Maximum Matched Pattern-based Topic Model for Information Filtering

Greeshma G S
PG Scholar
Computer Science and Engineering
Younus College of Engineering and Technology,
Kollam, India-691010

Abisha A
Assistant Professor
Information Technology
Younus College of Engineering and Technology,
Kollam, India-691010

***Abstract*—Many techniques have been used in the field of information filtering to generate user's information needs from a collection of documents. A fundamental assumption for all techniques is that the documents in the collection are all about one topic. But the User's interests can be diverse and the documents in the collection often involve multiple topics. Topic modelling algorithms are used to find the hidden topics in a collection of documents.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 (MMPBTM), is proposed to estimate the document relevance to the users information needs in order to filter out irrelevant documents.**

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

1. INTRODUCTION

Recent years have witnessed a dramatic increase in web information. Statistics from Google in social blog state that the web pages indexed by Google numbered around one million in 1998, quickly reached one billion in 2000 and had already exceeded one trillion in 2008. According to Google's latest report this number has reached 60 trillion. Hence, advanced programs and formulas are required to understand what exactly users need and to deliver the best results based on user's information needs. This process contains two dominant components: user interest modelling and relevance ranking.

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. Information filtering is concerned with the problem of selecting the information relevant to the needs of the individuals. Users of a filtering system specify their needs in a profile reflecting their long term wants, i.e. information needs, interests and preferences, relevant to their work, use these profiles to automatically match them with the incoming information. Filter profiles could be constructed to reflect the needs of a group of individuals as to cover their common fields of interests.

Traditional IF models were developed based on a term-based approach. The content of each item is represented as a set of descriptors or terms, typically the words that occur in a document. The user profile is represented with the same terms and built up by analyzing the content of items which have been seen by the user. Several issues have to be considered when implementing a term-based filtering system. First, terms can either be assigned automatically or manually. When terms are assigned automatically a method has to be chosen that can extract these terms from items. Second, the terms have to be represented such that both the user profile and the items can be compared in a meaningful way. Third, a learning algorithm has to be chosen that is able to learn the user profile based on seen items and can make recommendations based on this user profile.

The information source that term-based filtering systems are mostly used with is text documents. A standard approach for term parsing selects single words from documents. The vector space model and latent semantic indexing are two methods that use these terms to represent documents as vectors in a multi dimensional space. The advantage of term-based approach is the efficient computational performance, as well as mature theories for term weighting, like Rocchio, BM25. 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 for information filtering and achieved some improvements on effectiveness, since patterns carry more semantic meaning than terms. Also, data mining has developed some techniques (i.e., maximal patterns, closed patterns and master patterns) for removing the redundant and noisy pattern. One of the promising techniques is Pattern Taxonomy Model (PTM) that discovered closed sequential patterns in text classifica-

tion. It shows a certain extent improvement on effectiveness, but still faces one challenging issue which is low frequency of the patterns appearing in documents. In order to solve this problem, Wu proposed deploying pattern approach to weight terms by calculating their appearance in discovered patterns. All these data mining and text mining techniques hold the assumption that users interest is only related to a single topic. However, the reality is that multiple semantic topics are involved. Topic modelling has become one of the most popular probabilistic text modelling techniques and quickly been accepted by machine learning and text mining 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.

Topic modelling 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) and LDA. 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.

This paper proposes 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 MMPBTM is proposed for document representation and document relevance ranking. The patterns in the MMPBTM are well structured so that the maximum matched patterns can be efficiently and effectively selected and used to represent and rank documents.

2. LITERATURE SURVEY

Information filtering is concerned with the problem of selecting the information relevant to the needs of the individuals. Users of a filtering system specify their needs in a profile reflecting their long term wants, i.e. information needs, interests and preferences, relevant to their work, use these profiles to automatically match them with the incoming information. User profiles could be constructed to reflect the needs of a group of individuals as to cover their common fields of interests.

