Topical Pattern Based Document Modelling and Relevance Ranking

Yang Gao, Yue Xu, and Yuefeng Li

Science and Engineering,
Queensland University of Technology, Brisbane, Australia
{y21.gao,yue.xu,y2.li}@qut.edu.au

Abstract. For traditional information filtering (IF) models, it is often assumed that the documents in one collection are only related to one topic. However, in reality users' interests can be diverse and the documents in the collection often involve multiple topics. Topic modelling was proposed to generate statistical models to represent multiple topics in a collection of documents, but in a topic model, topics are represented by distributions over words which are limited to distinctively represent the semantics of topics. Patterns are always thought to be more discriminative than single terms and are able to reveal the inner relations between words. This paper proposes a novel information filtering model, Significant matched Pattern-based Topic Model (SPBTM). The SPBTM represents user information needs in terms of multiple topics and each topic is represented by patterns. More importantly, the patterns are organized into groups based on their statistical and taxonomic features, from which the more representative patterns, called Significant Matched Patterns, can be identified and used to estimate the document relevance. Experiments on benchmark data sets demonstrate that the SPBTM significantly outperforms the state-of-the-art models.

Keywords: Topic model, information filtering, significant matched pattern, relevance ranking.

1 Introduction

In information filtering (IF) models, relevance features are discovered from a training collection of documents and used to represent the user's information needs of the collection. Term-based approaches, such as Rocchio, BM25, etc [2,9], are popularly used to generate term-based features due to their efficient computational performance as well as mature theories. But the 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 utilise patterns to represent users' interest and achieved some improvements in effectiveness [5] since patterns carry more semantic meaning than terms. Also, some data mining techniques have been developed to remove redundant and noisy patterns for improving the quality of the discovered patterns, such as maximal patterns, closed patterns, master patterns, etc [1,17,19],

B. Benatallah et al. (Eds.): WISE 2014, Part I, LNCS 8786, pp. 186–201, 2014.

some of which have been used for representing user information needs in IF systems [21]. 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 policy, market, etc. At any time, new topics may be introduced in the document stream, which means the user's interest can be diverse and changeable. 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 [3, 13, 15] has become one of the most popular probabilistic text modelling techniques and has been quickly been accepted by many communities. The most inspiring contribution of topic modelling is that it automatically classifies documents in a collection by a number of topics and represents every document with multiple topics and their corresponding distribution. Latent Dirichlet Allocation (LDA) [3] is the most effective topic modelling method. It is reasonable to expect that applying LDA to IF could create a breakthrough for current IF models. However, there are two problems in directly applying LDA to IF. 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 LDA 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 not often representative. Our previous work [7] incorporated data mining into topic modelling and generated pattern-based topic representation, which discovers the associations of words inner topics and alleviates the problem of semantic ambiguity of the topic representations in LDA model. However, the pattern-based topic representation can only represent the collection rather than modelling individual documents. How to utilize the pattern-based topic modelling for document representation is still an open question.

In this paper, we propose a new model, called Significant matched Pattern-based Topic Model (SPBTM), in which two parts are involved, user interest modelling (also called "document modelling" since the user interest is generated based on a collection of documents) and document relevance ranking. The user interest model is represented in terms of multiple topics and each topic is represented by patterns. More importantly, the patterns are organized into groups, called equivalence classes, based on their statistical and taxonomic features. With the structured representation, the set of more representative patterns can be identified to represent the user's information needs. Based on the user's interest model, significant matched patterns are selected to determine the relevance of a new coming document.

The remainder of this paper is organized as follows. Section 2 provides a brief background of work LDA. Section 3 and 4 presents the details of our proposed model. Then, we describe data sets, baseline models and empirical results in

Section 5. Section 6 reports related discussions, followed by related work. At last, Section 8 concludes the whole work and presents the future work.

2 Background

Latent Dirichlet Allocation (LDA) [3] 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 every 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. In LDA model, Gibbs sampling method is a very effective strategy for hidden parameters estimation [11] that is used in this paper. 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 $\theta_{d_i} = (\vartheta_{d_i,1}, \vartheta_{d_i,2}, \cdots, \vartheta_{d_i,V})$. 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. 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. Assuming the documents in D involve 3 topics, Z_1, Z_2 and Z_3. Table 1 illustrates the topic distribution over documents and 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.

