engineering*, vol. 70, no. 6, pp. 555–575, 2011.
[9] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2006, pp. 178–185.
[10] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2011, pp. 448–456.
[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[12] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 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*. Springer, 2013, pp. 221–232.
[14] Y. Gao, Y. Xu, and Y. Li, "Pattern-based topic models for information filtering," in *Proceedings of International Conference on Data Mining Workshop SENTIRE, ICDM'2013*. IEEE, 2013.
[15] J. Mostafa, S. Mukhopadhyay, M. Palakal, and W. Lam, "A multilevel approach to intelligent information filtering: model, system, and evaluation," *ACM Transactions on Information Systems (TOIS)*, vol. 15, no. 4, pp. 368–399, 1997.
[16] S. E. Robertson and I. Soboroff, "The TREC 2002 filtering track report." in *TREC*, vol. 2002, no. 3, 2002, 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 *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2006, pp. 186–193.
[18] S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in *6th International Conference on Data Mining, ICDM'06*. IEEE, 2006, pp. 1157–1161.
[19] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 1, pp. 30–44, 2012.
[20] J. Lafferty and C. Zhai, "Probabilistic relevance models based on document and query generation," in *Language Modeling for Information Retrieval*. Springer, 2003, pp. 1–10.
[21] L. Azzopardi, M. Girolami, and C. Van Rijsbergen, "Topic based language models for ad hoc information retrieval," in *Neural Networks, 2004. Proceedings. IEEE International Joint Conference on*, vol. 4. IEEE, 2004, pp. 3281–3286.
[22] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data," in *IJCAI*, vol. 3, 2003, pp. 587–592.
[23] J. Fürnkranz, "A study using n-gram features for text categorization," *Austrian Research Institute for Artificial Intelligence*, vol. 3, no. 1998, pp. 1–10, 1998.
[24] W. B. Cavnar, J. M. Trenkle *et al.*, "N-gram-based text categorization," *Ann Arbor MI*, vol. 48113, no. 2, pp. 161–175, 1994.
[25] J.-F. Boulicaut, A. Bykowski, and C. Rigotti, "Free-sets: a condensed representation of boolean data for the approximation of frequency queries," *Data Mining and Knowledge Discovery*, vol. 7, no. 1, pp. 5–22, 2003.
[26] A. Bykowski and C. Rigotti, "Dbc: a condensed representation of frequent patterns for efficient mining," *Information Systems*, vol. 28, no. 8, pp. 949–977, 2003.
[27] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 55–86, 2007.
[28] X. Yi and J. Allan, "A comparative study of utilizing topic models for information retrieval," in *Advances in Information Retrieval*. Springer, 2009, pp. 29–41.
[29] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2001, pp. 334–342.
[30] Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2002, pp. 81–88.
[31] C. Zhai, "Statistical language models for information retrieval," *Synthesis Lectures on Human Language Technologies*, vol. 1, no. 1, pp. 1–141, 2008.
[32] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2007, pp. 490–499.
[33] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 977–984.
[34] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in *7th IEEE International Conference on Data Mining, ICDM'2007*. IEEE, 2007, pp. 697–702.
[35] H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, "Enriching text representation with frequent pattern mining for probabilistic topic modeling," *Proceedings of the American Society for Information Science and Technology*, vol. 49, no. 1, pp. 1–10, 2012.
[36] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," in *Proceedings of the 23rd Annual International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2000, pp. 33–40.
[37] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.
[38] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic pattern-taxonomy extraction for web mining," in *Proceedings. IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE, 2004, pp. 242–248.
[39] H. S. Christopher D. Manning, Prabhakar Raghavan, *An Introduction to Information Retrieval*. Cambridge University Press, 2009.
[40] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.
[41] A. Tagarelli and G. Karypis, "A segment-based approach to clustering multi-topic documents," *Knowledge and information systems*, vol. 34, no. 3, pp. 563–595, 2013.



Yang Gao is a doctorate student of school of Electrical Engineering and Computer Science, Queensland University of Technology. Her research interests include data mining and web intelligence. In particular, her research focuses topic modelling associated with text mining in applications of information filtering, information retrieval and recommender system. Papers have been published in conferences such as WWW, WISE, ICDM and PAKDD.

Yue Xu is an Associate Professor in the School of Electrical Engineering and Computer Science, Queensland University of Technology, Australia. She has worked in knowledge based systems for many years. Her current research interests are focused on web intelligence and data mining. A/Prof. Xu has published over 140 refereed papers covering research areas of association rule and pattern mining, recommender systems, and trust and reputation management, etc. Some of the papers have been published in major conferences such as ICDM, CIKM, PAKDD, HT(Hypertext and hypermedia), WISE, WWW, Web Intelligence.

Yuefeng Li is a full professor in the School of Electrical Engineering and Computer Science, Queensland University of Technology, Australia. He has published over 150 refereed papers (including 43 journal papers). He has demonstrable experience in leading large-scale research projects and has achieved many established research outcomes that have been published and highly cited in top data mining journals and conferences (Highest citation per paper = 188). He is the Managing Editor of Web Intelligence and Agent Systems, an Associate Editor of the International Journal of Pattern Recognition and Artificial Intelligence.