IF systems were originally considered to have the same function as IR systems did. Different from IR systems, IF systems were commonly personalized to support long-term information needs of users [1]. The main distinction between IR and IF was that IR systems used queries but IF systems acquire user information needs from user profiles. The representation of the user information need is variously referred to as user profiles, or topic profiles. As the quality of the profiles directly influences the quality of information filtering, the issue of how to built accurate, reliable profiles is a crucial concern [2]. The tasks of the filtering track in TREC included batch and routing filtering, and adaptive filtering. A batch filtering system uses a retrieval algorithm to score each incoming document. If the score is greater than a specified threshold, then the document is delivered to the user. The routing filtering systems are more similar to the retrieval systems, the profile remains constant and the task is to match an incoming stream of documents to a set of profiles. Both systems need to return a ranked list of documents.

The term-based IF systems used terms to represent the user profiles. Such profiles are the most simplest and common representation of the profiles. For examples: the probabilistic models [4], BM25 [5], rough set-base models [6], [7], and ranking SVM [8] based filtering models used the term-based user profiles. The advantage of term- based model is efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term-based models suffer from the problems, such as; the relationship among the words cannot be reflected [8] and also, only considering single words as features is the semantic ambiguity. For example: the synonym problem is a word that shares the same meaning as another word (for example,“taxi”and “cab”), and the homonym problem is a word that is pronounced, and sometimes spelled, in the same way as another word but has a different meaning (for example, “there”and “their”).

Phrase-based method is therefore proposed. This method used the multiple words (phrases) as features to solve the semantic ambiguity problem. It is believed that the simple term-based representation of the profile is usually inadequate, because single words are rarely sufficiently specific for accurate discrimination. However, Fuhr [9] investigated the probabilistic models in IR and pointed out that a dependent model for phrases is not sufficient, because only the occurrence of the phrase components in a document is considered, but not the syntactical structure of the phrases. Moreover, the certainty of identification should also be regarded, such as, whether the words occur adjacent or only within the same paragraph.

Data mining techniques were applied to text mining and classification by using word sequences as descriptive phrases (n-Gram) from document collections [8]. But the performance of n-Gram is highly restricted due to low frequency of phrases. Pattern mining has been extensively studied for many years. A variety of efficient algorithms such as Apriori-like algorithms [3], FP-tree has been proposed. In the field of text mining, pattern mining techniques can be used to find various text patterns, such as co-occurring terms and multiple grams, maximal frequent patterns, and closed patterns, for building up a representation with these new types of features.

Data mining techniques have been used for text analysis

by extracting co-occurring terms as descriptive phrases from document collections[16]. However, the effectiveness of the text mining systems using phrases as text representation showed no significant improvement. Mining maximal frequent patterns was also proposed to reduce the time complexity of mining all frequent patterns, where an itemset (or a pattern) was maximal frequent if it had no superset that was frequent. The similar idea, maximal association rules, was also used for text mining where users provided categories for finding maximal rules they wanted.

Maximal association mining ignored all of small patterns. However, some small patterns can be very useful. Closed patterns were used to prune some smaller useless patterns and that have been used for improve the effectiveness of text mining. Typically, text mining discusses associations between terms at a broad spectrum level, paying little heed to duplications of terms, and labeled information in the training set. Usually, the existing data mining techniques return numerous discovered patterns (e.g., sets of terms) from a training set. Not surprisingly, among these patterns, there are many redundant patterns [17]. Nevertheless, the challenging issue is how to effectively deal with the large amount of discovered patterns.

The LDA-based document models [6] are state of- art topic modelling approaches. Information retrieval systems based on these models achieved good performance. The authors claimed the retrieval performance achieved by [18] not only because of the multiple topics document model, but also because that each topic in the topic model is represented by a group of semantically similar words, which solve the synonymy problem of single words. The relevant documents are determined by user-specific topic model that has been extracted from user information needs. These topic model based applications are all related to long-term user needs extraction and related to the task of this paper. But, there is a lack of explicit discrimination in most of the language models based approaches and probabilistic topic models.

3. PROPOSED SYSTEM

This paper proposes 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 MMPBTM is proposed for document representation and document relevance ranking. The patterns in the MMPBTM are well structured so that the maximum matched patterns can be efficiently and effectively selected and used to represent and rank documents. The architecture of the proposed system is shown in fig 1.

