Pattern-based Topic Modelling for Query Expansion

Yang Gao **Yue Xu** **Yuefeng Li**

School of Electrical Electronics and Computer Science
Queensland University of Technology
Brisbane, Queensland, 4000
Email: [y10.gao@connect,yue.xu@,y2.li@]qut.edu.au

Abstract

One big problem with information retrieval (IR) is that the size of queries is usually short and the keywords in a query are very often ambiguous or inconsistent. Automatic query expansion is a widely recognized technique which is effective to deal with this problem. However, many query expansions methods require extra information such as explicit relevance feedback from users or pseudo relevance feedback from retrieval results. In this paper, we propose an unsupervised query expansion method, called Topical Query Expansion (TQE), which does not require extra information. The proposed TQE method expands a given query based on the topical patterns which can create links among those more associated and semantic words in each topic. This model also discovers related topics that are related to the original query. Based on the expanded terms and related topics, we propose to rank the document relevance with different ranking strategies. We conduct experiments on popularly used datasets, TREC datasets, to evaluate the proposed methods. The results demonstrate outstanding results against several state-of-the-art models.

Keywords: Topical Pattern, Information Retrieval, Query Expansion

1 Introduction

Standard bag-of-words retrieval models, such as BM25 (Lv & Zhai 2011, Robertson et al. 2004), have the benefits of mature statistical theories and their efficient computational performance which produce reasonable good results. However, these models are restricted on limited number of features that are from query terms. Besides, the words in a query may not be consistent and can be ambiguously understood or interpreted.

Automatic Query Expansion (AQE) is considered as an promising technique in IR to improve the effectiveness of document retrieval especially for short queries. Most researches of AQE techniques involve relevance feedback (Andrzejewski & Buttler 2011, Maxwell & Croft 2013) and are always combined with smoothing technique to improve the effectiveness(Yi & Allan 2009, Zhai & Lafferty 2001), etc. But one limitation of these methods in real Web search is their high cost in computation because fast response time required by Web search applications.

Targeting on the problems mentioned above, in this paper, we will propose a new IR model which includes a novel query expansion method and a novel document ranking method, both based on topical patterns generated from the document collection. The proposed query expansion method can precisely interpret the user interests even though only limited features are provided and also can precisely balance the query drift during the process of the query expansion. The proposed IR model can efficiently retrieve relevant documents because no online re-training is needed. In the real scenario, if users are not very familiar with the content related to their formulated queries, they will not be able to provide interactive refinement within a limited response time, therefore the relevance feedback models are not suitable under this situation. In contrast, in this case, the proposed IR model will be much more useful.

The new proposed Topical Query Expansion (TQE) model, can effectively expand queries with semantic topical patterns. In the model, the discovered relevant topical patterns are used to determine the certain topics and uncertain topics for a specific query. For document relevance ranking, this distinguish for the related topics can be reflected by different weighting mechanisms. The rest of this paper will be presented as follows. In Section 2, we review topic models and related state-of-the-art IR tehniques. Section 3 introduces the method of generating pattern-based topic model. Then we propose a new method for query expansion based on the pattern-based topic representation and the topic-based weighting system, which are described in Section 4. In Section 5 we describe the ad hoc retrieval experiments and prove the proposed model is significantly improve the effectiveness. According to the experimental results, we discuss the strengths of the proposed model from different perspectives in Section 6. At last, Section 7 concludes the whole work and presents the future direction.

2 Related Work

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)(Blei et al. 2003) 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

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}

$D = \{d_1, d_2, \cdots, d_M\}$ 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. 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:

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

$z_{d,i}$ 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. Let ϕ_j be the multinomial distribution over the words for Z_j, $\phi_j = (\varphi_{j,1}, \varphi_{j,2}, \cdots, \varphi_{j,n})$, $\sum_{k=1}^{n} \varphi_{j,k} = 1$. θ_d refers to multinomial distribution of the topics in document d. $\theta_d = (\vartheta_{d,1}, \vartheta_{d,2}, \cdots, \vartheta_{d,V})$, $\sum_{j=1}^{V} \vartheta_{d,j} = 1$. $\vartheta_{d,j}$ indicates the proportion of topic j in document d. LDA is a generative model in which the only observed variable is $w_{d,i}$, while the others are all latent variables that need to be estimated. Gibbs sampling method is an effective strategy for hidden parameters estimation(Steyvers & Griffiths 2007) that is used in this paper.

The resulting representations of the LDA model are at two levels, document level and collection level. Apart from these, 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.