Table 1. Example results of LDA: word-topic assignments

Topic		Z_1		Z_2		Z_3
d	$\vartheta_{d,1}$	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}

3 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.

3.1 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, 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.

3.2 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 [7], 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 $\frac{supp(X)}{|\Gamma_j|}$. 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 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 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

Table 3. 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

4 Information Filtering Model Based on Pattern Enhanced LDA

The representations generated by the pattern enhanced LDA model, discussed in Section 3, 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, Significant matched Pattern-based Topic Model (SPBTM), is proposed based on the pattern enhanced topic representations. The proposed model consists of topics distribution describing topic preferences of documents or a document collection and structured pattern-based topic representations representing the semantic meaning of topics in documents. Moreover, the proposed model estimates the relevance of incoming documents based on Significant Matched Patterns, which are the more relevant and representative patterns, as proposed in this paper. The details are described in the following subsections.

4.1 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 [1] and closed patterns. 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 [17,19].

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 $supp(X) = supp(g)$. A generator g of X is said a minimal generator of X if $\nexists g' \subset g$ and g' is a generator of X.

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.

Table 4. The equivalence classes in Z_2

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

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

4.2 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 mentioned 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 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 3, 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, and the patterns in each $\mathbf{X}_{Z_i}$ represent what the user is interested in terms of topic Z_i.

Frequent patterns can be well organized into groups based on their statistics and coverage. As discussed in Section 4.1, equivalence class is a useful structure which collects the frequent patterns with the same frequency in 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 SPBTM, the equivalence classes $\mathbb{E}(Z_i)$ are used to represent user interests which are denoted as $\mathbb{U}_E = \{\mathbb{E}(Z_1), \cdots, \mathbb{E}(Z_V)\}$.

4.3 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. For a new incoming document d, the basic way to determine the relevance of d to the user interests is firstly to identify significant 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 pattern has 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 and thus the size of the pattern. 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 $0 < m < 1$, $|X|$ is the length of X, i.e. the number of words in X.

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

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

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 $a = 1$).

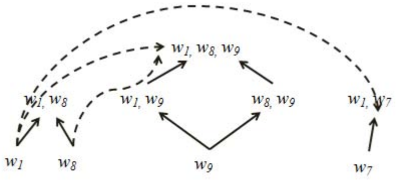

Fig. 1. Pattern Taxonomy in Z_2

In the SPBTM model, the topic significance is determined by significant matched pattern which is defined below.

Definition 6. *Significant Matched Patterns (SMPatterns)*: 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 X in d is considered a *matched pattern* to equivalence class to equivalence class EC_{jk}, if $X \in EC_{jk}$. Let c_{jk} be the closed pattern in EC_{jk}, a matched pattern X to EC_{jk} is considered a *significant matched pattern* to EC_{jk} if $\eta_X = \dfrac{|X|}{|c_{jk}|} \geq \varepsilon$, where $\varepsilon \in [0, 1]$ is the threshold for determining the significant pattern, the higher the η_X, the more significant the significant pattern is.

The set of all SMPatterns, denoted as SM^d_{jk}, to equivalence class EC_{jk} are those matched patterns which are significantly close to the closed pattern and only a proportion (controlled by ε) of all the matched patterns in EC_{jk} are selected. Therefore, the SMPatterns SM^d_{jk}, where $k = 1, \cdots, n_j$ are considered the significant patterns in d which can represent the relevant topic Z_j.

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 SPBTM, the patterns PA^d_{jk} in the topic significance $sig\left(Z_j, d\right)$ are SMPatterns in $\mathbb{U}_E$. And the specificity is calculated by the closed pattern c_{jk} in E_{jk} and η_X which represents the degree of the significance of the matched pattern X in the specific equivalence class. By incorporating Equation (1) into Equation (2), the relevance ranking of d, denoted as $Rank_E(d)$, is estimated by the following equation:

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

where V is the total number of topics, SM^d_{jk} is the set of significant 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 $Rank(d)$, the more likely the document is relevant to the user's interest.