The original contributions of the proposed MMPBTM to the field of IF can be described as follows:

• To model users interest with multiple topics rather than a single topic under the assumption that users information interests can be diverse.

• 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 MMPBTM 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.

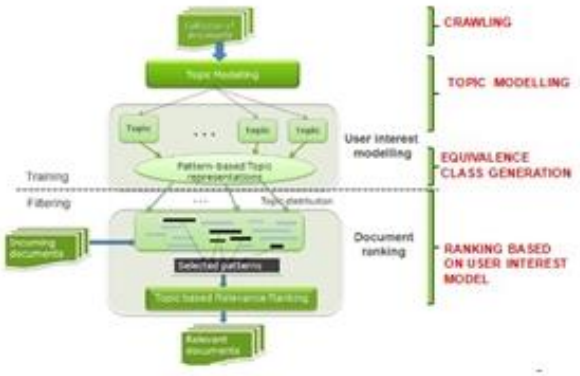

Figure 1. Architecture of Proposed System

• Proposes 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.

• Proposes 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 users interest. The maximum matched patterns are the most representative and discriminative patterns to determine the relevance of incoming documents.

4. TOPIC MODELLING

Topic modelling algorithms are used to discover a set of hidden topics from collections of documents, where a topic is a distribution over terms. Topic models provide an interpretable low-dimensional representation of documents.

Latent Dirichlet Allocation (LDA) [20] 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 with the appearing words. Let $D = \{d_1, d_2, ...d_M\}$ 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 involving multiple topics and each topic can be defined as a distribution over fixed vocabulary of terms that appear

Topic	DataW arehouse(Z_1)		DataMining(Z_2)		Database(Z_3)	
Document	$\vartheta_{d,1}$	words	$\vartheta_{d,2}$	words	$\vartheta_{d,3}$	words
d_1	4.25	Analysis, data, multidimensional w_5, w_1, w_{10}	2.5	Analysis, mining, knowledge w_5, w_4, w_6	3.25	Database, information w_7, w_3
d_2	2.86	Data, query w_1, w_9	3.93	Information, analysis, knowledge w_3, w_5, w_6	3.25	Query, language, relational w_9, w_8, w_{11}
d_3	4.36	Warehouse, analysis, information w_1, w_{12}, w_5, w_3	2.18	Information, data, multidimensional w_3, w_1, w_{10}	3.45	Database, information, query w_7, w_3, w_9, w_8
d_4	4.44	Data, heterogenous, information w_1, w_2, w_3	2.5	Mining, knowledge, analysis w_4, w_5, w_6	3.06	Database, query. Language w_7, w_9, w_8

TABLE 1. EXAMPLE RESULTS OF LDA: WORD-TOPIC ASSIGNMENTS

in documents. Specifically, LDA models a document as a probabilistic mixture of topics and treats each topic as a probability distribution over words. For the ith word in document d, denoted as $w_{d,i}$, the probability of $w_{d,i}$, $P(w_{d,i})$ is defined as:

$$P(w_{d,i}) = \sum_{j=1}^{V} P(w_{d,i} \mid z_{d,i} = Z_j) \times P(z_{d,i} = Z_j) \quad (1)$$

$z_{d,i}$ is the topic assignment for $w_{d,i}$, $z_{d,i} = Z_j$ means that the word $w_{d,i}$ is assigned to topic j and the V represents the total number of topics. The resulting representations of LDA are at two levels, collection level and document level. At document level, each document d_i is represented by topic distribution $\theta_{d,i}$. 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 $\Phi=\{\varphi_1, \varphi_2, \ldots \varphi_V\}$ for all topics. Based on the distribution of Φ for the whole collection, D can be represented by topics distribution, $\theta_d = (\vartheta_{D,1}, \vartheta_{D,2}, \ldots \vartheta_{D,V})$ $\vartheta_{D,j}$ indicates the proportion of topic j in the collection D.