LDA has been widely accepted by IR community. It was combined with language model for document smoothing (Mei et al. 2008, Yi & Allan 2009) (typically like LBDM (Wei & Croft 2006)) and used for query expansion (i.e., model-based feedback (Zhai & Lafferty 2001) and relevance model with Markov random fields (Lavrenko & Croft 2001, Metzler & Croft 2007)) techniques. The combination works mainly because that it takes the advantages from LDA's multiple topic representation for document modelling and relevance feedback viewing each document as topics to discover better query-specific topics. However, most of these works assume that query terms are independent given documents, which makes query expansion can not always keep good performance.

The topical n-Gram model (TNG) proposed in (Wang et al. 2007) automatically discovers term relationships within topics and extracts topically relevant and flexible phrases. Also topical PageRank can extract keyphrases in (Liu et al. 2010). But syntactically valid phrases often share low frequency in documents which cause poor performance for some queries. In (Bai et al. 2005), dependence models have been incorporated to extract term relationships for query expansion. This combination of terms are more flexible than phrases and the expanded terms are dependent to query and document. However, the risk of query drift is still a problem in query expanding area.

To solve the problem, the concept of optimization is prevalent for choosing relevant information. For example, optimization is treat as a classification task (Cao et al. 2008) that discriminates relevant from irrelevant expansion terms depending on whether they improve the performance. The approach proposed in (Maxwell & Croft 2013) focus on in-query terms selection by defining informative words and incorporating global statistics and local syntactic phrase to improve the performance. Optimised smoothing (Mei et al. 2008) technique is a general unified optimization framework for smoothing language models on graph structures. Collins-Thompson (Collins-Thompson 2009) defines a uncertainty set and models constraints to minimise the optimal loss over this set. Our approach partially inherits the idea of "mitigate risk-reward tradeoff", but we additionally provide more concrete and meaningful categories to estimate the expanded query and the methods are proposed from different perspectives.

In this paper, we propose a new approach that incorporates multiple topics from topic model and association rule mining techniques, in the sense that discovers semantic meaning of topics and inferentially exploits related terms for original query. And we also present a new ranking method to systematically weight the different importances for all expanded and original queries.

3 Pattern-based Topic Representation

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 (Gao, Xu, Li & Liu 2013), 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 (also called relative support) of the itemset X is defined $\frac{supp(X)}{|\Gamma_j|}$. Topic Z_i 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 $\mathbf{X}_{Z_i}$. For all topics, we have $\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}$ and V is the total number of topics. $\mathbf{U} = \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, ..., \mathbf{X}_{Z_V}\}$ is the pattern-based topic model generated from the given collection of documents. 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: 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

4 Topical Query Expansion (TQE)

The representations generated by the pattern-based LDA model, discussed in Section 3, carry more concrete and identifiable meaning than the word-based representations generated using the original LDA model. Based on the pattern-based topic model with layered structure and interpretable representation, we will present the method of expanding queries by our proposed novel Topical Query Expansion (TQE) model for IR. The details are described in the following subsections.

4.1 Related Topics Selection and Query Expansion

Given a collection of documents D, V pre-specified latent topics can be generated and represented by patterns according to the approach described in Section 3. 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, $\mathbf{U} = \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$, 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 the transactional dataset Γ_i, where m_i is the total number of patterns in topic Z_i and each X_{ij} in $\mathbf{X}_{Z_i}$ is a unique pattern with corresponding weight f_{ij}. This pattern-based representation enhances the semantic meaning of topics, which can also be useful for selecting right topics that a query may involve.

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 (Bayardo Jr 1998) and closed patterns. The number of these concise patterns is significantly smaller than the number of frequent patterns for a dataset. Based on the results on information filtering (Gao, Xu & Li 2013, Gao et al. 2014), the topics that are discovered from the collection of documents can be better represented by closed patterns.

For a transactional dataset, an itemset X is a *closed itemset* if there exists no itemset X' such that (1) $X \subset X'$, (2) $supp(X) = supp(X')$. A closed pattern reveals 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.

Association rule mining is to find associations between itemsets, called *association rules*. An *association rule* is an implication in the form of $X \Rightarrow Y$, where X and Y are disjoint itemsets, i.e., $X \cap Y = \emptyset$. The strength of an association rule can be measured in terms of its *support* and *confidence*. *Support* determines how often a rule is applicable to a given dataset, while *confidence* determines how interesting or strong the correlations of this rule. X and Y are itemsets, X is called antecedent and Y is called consequent, the *rule* means that X implies Y. The *relative support* of a rule is the percentage of transactions that contain X and Y, the *confidence* of a rule is the ratio between the support of the rule and the support of X.

Confidence, on the other hand, measures the reliability of the inference made by a rule. For a given $X \Rightarrow Y$, the higher the confidence, the more likely it is for Y to be present in transactions that contain X. The association rule suggests co-occurrence relationship between items in the antecedent and consequent of the rule.