5 Evaluation

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

5.1 Data and Measures

The Reuters Corpus Volume 1 (RCV1) dataset was collected by Reuter's journals between August 20, 1996, and August 19, 1997, incorporating a total of 806,791 documents that cover 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 [4], 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, which are used for experiments in this paper, and the 'title' and 'text' of the documents are used by all the models.

The effectiveness is assessed by five different measures: Mean Average Precision (MAP), average precision of the top K ($K = 20$) documents, break-even point (b/p), $F_\beta(\beta = 1)$ measure 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.

5.2 Baseline Models and Settings

The experiments were conducted extensively covering all major representations such as phrases and patterns in order to evaluate the effectiveness of the proposed topic-based IF model. The evaluations were conducted in terms of two technical categories: topic modelling methods and pattern mining methods. For each category, some state-of-the-art methods are chosen as the baseline models. More details about these baseline models are given below.

(1) *Topic modelling based category*

TNG: In the phrase-based topic model, n-gram phrases that are generated by using the TNG model [14], which can be used to represent user interest needs and phrase frequency is used to represent topic relevance. Readers who are interested in the details can refer to [14].

PBTM: We have proposed a topic-based model PBTM_FCP [6] which uses closed patterns to represent topics and uses patterns' support to represent topic relevance. PBTM_FCP is chosen as a baseline model for the pattern-based topic models. The following equation is used to calculate the relevance of a document d with PBTM_FCP:

$$Rank_C(d) = \sum_{j=1}^{V} \sum_{k=1}^{n_j} |c_{jk}|^{0.5} \times \delta(c_{jk}, d) \times f_{jk} \times \vartheta_{D,j} \tag{5}$$

where c_{jk} is a closed pattern in PBTM_FCP and n_j is the total number of closed patterns in topic j.

The parameters for all 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 $\beta = 0.01$. Our experience shows that filtered results are not very sensitive to the settings of these parameters. But the number of topics V affects the results depending on various data collections. In this paper, V is set to 10.

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 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 :

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

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 information needs. The minimum support in the pattern-based models, including phrases, sequential closed patterns, is set to 0.2.

Sequential Closed Patterns(SCP): The Pattern Taxonomy Model [21] is one of the state-of-the-art pattern-based model. It was developed to discover sequential closed patterns from the training dataset and rank the incoming documents in the filtering stage with the relative supports of the discovered patterns that appear in the documents.

***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 an word-based representation. In our experiments, we use n-Gram phrases to represent a document collection (i.e. user information needs), where $n = 3$.

5.3 Results

Five different thresholds ($\varepsilon = 0.3, 0.4, 0.5, 0.6, 0.7$) are used in order to find proper SMPatterns in the proposed SPBTM using the 50 human assessed collections and the results are shown in Table 5. Based on the comparison in Table 5, SPBTM achieves the best result when $\varepsilon = 0.5$ for this dataset. Therefore, this result is used to compared with all the baseline models mentioned above. The results are depicted in Table 6 and evaluated using the measures in Section 5.1.

Table 6 consists of two parts. The top and bottom parts in Table 6 provide the results of the topic modelling methods and the pattern mining methods, respectively. The *improvement%* line at the bottom of each part provides the percentage of improvement achieved by the SPBTM which consistently performs the best among all models against the second best model in that part for each measure.

We also conducted the T-Test to compare the SPBTM with all baseline models. The results are listed in Table 7. The statistical results indicate that the

Table 5. Comparison of SPBTM results with different values of threshold ε, using the first 50 collections of RCV1

Threshold ε	*MAP*	*b/p*	*top20*	F_1
0.3	0.452	0.436	0.513	0.445
0.4	0.455	0.436	0.521	0.445
0.5	0.456	0.446	0.524	0.446
0.6	0.449	0.433	0.513	0.439
0.7	0.442	0.425	0.515	0.435