Apart from these two level outcomes, LDA also generates word-topic assignment, 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, data warehouse(Z_1), data mining(Z_2) and database(Z_3).Table 1 illustrates the topic distribution over the documents and the word-topic assignments in this small collection.

5. PATTERN BASED TOPIC MODELLING

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

5.1. Construct Transactional Dataset

Let $R_d i, Zj$ represent the word-topic assignment to topic Z_j in document d_i. $R_d i, Zj$ is a sequence of words assigned to topic Z_j. Construct a set of words from each word-topic assignment $R_d i, Zj$ instead of using the sequence of words in $R_d i, Zj$, because for pattern mining, the frequency of a word within a transaction is insignificant.Let $I_{i,j}$ be a set of words which occur in $R_d i, Zj$. $I_{i,j}=\{w|w\varepsilon R_d i, Zj\}$ $I_{i,j}$ called a topical document transaction, is a set of words without any duplicates. From all the word-topic assignments $R_d i, Zj$ to Z_j we can construct a transactional dataset Γ_j. Let $D = \{d_1, \ldots d_M\}$ be the original document collection, the transactional dataset Γ_j for topic Z_j is defined as $\Gamma_j = \{I_{1j}, I_{2j}, \ldots I_{Mj}\}$. For the topics in D, we can construct V transactional datasets. An example of transactional datasets is illustrated in table 2, which is generated from the example in table 1.

transaction	Data warehouse (Γ_1)	Data mining (Γ_2)	Database (Γ_3)
1	$\{w_5, w_1, w_{10}\}$	$\{w_5, w_4, w_6\}$	$\{w_7, w_3\}$
2	$\{w_1, w_9\}$	$\{w_3, w_5, w_6\}$	$\{w_9, w_8, w_{11}\}$
3	$\{w_1, w_{12}, w_5, w_3\}$	$\{w_3, w_1, w_{10}\}$	$\{w_7, w_3, w_9, w_8\}$
4	$\{w_1, w_2, w_3\}$	$\{w_4, w_5, w_6\}$	$\{w_7, w_9, w_8\}$

TABLE 2. TRANSACTIONAL DATASETS GENERATED FROM WORD TOPIC ASSIGNMENTS

5.2. Generate Pattern Based Representation

The basic idea of the proposed pattern based method is to use patterns generated from each transactional dataset Γ_j to represent Z_j. In the two-stage topic model, frequent patterns are generated in this step. For a given minimal support threshold σ an item set 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. 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.

Patterns	supp
$\{w_5\}, \{w_6\}, \{w_5, w_6\}$	3
$\{w_3\}, \{w_4\}, \{w_4, w_5\}, \{w_4, w_6\}, \{w_4, w_5, w_6\}$	2

TABLE 3. THE FREQUENT PATTERNS

6. STRUCTURED PATTERN BASED TOPIC MODEL FOR INFORMATION FILTERING

Representations generated by pattern based LDA carry more concrete and identifiable meaning than the word based representations generated using original LDA. However, the number of patterns in some of the topics can be huge and many of the patterns are not discriminative enough to represent different topics. As a result, documents are hardly accurately represented by these topic representations. That means, these pattern based topic representations which represent user interests may be not 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, MMPBTM, 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.

6.1. Pattern Equivalence Classes

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 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[].

Closed Itemset: for a transactional dataset, an itemset X is a closed itemset if there exists no itemset X' such that $X \subset X'$ and supp(X)=supp(X').Closed pattern reveals the relations of the largest range of the associated terms.

It covers all information that its subsets describe. Closed patterns are more effective and efficient to represent topics than frequent patterns.

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

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)∪{X}

Let EC_1 and EC_2 be two different equivalence classes of the same transactional dataset. Then $EC_1 \cap EC_2 = \varphi$, 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 fn indicates the statistical significance of each class.