For the pattern-based topic model described in

Section 3, from each transactional dataset 'Γ_i' for topic i, we can generate a set of association rules which satisfy the predefined minimum support σ and confidence η from a given transactional dataset, and denoted as $\mathbb{R}_i$. Based on the discovered rules from pattern-based topic $\{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$, for a given query $Q = \{q_1, q_2, \cdots, q_n\}$, where q_w is one of the terms in the query Q, $w = 1, 2, \cdots, n$, we can discover which topics that the query is related to, as well as the expanded terms of original queries.

The rational behind using pattern based topic models to expand queries is topical patterns contain more strong relations among terms and these associations create a reliable links between original query terms and their related topics from which the best terms can be determined to expand the query. The detailed process is described as follows.

As mentioned before, the pattern-based topic representation is $\mathbf{X}_{Z_i} = \{X_{i1}, X_{i2}, \cdots, X_{im_i}\}$ for topic Z_i in which the pattern $X_{ij} = \left\{x_{ij}^1, x_{ij}^2, \cdots, x_{ij}^{l_{ij}}\right\}$ is a set of terms, l_{ij} is the length of this pattern X_{ij}, $supp(X_{ij})$ is the support of X_{ij}. If there is a term $x_{ij}^k \in X_{ij}, k \in \{1, \cdots, l_{ij}\}$, $q_w = x_{ij}^k$, i.e., $q_w \in X_{ij}$, and $q_w \Rightarrow X_{ij} \backslash q_w$ is a rule in $\mathbb{R}_i$, topic Z_i is considered as a related topic of q_w. The pattern X_{ij} is called a topic-related patten of q_w. The set of related topics of the query word q_w, denoted as RT_{q_w} can be defined below:

$$RT_{q_w} = \{Z_i | \exists (q_w \Rightarrow X_{ij} \backslash q_w) \in \mathbb{R}_i, q_w \in X_{ij}\} \quad (2)$$

The set of related topics for a query Q is defined as:

$$RT_Q = \bigcup_{w=1}^{n} RT_{q_w} \quad (3)$$

For a query word q_w, there could be multiple topic-related patterns in X_{Z_i} that make the topic Z_i a related topic of q_w. Let $X_{q_w}^i$ be a set of topic-related patterns in X_{Z_i} for q_w, $X_{q_w}^i$ can be used to represent topic Z_i in terms of q_w. $X_{q_w}^i$ is defined below:

$$X_{q_w}^i = \{X | X \in X_{Z_i}, \exists (q_w \Rightarrow X \backslash q_w) \in \mathbb{R}_i, q_w \in X\} \quad (4)$$

For each pattern $X \in X_{q_w}^i$, the expanded query is $X \backslash q_w$ and the relevance of X to q_w with respect to the topic Z_i is defined as:

$$f_{q_w}^i(X) = \frac{supp(X)}{\sum_{Y \in X_{q_w}^i} supp(Y)} \quad (5)$$

The relevance $f_{q_w}^i$ will be used to determine the relevance of a document to a query in the retrieval stage which will be discussed in Section 4.3.

A simple example is given in Figure 1 where we describe how to find related topics for word "trade" in query 55. The minimum confidence in our experiment $\eta = 0.25$. The related patterns in topic 59 are "trade stock", "trade market" and "trade exchange" in Figure 1. The relative support of the pattern "trade stock" is 0.44, and the relative support of "trade" is 0.59 in topic 59, and $0.44/0.59 = 0.75$ is the confidence of the rule "trade⇒stock" which is larger than η. As a result, topic 59 is one of the related topics for the keyword "trade". Similarly, Z_{28} is another related topic and thus $RT_q = \{Z_{59}, Z_{28}\}$ is the set of related topics of keyword "trade".

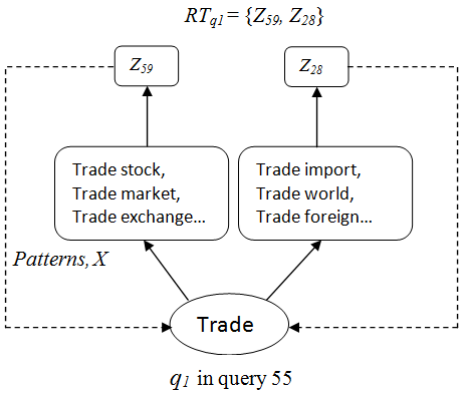

Figure 1: An example of finding related topics given a keyword "trade" from query 55

As results, the corresponding expanded queries contain "stock, market, exchange, import, world, foreign". But among all the related topics, the association degree with one word in the query is different. Some related topics are highly related to the user's interests while some other discovered related topics are actually rarely associated. So we need topic optimization process to differentiate the various importance of all the related topics to the user's real interests.