Table 6. Comparison of all models using the first 50 collections of RCV1

Methods	*MAP*	*b/p*	*top20*	F_1
SPBTM	**0.456**	**0.446**	**0.524**	**0.446**
TNG	0.374	0.367	0.446	0.388
PBTM_FCP	0.424	0.420	0.494	0.424
improvement%	7.5	6.2	6.1	5.2
SCP	0.364	0.353	0.406	0.390
n-Gram	0.361	0.342	0.401	0.386
FCP	0.361	0.346	0.428	0.385
improvement%	25.3	26.3	22.4	14.4

Table 7. T-Test *p*-values for all modes compared with the SPBTM model

Methods	*MAP*	*b/p*	*top20*	F_1
TNG	0.0003	0.0005	0.0066	0.0002
PBTM_FCP	0.0005	0.0299	0.0267	0.0002
SCP	0.00004	0.0001	0.0002	0.0002
n-Gram	0.0001	0.0001	0.0004	0.0002
FCP	0.00002	0.00004	0.0031	0.0001

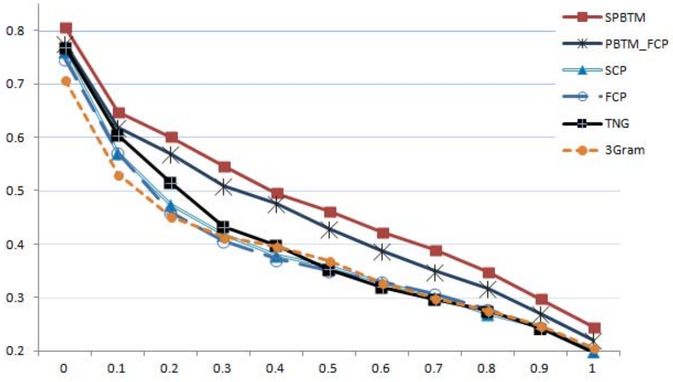

Fig. 2. 11 point results of comparison between the proposed SPBTM and baseline models

proposed SPBTM significantly outperforms all the other models (all values in Table 6 are less than 0.05) and the improvements are consistent on all four measures. Therefore, we conclude that the SPBTM 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 SPBTM model has achieved the best performance compared with all the other baseline models.

6 Discussion

As we can see from the experiment results, taking topics 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 SPBTM 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. These comparisons can strongly validate the first hypothesis. Moreover, the outstanding performance of the SPBTM over the PBTM_FCP indicates the significant benefit of using the proposed SMPatterns in estimating document relevance over using frequent closed patterns. This result clearly supports the second hypothesis.

6.1 Significant Matched Patterns

In the SPBTM, 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. However, among all matched patterns in each equivalence class, not all of them are useful for estimating the document relevance. The results in Table 5 show that the best performance achieved by the SPBTM is when the threshold ε is 0.5. This result indicates that, selecting more matched patterns as SMPatterns (i.e., $\varepsilon < 0.5$) actually hurts the performance of document relevance ranking. When ε is small, some short matched patterns would be selected. These short patterns are much less specific than longer patterns to represent the documents and also possibly brings bias to the document relevance ranking. Similarly, the performance also deteriorates when selecting less matched patterns (i.e., $\varepsilon > 0.5$). This is because some useful matched patterns will not be selected due to the high threshold, which will negatively affect the quality of the selected SMPatterns.

From Table 6, we can see that the PBTM_FCP achieved better performance than all the other models but SPBTM. SPBTM is the only model which outperforms the PBTM_FCP. This result is an excellent example to show the quality of closed patterns as well as SMPatterns.

6.2 Topic-Based Relevance Estimation

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

6.3 Complexity

As discussed in Section 4, there are two algorithms in the proposed model, i.e. user profiling and document filtering. The complexity of the two algorithms 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 is linear with the number of topics (V) and the number of documents (N), i.e. $O(V * N)$ [15]. For pattern mining, the efficiency of the FP-Tree algorithm for generating frequent patterns has been widely accepted in the field of data mining. 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 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), $O(S)$ 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 in each topic representation. Theoretically, the complexity of the topic-based methods is higher than the pattern-based or term-based methods but practically, the number of SMPatterns is much smaller than the number of frequent patterns. Therefore, the complexity of the SPBTM model is very often acceptable.