EC_{21} (f_{21} = 0.75)	EC_{22} (f_{22} = 0.5)	EC_{23} (f_{23} = 0.5)
{w_5 },{w_6 }	{w_4 ,w_5 ,w_6 }	{w_3 }
{w_5	{w_4 ,w_5 }	
{w_6 }	{w_4 ,w_6 }	
	{w_4 }	

TABLE 4. THE EQUIVALENCE CLASS IN Z_2

6.2. User Interest Modelling

For a collection of documents D, by using the pattern based model we can generate the user's interests $U = \{X_{Z1}, X_{Z2}, ..., X_{ZV}\}$, $X_{Zi} = \{X_{i1}, X_{i2}, ..., X_{imi}\}$, where X_{Zi} is the pattern based representations for Z_i and m_i is the total number of patterns in X_{Zi} V is the total number of topics.

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.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 X_{Zi}, which are $c_{i1}; \ldots; c_{in}i$, and that X_{Zi} can be partitioned into n_i equivalence classes, EC (c_{i1}), . . ., EC ($c_{in}i$). For simplicity, the equivalence classes are denoted as EC_{i1}, . ., EC_{in} for X_{Z1}, or simply for topic Z_i. Let $E(Z_i)$ denote the set of equivalence classes for topic Z_i, i.e. $E(Z_i) = \{EC_{i1}, . . ., EC_{in}i\}$. In the model MMPBTM, the equivalence classes $E(Z_i)$ are used to represent user interests which are denoted as $U_E = \{E(Z_1), ..., E(Z_V)\}$.

6.3. Document 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. 2 depicts the taxonomy constructed for X_{Z_2} in Table 3.

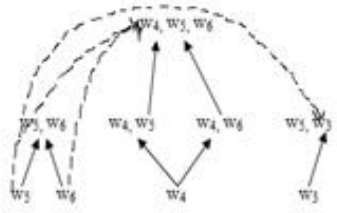

Figure 2. Pattern Taxonomy in Z_2

This tree-like structure demonstrates the subsumption relationship between the discovered patterns in Z_2. The longest pattern in a pattern taxonomy, such as $\{w_4, w_5, w_6\}$ in Fig. 2, is the most specific pattern that describes a user's interests since longer patterns have more specific meanings, while single words, such as w1 in Fig. 2, 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 $\{w_4, w_5, w_6\}$. The pattern taxonomy presents different specificities of patterns according to the level in the taxonomy structure and thus the size of the pattern.

As mentioned in pattern taxonomy, the longer the pattern is, the more specific it is. As the result, the specificity of a pattern can be estimated as a function of pattern length. For example, the 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. Pattern 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), $spe(X) = a|X|^m$ a and m are constant real numbers and $0 < m > 1$.

Topic Significance: Let d be a document, Z_j be a topic in the user interest model, PA^{jk} be matched patterns, $k = 1,...,$ n_j, to document d, and $f_{j1}, ..., f_{jn_j}$ be the corresponding supports of the matched patterns within Z_j, the topic significance of Z_j to d is defined as:

$$sig(Z_j, d) = \sum_{k=1}^{n_j} spe(PA^{jk}) \times f_{jk} = \sum_{k=1}^{n_j} |PA^{jk}|^m \times f_{jk} \quad (2)$$

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

Maximum Matched Pattern: Let d be a document, Z_j be a topic in the user interest model, $EC_{j1}, \ldots, EC_{jn_j}$ 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^d_{jk}, if the following conditions are satisfied:

- $MC^d_{jk} \in EC_{jk}$
- $MC^d_{jk} \subset d$ and $MC^d_{jk} \subset X$
- Exist X such that $X \in EC_{ik}$, $X \subseteq d$ and MC^d_{jk} to equivalence

The maximum matched pattern MC^d_{jk} 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^d_{jk}. Therefore, the maximum matched patterns MC^d_{jk}, where $k = 1,..., n_j$ 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 $d=\{w_1, w_2, w_4, w_5, w_{11}\}$, the maximum matched patterns would be $MC^d_{22} = \{w_4, w_5\}$.