4.2 Topic Optimization

In traditional topic models, topics are represented by single words. One word could appear in different topics since words suffer from the polysemy problem. In our pattern-based topic model, the topic representation is more discriminative by patterns. Topical patterns can capture more stable and meaningful associations between words. However, different combination with the same word can often represent different topics. For example, "south" in "south Africa" is country name but it in "south west" is a direction. As a result, not all the related topics generated using Equation (3) can 'truly' represent the user's interests. Topic optimization is such a process that helps to choose those related topics which are more closed to the user's interests. In this section, we will optimise the selected related topics, RT_Q, for the query, and define a level of certainty for these related topics.

A topic Z_i is considered as the user's certain topic if it meets the following conditions:

- Z_i is a related topic of at least keyword in Q, i.e., $\exists q_k \in Q, Z_i \in RT_{q_k}$;
- Z_i is a common related topic of at least two different keywords in Q, i.e., $\exists q_h \in Q, Z_i \in RT_{q_h}$

Formally, the *certain topics* of the query Q , denoted as T_Q^c, is defined by the equation (6).

$$T_Q^c = \left\{Z_i | Z_i \in RT_{q_k} \bigcap RT_{q_h}, \exists k, h \in \{1, \cdots, n\}, k \neq h\right\} \quad (6)$$

The set of *certain topics* can be considered as the closest topics to the user's interest because they are related to at least two query words in the user' query. This feature is very important because two or more words actually form a pattern. The pattern consisting of query words can be considered a 'user-specific pattern'. It is because of the 'user-specific pattern' that the topics in T_Q^p are more certain to represent the user's interest.

The other related topics other than the certain topics in T_Q^p are considered as a set of *uncertain topics*, T_Q^u:

$$T_Q^u = RT_Q \backslash T_Q^p \tag{7}$$

A related topic in the set of *uncertain topics* contains only one original query keyword and it only satisfies the first condition.

Let $RT_{q_w}^p$ be the set of certain topics of q_w and $RT_{q_w}^u$ be the set of uncertain topics of q_w:

$$RT_{q_w}^p = \{Z | Z \in RT_{q_w}, Z \in T_Q^p\} \tag{8}$$

$$RT_{q_w}^u = \{Z | Z \in RT_{q_w}, Z \in T_Q^u\} \tag{9}$$

4.3 Document Ranking

Original query Q can be expanded by patterns from the representation of related topics RT_Q based on their property (i.e., certain topic or uncertain topic). Expanded patterns of related topic (in section 4.1) which belongs to the certain topics are called certain expansion, and patterns of related topic which belongs to uncertain topics are uncertain expansion. For different expansions, we assign different weights to indicate their importance using a trade-off parameter $\lambda \in [0, 1]$. A constant $\lambda_s \in [0, 1]$ controls the relative proportion between original query and the extended terms, and a constant $\lambda_p \in [0, 1]$ controls the weighting trade-off between extended terms in T_Q^c and the ones in T_Q^u. Also, the expanded terms have their topical frequencies that can represent the their specificities in describing the meaning of the related topic, which is calculated by Equation (5). The higher relevance of the pattern indicates the more specific and important of this pattern in this topic.

For example, query 58 is "rail striker", the related topic for "rail" is topic66 in which the expanded patterns are "transport rail" with relative support 0.05918 and "rail train" with relative support 0.05477; the related topic for "striker" is topic84 in which the expanded patters are "striker worker" with relative support 0.06155 and "striker union" with relative support 0.05352. According to Equation (5), the relevance of the expanded patterns will be "transport rail"(0.5194) and "rail train"(0.4806) in topic66, "striker worker"(0.5349) and "striker union"(0.4651) in topic84. However, queries are always related to multiple topics. The more related topics a query has, the more diverse the query is, thus more uncertain for this query. The number of related topics in RT_{q_w}, is defined as word *diversity* of q_w, denoted as $div(q_w)$. For the set of certain topics, $div_{q_w}^c = |RT_{q_w}^c|$ and for the set of uncertain topic, $div_{q_w}^u = |RT_{q_w}^u|$. If a word has high diversity, it will not be discriminative in delivering the user interest therefore the importance weight should be lower than the word with low diversity.

The principles of document relevance ranking are: 1) increase weights of certain related topics to user's interests; 2) balance the weights of uncertain related topics to user's interests. The expanded terms in certain topics are assigned higher weight compared with expanded terms in uncertain topics in which the proportion is controlled by λ_f. At more specific level, diversity (div)of the keyword is used to balance every possible meaning of it; the relevance of the pattern $f(X)$ indicates its importance in one topic. $\#div \times f(X)\#$ together presents the specificity of the expanded terms.

