Query Expansion Based on Formal Concept Analysis From Retrieved Documents

Haibin Yu[1], Chongyang Shi[1], Yu Bai[2], Chunxia Zhang[1], Ryan Hearne[1]

[1] School of Computer Science, Beijing Institute of Technology, China
[2] QCIS, University of Technology, Australia
{yuhaibin, cy_shi}@bit.edu.cn, yu.bai-3@student.uts.edu.au, cxzhang@bit.edu.cn, itsryanhearne@gmail.com

Abstract

In this paper, we propose a new formal concept analysis (FCA)-based query expansion approach, which uses the set of retrieved document collection against the whole document set. In this approach, description topics (DTs) are extracted from the documents and organized to denote precisely the user's information need. For a new query, we build a concept lattice from the extracted DTs, using the retrieved document collection as the formal context, and choose the most probable interpretations as query concepts. Our experiments are performed on two collections (data sets from TREC-7, TREC-8 and AP89). The experimental evaluation shows that our approach can reduce the overall computational overhead, and is as good as some typical query expansion approaches.

Keywords: Query expansion, Formal concept analysis, Information retrieval, Query concepts

1 Introduction

Until now, knowing how to express a user's informational requirement in a query was a continually occurring question within the information retrieval (IR) field [1]. One of the pivotal problems is how to determine the set of words or terms that can express and expand the query semantically, except for refining the search query [2]. In some previous IR systems, after several general text preprocessing steps, a document in a data set is always represented by the terms or words that appear in it, and, usually, the important terms can be chosen to represent the user's informational need, which always results in the well-known vocabulary mismatch problem.

Consequently, traditional keyword-based retrieval is becoming more difficult to satisfy users' search demands, and so concept-based retrieval approaches have been developed. Concept-based approaches treat query words as concepts, rather than as literal strings of letters, and take into account some domain knowledge in determining an appropriate expression and expansion of the initial query [3-7]. Also, in some traditional research [8], particular related terms pre-extracted from a whole set of documents, regarded as query concepts, are selected to expand the user's information needs. But the construction of query concepts, which deals with the whole corpus, is very time consuming. However, these means are useful in that they get the experiences in bridging the gap between the terminology used in defining initial queries and the terminology used in representing documents. Therefore, the concept-based retrieval approaches are able to retrieve relevant documents even if they do not contain the specific words used in the initial query.

Wille proposed FCA in 1982 [9], and, since the 90s, it has been integrated with basic information retrieval techniques to build more comprehensive systems for the information-accessing field. Concept lattices were used as a support structure for IR. A number of researchers have proposed lattice-based structures for IR [10-15], but most of which only made researches or practices using a variety of small datasets showing the mathematical aspect of FCA. The search applications using FCA aiming at middling or more amounts of documents could not be found. That's because when having large number documents and attributes(always using the terms extracted from the documents), the lattice's building is time and space consuming.

In order to provide FCA a path to deal with more amounts of documents, minimize the scope of query in the large numbers of documents, and provide controlled ways of eliminating query expansion diversity, we develop a new FCA-based query expansion approach, as follows:

(1) Based on the query, which gets the set of retrieved documents against the whole document set to do the expansion.

(2) Description topics (DTs) defined as intrinsic concepts existing in a document are extracted from the retrieved documents.

(3) Using the retrieved documents as objects and DTs as attributes, a concept lattice is bulit as the

*Corresponding Author: Chongyang Shi; E-mail: cy_shi@bit.edu.cn
DOI: 10.3966/160792642019032002010

possible expansion space.

(4) Generating the expanded query by selected lattice node(s)

In our work, rather than using a thesaurus (e.g. WordNet [16]), we use the set of retrieved documents sets which is a kind of local technique consuming less time and space compared to global technique. Furthermore, to improve the expansion performance, we use the DTs to capture the latent semantic meanings in the documents instead of simply extracted terms, based on which the concept lattice is used to organize the documents and DTs. Finally, we get the query concepts from the concept lattice, which could denote the user's information need appropriately, not using traditional terms.