7 Related Work

Documents can be modelled by various approaches that primarily include term-based models [2,9], pattern-based models [16,21] and probabilistic models [8,10]. 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. Aiming at representing documents with multiple topics in a more detailed way, topic models are incorporated in the frame of language model and achieve successful retrieval results [15, 18]. Also, topic models [12, 13, 20] can extract user information needs by analysing content and represent them in terms of latent topics discovered from user profiles. But in all of these topic models, a fundamental assumption is a topic can be represented by a word-based multinomial distribution. Thus, it is desirable to interpret topics or documents with coherent and discriminative representations. TNG model generated topical phrases has achieved a slight improvement on IR task [14] and IF task from our experiment results in this paper, which mainly because of too limited occurrences of the discovered phrases to represent the document relevance.

8 Conclusion

This paper presents an innovative pattern enhanced topic model for information filtering including user interest modelling and document relevance ranking. The SPBTM generates pattern-based topic representations to model user's information interests across multiple topics. In the filtering stage, the SPBTM selects SMPatterns, 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 SMPatterns. The proposed model has been evaluated by using the RCV1 and TREC collections for the task of information filtering. Compared 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 topic modelling techniques and data mining techniques. Moreover, the significant topical patterns for incoming documents can effectively represent user's interests. The technique not only can be used for information filtering, but also can be applied to many content-based user interest modelling tasks.

References

1. Bayardo Jr., R.J.: Efficiently mining long patterns from databases. ACM Sigmod Record 27, 85–93 (1998)
2. Beil, F., Ester, M., Xu, X.: Frequent term-based text clustering. In: KDD 2002, pp. 436–442. ACM (2002)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
4. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: SIGIR 2000, pp. 33–40. ACM (2000)
5. Cheng, H., Yan, X., Han, J., Hsu, C.-W.: Discriminative frequent pattern analysis for effective classiffication. In: ICDE 2007, pp. 716–725. IEEE (2007)
6. Gao, Y., Xu, Y., Li, Y.: Pattern-based topic models for information filtering. In: Proceedings of International Conference on Data Mining Workshop SENTIRE, ICDM 2013. IEEE (2013)

7. Gao, Y., Xu, Y., Li, Y., Liu, B.: A two-stage approach for generating topic models. In: PADKDD 2013, pp. 221–232 (2013)
8. Lafferty, J., Zhai, C.: Probabilistic relevance models based on document and query generation. In: Language modeling for information retrieval, pp. 1–10. Springer, Heidelberg (2003)
9. Robertson, S., Zaragoza, H., Taylor, M.: Simple bm25 extension to multiple weighted fields. In: CIKM 2004, pp. 42–49. ACM (2004)
10. Sparck Jones, K., Walker, S., Robertson, S.E.: A probabilistic model of infor- mation retrieval: development and comparative experiments: Part 2. Information Processing & Management 36(6), 809–840 (2000)
11. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of Latent Semantic Analysis 427(7), 424–440 (2007)
12. Tang, J., Wu, S., Sun, J., Su, H.: Cross-domain collaboration recommendation. In: KDD 2012, pp. 1285–1293. ACM (2012)
13. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: KDD 2011, pp. 448–456. ACM (2011)
14. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: ICDM 2007, pp. 697–702. IEEE (2007)
15. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: SIGIR 2006, pp. 178–185. ACM (2006)
16. Wu, S.-T., Li, Y., Xu, Y.: Deploying approaches for pattern refinement in text mining. In: ICDM 2006, pp. 1157–1161. IEEE (2006)
17. Xu, Y., Li, Y., Shaw, G.: Reliable representations for association rules. Data & Knowledge Engineering 70(6), 555–575 (2011)
18. Yi, X., Allan, J.: A comparative study of utilizing topic models for information retrieval. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 29–41. Springer, Heidelberg (2009)
19. Zaki, M.J., Hsiao, C.-J.: CHARM: An efficient algorithm for closed itemset mining. In: SDM, vol. 2, pp. 457–473 (2002)
20. Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering. In: SIGIR 2002, pp. 81–88. ACM (2002)
21. Zhong, N., Li, Y., Wu, S.-T.: Effective pattern discovery for text mining. IEEE Transactions on Knowledge and Data Engineering 24(1), 30–44 (2012)