Table 4: Important Notations

Q	the original query
T_Q^c	set of certain topics for the query Q
T_Q^u	set of uncertain topics for the query Q
$RT_{q_w}^p$	set of certain topics of q_w
$RT_{q_w}^u$	set of uncertain topics of q_w
$X_{q_w}^i$	set of topic-related patterns in X_{Z_i} for q_w
$f_{q_w}^i(X)$	the relevance of pattern X to q_w with respect to the topic Z_i
λ_p, λ_s	trade-off parameters
$bm(q)$	the BM25 score of term q

One unique characteristic for topical pattern expansion is that the expanded terms particularly for an original query term are derived from one pattern, and they only make sense when the original query exists in a document since the original keyword and expanded terms together represent a completely meaningful pattern. Another reason is the original query is the antecedent of a rule which is a dominant element. Therefore, our proposed document ranking requires co-occurrence of the original query and its expanded terms in a document d. The ranking score of document d will linearly combine all these elements introduced above. The document ranking is calculated by Equation (10), the relevant notations in the equations are given in Table 4.

$$\begin{aligned} score(d|Q) = \sum_{\substack{q_w \in Q \\ q_w \in d}} \{\lambda_s bm(q_w) + (1-\lambda_s)\{ \\ \frac{\lambda_p}{|RT^c(q_w)|} \sum_{Z_i \in T_Q^c} \sum_{X \in X_{q_w}^i} f_{q_w}^i(X) bm(X \backslash q_w) + \\ \frac{1-\lambda_p}{|RT^u(q_w)|} \sum_{Z_i \in T_Q^u} \sum_{X \in X_{q_w}^i} f_{q_w}^i(X) bm(X \backslash q_w))\}\} \end{aligned} \tag{10}$$

4.4 Algorithm

To understand this process clearly, we formally describe the process in two algorithms: Offline training (i.e., generating pattern-based topic model for the collection) Algorithm and Online Retrieval (i.e., expanding queries and document relevance ranking) Algorithm. The former generates pattern-based topic representations and paves the way for the latter expansion system. Given a query online, the latter TQE model expands the query with related topics and computes the document ranking with topic-related patterns.

5 Evaluation

In order to emphasize the effectiveness of our proposed weighted query expansion model, in the experiments, we only use query itself as the input without user interactive information such as relevance feedbacks. The hypothesis to be verified in the experiments to be discussed in the following sections is that the proposed query expansion based on the related topical patterns generated from the collection is effective. This section discusses the experiments and evaluation in terms of data collection, baseline models, measures and results. The results show that the

Algorithm 1 *Offline Training*

Input: a collection of training documents D; minimum support σ_j as threshold for topic Z_j; number of topics V
Output: pattern-based topic representations for the collection, $\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}$

```
Generate topic representation φ and word-topic assignment z_{d,i} by applying LDA to D
for each topic Z_i ∈ [Z_1, Z_V] do
    Construct transactional dataset Γ_i based on φ and z_{d,i}
    Construct user interest model X_{Z_i} for topic Z_i using a pattern mining technique so that for each pattern X in X_{Z_i}, supp(X) > σ_i
end for
```

Algorithm 2 *Online Retrieval*

Input: topic representations $\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}$; a query $Q = \{q_1, q_2, \cdots, q_n\}$; a list of documents D'; a minimum confidence η;
Output: $score(d|Q), d \in D'$

```
for each q_w ∈ Q do
    RT_{q_w} := ∅
    for each topic Z_i ∈ [Z_1, Z_V] do
        X^i_{q_w} := ∅
        for each pattern X_{ij} ∈ X_{Z_i} do
            Scan patterns and find q_w = x_{ij}, x_{ij} ∈ X_{ij}
            if q_w ⇒ X_{ij}\q_w ∈ R_i then
                RT_{q_w} = Z_i ∪ RT_{q_w}
                X^i_{q_w} = X_{ij}\q_w ∪ X^i_{q_w}
            end if
        end for
    end for
end for
for all q_w ∈ Q do
    Find T^c_Q by equation (6)
end for
for each d ∈ D' do
    score(Q|d) := 0
    score(Q|d) is calculated by equation (10)
end for
```

proposed TQE model significantly outperforms the baseline models in terms of electiveness.

5.1 Data

The benchmark datasets from Text REtrieval Conference (TREC)[1] (Voorhees et al. 2005) are used to evaluate our proposed model for IR. For each dataset, the queries are taken from the "title" field of TREC topics. A collection of text documents and relevance judgements for each query are involved in each dataset. Table 5 shows the datasets (AP, SJMN, WSJ) details.