Therefore, our research focused on two issues: (1) how to get the DTs from the set of retrieved documents sets by a searching query, and (2) how to calculate the select the query concepts from the concept lattice of DTs. Moreover, we proposed an automatic query expansion framework, called the FCA-based query expansion model, to construct appropriate query concepts. Finally, we examined whether our framework is capable of using the retrieved set of documents to construct expanded query concepts.

This article is organized as follows. Section 2 summarizes related works of query expansion. Section 3 describes the query expansion measure based on formal concept analysis. Section 4 shows experimental results. Section 5 concludes the article.

2 Related Works of Query Expansion

Query expansion is a method for improving retrieval effectiveness. It is based on the assumption that the user's initial queries often do not entirely satisfy their information needs. Query expansion techniques can then be applied to modify the original user's queries, thereby aiming to improve the retrieval results. In this section, we will introduce the global analysis, local analysis and ontology-based query expansion.

2.1 Query Expansion by Global Analysis and Local Analysis

There are two kinds of method for query expansion techniques: global analysis (or corpus specific query expansion) and local analysis (or query specific query expansion) [17-20]. In the global analysis method, new terms are added to an original query before searching, which requires the use of external resources such as a thesaurus [21], or WordNet [22-23]. In the local analysis method, a new query is formulated on the basis of some retrieved documents of search with the original query [24].

2.1.1 Global Query Expansion Techniques

Global query expansion uses the knowledge embedded in the corpus, such as the lexical association or co-occurrence information, to determine useful terms to add to the user query. Over the years, many techniques have been developed by researchers to use this extracted knowledge.

The query expansion model, proposed in [25], gives a novel framework for query expansion, which generates a set of expanded queries that provides a classification of the original query result set. Specifically, the expanded queries maximally retrieve the results of the original query, and the results retrieved by different expanded queries are different. Shiri [21] focused on end-user query-expansion behavior within the context of a thesaurus-enhanced search setting. The results indicated that thesauri are capable of assisting end users in the selection of search terms for query formulation and expansion; in particular, by providing new terms and ideas.

With the development of the language-modeling (LM) framework, query expansion techniques, such as the hidden Markov model (HMM) [26], have been developed by researchers. These techniques can be categorized as global techniques. Bai [27] proposed to integrate various contextual factors, such as the topic domain of the query, the characteristics of the document collection, and the context words, within the query to generate a new query language model.

The concept-based query expansion (CQE) [28-30] presents a mechanism to help searchers navigate their way into the semantical richness of the meaning of a query, or the resource collection. The Markov chain translation model (MCTM) [31] indicates a method based on MC models, which integrates several types of monolingual term relation in addition to the translation relation. As a result, query translation is extended to cross-lingual query expansion. The CQE and MCTM are well-known global query expansion models.

2.1.2 Local Techniques for Query Expansion

Local techniques for query expansion basically use the pseudo relevance feedback. A set of top-ranked documents retrieved for the query is used to expand the query. Local feedback [32] and local context analysis [33] are some of the well-known local techniques. Carpineto [34-35] proposed an information-theoretic approach to select and weigh the expansion terms within Rocchio's framework of query reweighing. The relevance model [36] and model-based feedback [37] are some other popular local techniques for query expansion. Metzler [38-39] extended the Markov random field model using the latent concept expansion in a pseudo relevance feedback setting. Collins [40] used co-occurrence in the top retrieved documents, along with general word association, co-occurrence in a large Web corpus, and synonyms in his random walk model for query expansion. Cao [41] used supervised learning to separate good expansion terms from others,

as obtained through pseudo relevance feedback.

In our work, we use the local technique and develop an FCA-based query expansion approach, which uses the set of retrieved documents against the whole document set.

2.2 Ontology-based Query Expansion

Query expansion with the use of ontologies has been recently well researched to help the users clarify their information needs, and come up with semantic representations of documents. Various approaches and models exist for conducting query expansion using ontological information in the last decade [42-46].