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:

$$rank(d) = \sum_{j=1}^{V} sig(Z_j, d) \times \vartheta_{D,j} \quad (3)$$

For the MMPBTM, the patterns PA^{jk} in the topic significance $sig(Z_j, d$ are maximum matched patterns in U_E. By incorporating Equation (2) into Equation (3), the relevance ranking of d, denoted as $Rank_E(d)$, is estimated by the following equation:

$$Rank_E(d) = \Sigma^{j=1} \Sigma^{k_j} {}^{=1} |MC^d_{jk}|^{0.5} \times \delta(MC^d_{jk}, d) \times f_{jk} \times \vartheta_{D,j} \quad (4)$$

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

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

7. CONCLUSION

This paper presents an innovative pattern enhanced topic model for information filtering including user interest modelling and document relevance ranking. The proposed MMPBTM model generates pattern enhanced topic representations to model user's interests accross multiple topics. In the filtering stage, the MMPBTM 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 automatically generates discriminative and semantic rich representations for modelling topics and documents by combining statistical topic modelling techniques and data mining techniques.

8. FUTURE WORK

In future along with user interest model rating by the user can be used to estimate the relevance of a document to the user.

REFERENCES

[1] F. Beil, M. Ester, and X. Xu, Frequent term-based text clustering, in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002, pp. 436442.
[2] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, Discriminative frequent pattern anal- ysis for eFFective classi cation, in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007, pp. 716725.
[3] R. J. Bayardo Jr, Efficiently mining long patterns from databases, in ACM Sigmod Record, vol. 27, no. 2. ACM, 1998, pp. 8593.
[4] S.-T. Wu, Y. Li, and Y. Xu, Deploying approaches for pattern refinement in text mining, in Data Mining, 2006. ICDM06. Sixth International Conference on. IEEE, 2006, pp. 11571161.
[5] N. Zhong, Y. Li, and S.-T. Wu, Effective pattern discovery for text mining, Knowl- edge and Data Engineering, IEEE Transactions on, vol. 24, no. 1, pp. 3044, 2012.
[6] 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. 178185.
[7] K. Sparck Jones, S. Walker, and S. E. Robertson, A probabilistic model of infor- mation retrieval: development and comparative experiments: Part 2, Information Processing and Management, vol. 36, no. 6, pp. 809840,2000.
[8] W. B. Cavnar, J. M. Trenkle et al., N-gram-based text categorization, Ann Arbor MI, vol. 48113, no. 2, pp. 161175, 1994.
[9] Y. Zhang, J. Callan, and T. Minka, Novelty and redundancy detection in adaptive ltering, in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002, pp. 8188.
[10] 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. 32813286.
[11] Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 490499. ACM.
[12] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pages 113120. ACM.
[13] Yang, Y., Aufaure, M., and Claramunt, C. (2007). Towards a DL-based semantic user model for web personalization. In Autonomic and Autonomous Systems, 2007. ICAS07. Third International Conference on, pages 6161. IEEE.
[14] Jung, S. Y., Hong, J.-H., and Kim, T.-S. (2005). A statistical model for user preference. Knowledge and Data Engineering, IEEE Transactions on, 17(6):834843.
[15] Li, Y. and Zhong, N. (2006). Mining ontology for automatically acquiring web user information needs. Knowledge and Data Engineering, IEEE Transactions on, 18(4):554568.
[16] Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Data Mining, 2007.ICDM 2007. Seventh IEEE International Con- ference on, pages 697-702. IEEE, 2007.
[17] Xujuan Zhou, Yuefeng Li, Peter Bruza, Sheng-Tang Wu, Yue Xu, and Raymond YK Lau. Using information ltering in web data mining process. In Web Intelli- gence, IEEE/WIC/ACM International Conference on, pages 163-169. IEEE, 2007.
[18] Ah-Hwee Tan et al. Text mining: The state of the art and the challenges. In Pro- ceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, volume 8, page 65, 1999.
[19] C. Buckley and E. M. Voorhees, Evaluating evaluation measure stability, in Pro- ceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000, pp. 3340.
[20] Y. Gao, Y. Xu, and Y. Li, Pattern-based topic models for information filtering, in Proc. Int. Conf. Data Min. Workshop SENTIRE, 2013, pp. 921928.