5.2 Measures

The effectiveness is assessed by five different measures: average precision of the top K ($K = 5$, and $K = 10$) documents, $F_\beta(\beta = 1)$ measure, Mean Average Precision (MAP). 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

[1] http://trec.nist.gov/

Table 5: The Statistics of TREC corpora and topics. The number of documents D is given in thousands

Collection	Abbrev	$\#D$	TREC topics
Associated Press	AP	243	51-150
San Jose Mercury News	SJMN	90	51-150
Wall Street Journal	WSJ	173	51-100 151-200

top5, *top10*, MAP or F_1 score, the better the system performs.

5.3 Settings

Firstly, we apply the LDA model to construct topic models with $V = 100$ latent topics for SJMN, 200 topics for WSJ and 300 topics for AP according to the size of each data collection, using the MALLET topic modelling toolkit[2]. Our experiments show that insufficient number of topics will largely bring in abundant expanded patterns into the topic model. We run collapsed Gibbs inference for 1000 samplings, the hyper-parameters of the LDA model are $\alpha = 50/V$ and $\beta = 0.01$. We also pre-process all documents by removing standard stopping words, removing numbers and punctuation.

In the process of generating pattern-based topic representations, the minimum support σ for every topic in each collection is set to 0.05.

For selecting the related topics, the minimum confidence of a topical association rule is set to $\eta = 0.25$.

The trade-off parameters in document relevance ranking are set as $\lambda_p = 0.9$, $\lambda_s = 0.7$ in experiments for all the three datasets.

5.4 Baseline IR models

Three existing IR models are chosen as baseline models in the experiments, including one term-based ranking model BM25 and two state-of-the-art topic-based IR models, which integrate topic model with language model and have achieved relatively successful performance.

5.4.1 BM25

BM25 (Robertson et al. 2004) is one of the state-of-the-art term-based document ranking approaches. In this paper, the original and expanded queries are all scored by BM25 weights. The term weights are estimated using the following equation:

$$W(t) = \frac{tf \times (k+1)}{k_1 \times ((1-b) + b\frac{DL}{AVDL}) + tf} \times \log(\delta + \frac{N - n + 0.5}{n + 0.5}) \quad (11)$$

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_1 and b are the parameters, which are set as 1.2 and 0.75 in this paper. Notice that, in this paper, we use the modified BM25 in the part

[2] http://mallet.cs.umass.edu/

of inverse document frequency (idf). The constant δ is added to avoid the negative value of idf for the common terms in the collection, we set $\delta = 1$.

5.4.2 Topical n-gram model

The topical n-Gram model (TNG) proposed in (Wang et al. 2007) automatically and simultaneously discovers topics and extracts topically relevant phrases. It has been seamlessly integrated into the language modelling based IR task (Wang et al. 2007). The generative process can be described as follow:
1) draw discrete ϕ_z from Dirichlet β for each topic z;
2) draw discrete θ_d from Dirichlet α;
3) the difference of TNG from normal LDA model is to draw Bernoulli φ_{zw} from Beta γ for each topic z and each word w; and
4) draw discrete σ_{zw} from Dirichlet δ for each topic z and each word w;
Bernoulli chooses the assignment of the word $w_{d,i}$ to a topic ϕ_j, which is used to determine whether nearby content can be composed as phrases. Readers can refer to (Wang et al. 2007) for more details.

5.4.3 LDA-based Document Model (LBDM)

LDA-based document model smoothing technique (Wei & Croft 2006) utilizes Dirichlet smoothing to smooth $P_{ML}(w|d)$ with $P(w|coll)$, then further smooth the result with $P_{LDA}(w|d)$:

$$P(w|d) = \lambda(\frac{N_d}{N_d+\mu}P_{ML}(w|d) + (1-\frac{N_d}{N_d+\mu}) \\ P_{ML}(w|coll)) + (1-\lambda)P_{LDA}(w|d) \quad (12)$$

where $P(w|d)$ is the maximum likelihood estimate of word w in the document d, and $P(w|coll)$ is the maximum likelihood estimate of the word w in the whole collection. μ is the Dirichlet prior, which is set to 1000 , and $\lambda = 0.7$ in the experiment.

5.5 Results

Table 7: Comparison of the TQE model with TNG and LDBM models. The evaluation measure is average precision. *impr* indicates the percentage of improvement of TQE over best performance of TNG and LBDM

Collection	TNG	LDBM	TQE	*impr*
AP	0.2423	0.2651	0.3390	27.8%
SJMN	0.2122	0.2307	0.2375	3.0%
WSJ	0.2958	0.3253	0.3502	7.7%

We can see that TQE model outperforms the other two baseline models on MAP value, achieves dramatical increase for AP collection which is 27.8% and minimum increase of 3.0% for SJMN. The significant improvement is solid evidence that supports the hypothesis.