Nenad Stojanovic [42] presented an approach for calculating relevance in the ontology-based retrieval. The approach extends the notion of syntactical relevance, which can be found in traditional IR, to the semantic relevance that takes into account the conceptualization of the retrieval process. It combines standard IR techniques for representing uncertainty with the rich domain model. The main advantage is the possibility to define relevance in a flexible way, so that it can be adapted for different contexts. In [43], an ontology-based information retrieval approach is proposed based on the existence of a conceptual hierarchical structure. This approach encodes the contents of the domain to which the considered collection of documents belongs. Both documents and queries are represented as weighted trees. Dragoni et al. [44] designed a vector space model approach to representing documents and queries, which is based on concepts instead of terms and uses WordNet as a light ontology. Such representation reduces the information overlap with respect to classic semantic expansion techniques. Experiments carried out on the MuchMore benchmark and on the TREC-7 and TREC-8 ad-hoc collections demonstrate the effectiveness of the proposed approach. The approaches to semantic searches by incorporating Linked Data annotations of documents into a generalized vector space model are illustrated in [45]. One model exploits taxonomic relationships among entities in documents and queries, while the other model computes term weights based on semantic relationships within a document. Corcoglioniti et al. [46] investigated the benefits of using the semantic content automatically extracted from text, in which both queries and documents are processed to extract semantic content pertaining to the semantic layers. They are: entities, types, frames and temporal information.

Experiments undertaken in the above researches illustrate the ontology-based query expansion methods' effectiveness of the approach. In our work, similar to [44-45], we use the vector space model and concept lattice to express and organize the documents and queries instead of using traditional terms.

3 The New Query Expansion Model Based on Formal Concept Analysis

In this paper, DTs are defined as intrinsic concepts existing in a document, or a document collection, which can represent the inclusive and important meaning of a document or a document collection. Since it is possible that there may be many repeated DTs in similar documents, we can construct these topics between documents systematically into concept lattices in order to make use of them in IR.

Since the DTs represent the main content of documents, if we apply our framework to the whole corpus, it becomes more time consuming because of the size of (1) the generated possible DTs and (2) the DT lattice underlying the whole set of documents. Therefore, we use the set of retrieved documents instead of the whole collection, which is expected to be less computationally expensive in obtaining the DTs and in building the lattice. The use of the set of retrieved documents makes it possible to derive the query concepts from the initial query. This approach is encouraged by earlier works [47-48], because the reliability of the retrieved documents, as a document collection, is quite good. In these earlier works, the authors prefer to use the information extracted from the retrieved set of documents for a query expansion. However, this approach has the disadvantage that we must construct the DTs and a build lattice for each new query expansion. Figure 1 shows the process of constructing DTs graphically.

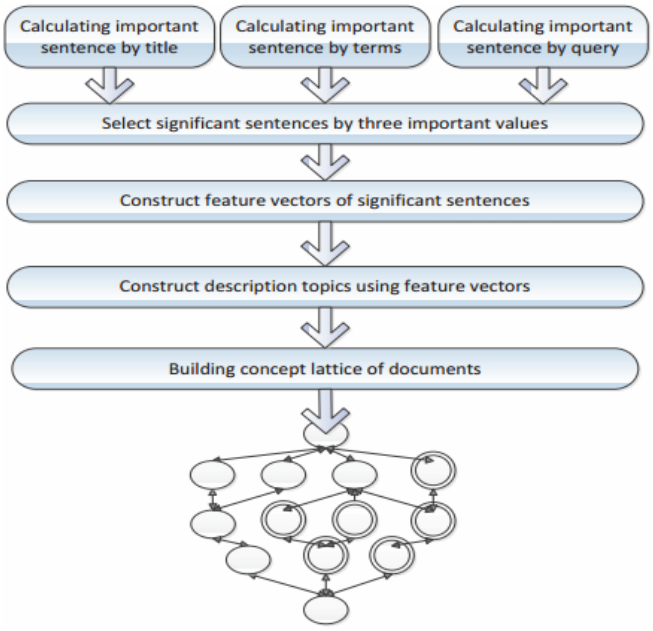


Figure 1. Construct description topics and building a concept lattice

3.1 Obtaining Important Sentences

In everyday life, we usually find that a certain sentence can be representative of a well-known book

or article. A representative sentence may easily remind people of the book containing this sentence. We adopt individual sentences as basic units for feature selection, since sentences are more representative than terms.