From Table 6, we can see that the number of net queries in AP, SJMN and WSJ are 99, 94 and 100, respectively. Among the valid queries, 80.8% (80/99) queries in AP, 90.4%(85/94) queries in SJMN and 83.0%(83/100) in WSJ can be expanded by our proposed approach. This demonstrates that topical patterns that are discovered by our model are the interpretable representations and also effective at selecting related topics which can be further used in query expansion. This results also support the hypothesis.

In the same table, we can find that in the expanded queries across the three collections, 23.5%-37.5% queries get improved under evaluation measure *top*5 ,25.9%-42.4% queries performs better under evaluation *top*10, 53.0%-66.3% queries under MAP and 51.8%-66.3% under F1 get improved. Although the improving number of queries under *top*5 is smallest but its average gain over all improved queries are the highest, while MAP and F1 evaluation have large number of improved queries but the improvement value is relatively the lowest. To summarise the performance, our proposed TQE consistently performs excellently across all evaluation measures, which proves the hypothesis that it is effective at query expansion.

6 Discussion

The results on benchmark TREC datasets show that this technique can result in major improvements for a large proportion of queries. The reason behind is that using topical patterns are to expand the query can obtain more accurate and semantically related terms. The significant improvements are also contributed by the divisions of focused queries and scattered queries at a topic level and different ranking mechanisms based on the different queries for ranking documents.

6.1 Topical Patterns for Query Expansion

Instead of using individual words as representation of topics, we use stronger terms relationships by topical patterns to select related topics as well as expanded terms from the related patterns. Pattern-based representations are effective at interpreting the semantic meaning, thus the topical pattern based expansion is more trustful in terms of semantics . At ranking stage, the topical patterns are also effective because they can deliver specific weightings with patterns (X) and their relevance ($f^i_{q_w}(X)$).

But based on our experimental experience, if the number of trained topics is too small (i.e., $V = 100$ for the larger collection AP) which can't fit the topic partitions within a collection, the performance will drop sometimes even worse than original BM25. The main reason is that very low dimension of topics leads to abundant patterns, which causes the related patterns contain many noises. Therefore, the number of topics directly affects the correctness of query expansion by topical patterns.

The query expansion approach TQE is quite flexible indeed. If the query itself is short, the related topics will complement the features; and if they query is verbose, the optimization of related topics will converge to focused topics and decrease the effects of noisy terms. Consequently, topical pattern for query expansion is a good strategy for IR and it also solve the mentioned problems.

6.2 Complexity

Many topical IR models require feedbacks which is less efficient in retrieving documents after user's inputs. Since our proposed model only need offline training once which requires no extra online processing time, it performs significantly outstanding on retrieval effectiveness and efficiency.

As discussed in Section 4.4, there are two algorithms in the proposed model, i.e. offline training and Online retrieval. The complexity of the online retrieival is related to the number of terms in a query

Table 6: Improvements of using TQE model compared with only using original queries. "Expanded Queries" column indicates number of the queries can be expanded . For each evaluation, "impr." shows the number of queries get improved, and "avg gain" means the average improvement over the improved queries.

Collection	Net Queries	Expanded Queries	*top*5		*top*10		MAP		F1	
			impr.	avg gain	impr.	avg gain	impr.	avg gain	impr.	avg gain
AP	99	80	30	0.238	36	0.196	53	0.099	53	0.070
SJMN	94	85	20	0.217	22	0.10	50	0.031	48	0.037
WSJ	100	83	23	0.236	28	0.123	44	0.072	43	0.044

and the expanded terms. However, the number is always small and the calculation time is often acceptable.

For offline training, 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. $O(V * N)$ (Wei & Croft 2006). 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 (Han et al. 2007) for generating frequent patterns has been widely accepted in the field of data mining and text mining. The transactional datasets used in the TQE model 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 TQE model combines the topic modelling and pattern mining linearly. Thus, in summary, the complexity of the TQE model can be determined by topic modelling or pattern mining. In most cases, the complexity of the TQE model would be the same as pattern mining since, in general, the complexity of pattern mining is greater than that of topic modelling. As them name indicates, offline training can be conducted off-line which means that the complexity of the offline training part will not affect the efficiency of the proposed IR model.

7 Conclusion and Future Work

This paper presents an innovative weighted query expansion for information retrieval including topical-pattern based query expansion and document relevance ranking. The TQE generates pattern-based topic representations for every topic in a collection. With these semantic topical patterns, a query belongs to certain topics or uncertain topics can be expanded with topic-related patterns. For the document ranking, the TQE estimates the relevance from the aspects of determining certainty of topics at general level, additionally analysing related patterns with more specific features. The proposed model has been evaluated by using the TREC collections for IR task. Compared with the state-of-the-art models, the TQE demonstrates excellent strengths both on query expansion and document relevance ranking.

The proposed model automatically generates discriminative and semantic rich representations for modelling topics and queries by combining topic modelling techniques and data mining techniques. The following work in future could take relevance feedback into consideration, the related topics and patterns discovered by the TQE will be assigned with more precise weights.

References

Andrzejewski, D. & Buttler, D. (2011), Latent topic feedback for information retrieval, *in* 'Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 600–608.

Bai, J., Song, D., Bruza, P., Nie, J.-Y. & Cao, G. (2005), Query expansion using term relationships in language models for information retrieval, *in* 'Proceedings of the 14th ACM International Conference on Information and Knowledge Management', CIKM '05, ACM, New York, USA, pp. 688–695.
URL: *http://doi.acm.org/10.1145/1099554.1099725*

Bayardo Jr, R. J. (1998), Efficiently mining long patterns from databases, *in* 'ACM Sigmod Record', Vol. 27, ACM, pp. 85–93.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *the Journal of machine Learning research* **3**, 993–1022.

Cao, G., Nie, J.-Y., Gao, J. & Robertson, S. (2008), Selecting good expansion terms for pseudo-relevance feedback, *in* 'Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 243–250.

Collins-Thompson, K. (2009), Reducing the risk of query expansion via robust constrained optimization, *in* 'Proceedings of the 18th ACM conference on Information and knowledge management', ACM, pp. 837–846.

Gao, Y., Xu, Y. & Li, Y. (2013), Pattern-based topic models for information filtering, *in* 'Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on', pp. 921–928.

Gao, Y., Xu, Y. & Li, Y. (2014), A topic based document relevance ranking model, *in* 'Proceedings of the companion publication of the 23rd international conference on World wide web companion', International World Wide Web Conferences Steering Committee, pp. 271–272.

Gao, Y., Xu, Y., Li, Y. & Liu, B. (2013), A two-stage approach for generating topic models, *in* 'Advances in Knowledge Discovery and Data Mining', Springer, pp. 221–232.

Han, J., Cheng, H., Xin, D. & Yan, X. (2007), 'Frequent pattern mining: current status and future directions', *Data Mining and Knowledge Discovery* **15**(1), 55–86.

Kumaran, G. & Carvalho, V. R. (2009), Reducing long queries using query quality predictors, *in* 'Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 564–571.

Lavrenko, V. & Croft, W. B. (2001), Relevance based language models, *in* 'Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 120–127.

Liu, Z., Huang, W., Zheng, Y. & Sun, M. (2010), Automatic keyphrase extraction via topic decomposition, *in* 'Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, pp. 366–376.

Lv, Y. & Zhai, C. (2011), Lower-bounding term frequency normalization, *in* 'Proceedings of the 20th ACM international conference on Information and knowledge management', ACM, pp. 7–16.

Maxwell, K. T. & Croft, W. B. (2013), Compact query term selection using topically related text, *in* 'Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 583–592.

Mei, Q., Zhang, D. & Zhai, C. (2008), A general optimization framework for smoothing language models on graph structures, *in* 'Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 611–618.

Metzler, D. & Croft, W. B. (2007), Latent concept expansion using markov random fields, *in* 'Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 311–318.

Robertson, S., Zaragoza, H. & Taylor, M. (2004), Simple bm25 extension to multiple weighted fields, *in* 'Proceedings of the thirteenth ACM international conference on Information and knowledge management', ACM, pp. 42–49.

Steyvers, M. & Griffiths, T. (2007), 'Probabilistic topic models', *Handbook of latent semantic analysis* **427**(7), 424–440.

Voorhees, E. M., Harman, D. K. et al. (2005), *TREC: Experiment and evaluation in information retrieval*, Vol. 63, MIT press Cambridge.

Wang, X., McCallum, A. & Wei, X. (2007), Topical n-grams: Phrase and topic discovery, with an application to information retrieval, *in* 'Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on', IEEE, pp. 697–702.

Wei, X. & Croft, W. B. (2006), 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, pp. 178–185.

Yi, X. & Allan, J. (2009), A comparative study of utilizing topic models for information retrieval, *in* 'Advances in Information Retrieval', Springer, pp. 29–41.

Zhai, C. & Lafferty, J. (2001), Model-based feedback in the language modeling approach to information retrieval, *in* 'Proceedings of the tenth international conference on Information and knowledge management', ACM, pp. 403–410.