Since summaries are composed using important sentences in documents, the basic idea of composing a salient summary is to include the sentences that contain more salient concepts. Moreover, to compose a summary with good coverage, a summary sentence should provide some salient concepts that are different from the other summary sentences. A variety of summarization methods have been proposed over the years. Currently, most methods are extractive methods for which sentence ranking is essential. Many sentence-ranking methods have been proposed in previous studies, including feature-based methods, graph-based methods, and learning-based methods [49-53]. Therefore, we select the significant sentences of each document by use of scoring in our work.

First, we retrieve a set of documents for a given query, and then take the top-ranked n documents from the retrieved documents. The importance of each sentence is measured here by a hybrid method. We compute the importance by the title, importance of term, query, and then the importance of a sentence is calculated by combination of these three kinds of importance.

3.1.1 The Importance of Sentences by the Title

Generally, we believe that a title summarizes the important content of a document [54]. By Mock [55] terms that occur in the title have higher weights. But the effectiveness of this method depends on the quality of the title. In many cases, the titles of documents from newsgroups or emails do not properly represent the contents of these documents. Hence, we use the similarity between each sentence and the title instead of directly using the terms in the title. Similar sentences to the title contain the important terms, generally. For example, let us consider the following title: "I have a question." This title does not contain any meaning about the contents of a document. Nevertheless, sentences with a question should be handled importantly because they can have key terms about the question.

We measure the similarity between the title and each sentence, and then we assign the higher importance to the sentences with the higher similarity. The title and each sentence of a document are represented as the vectors of content words. The similarity value of them is calculated by taking the inner product, and the calculated values are normalized into values between 0 and 1 by a maximum value. The similarity value between the title T and the sentence S_i in a document d is calculated by using the following formula:

$$Sim(T, S_i) = \frac{T \bullet S_i}{max_{s_j \in d}(\vec{T} \bullet \vec{S_j})} \quad (1)$$

where $\vec{T}$ denotes a vector of the title, and $\vec{S_i}$ denotes a vector of sentence.

3.1.2 The Importance of Sentences by the Importance of Terms

Since the method by the title depends on the quality of the title, it can be useless in a document with a meaningless title or no title at all. Besides, sentences that do not contain important terms need not be handled importantly; although, they are similar to the title. On the contrary, sentences with important terms must be handled importantly; although, they are dissimilar to the title. Considering these points, we first measure the importance values of terms by TF, IDF, and statistical values. Then, the sum of the importance values of terms in each sentence is assigned to the importance value of the sentence. In this method, the importance value of a sentence S_i in a document d is calculated as follows:

$$SoIT(S_i) = \frac{\sum_{t \in S_i} tf(t) \times idf(t)}{max_{s_j \in d}(\sum_{t \in S_i} tf(t) \times idf(t))} \quad (2)$$

where f (t) denotes the term frequency of term t, and id f (t) denotes the inverted document frequency.

3.1.3 The Importance of Sentences by the Query

Since most of the query-focused summarization systems base their work on incorporating features that are related to the given query (e.g. the relevance of a sentence to the query), a generic summarization system can be easily adapted to a query-focused one. These query analyses are investigated by a number of systems to either extract key words from the query based on the word weight calculated in different ways, or analyze the question type of a query if there exists, which indicates what kind of information the query is seeking [56].

Accordingly, we measure the similarity between the query and each sentence, and then we assign the higher importance to the sentences with the higher similarity. The query, and each sentence of a document, are represented as vectors of content words. The similarity value of them is calculated by taking the inner product, and the calculated values are normalized into values between 0 and 1 by a maximum value. The similarity value between the title q and the sentence S_i in a document d is calculated by the following formula:

$$Sim(q,S_i)=\frac{q\bullet S_i}{max_{s_j\in d}(\vec{q}\bullet\overrightarrow{S_j})} \quad (3)$$

where $\vec{q}$ denotes a vector of the query, and $\overrightarrow{S_j}$ denotes a vector of sentence.

3.1.4 The Combination of Three Sentence Importance Values

Three kinds of sentence importance are simply combined in the following formula:

$$Score(S_i)=w_1\times Sim(T,S_i)+w_2\times Sim(q,S_i)+SoIT(S_i) \quad (4)$$

In (4), the w_1, w_2 and w_3 are constant weights, which control the rates of reflecting the three importance values. Also, $w_1 + w_2 + w_3 = 1$.

3.2 Important Sentences into Feature Vector

For each document, we score the sentences with the above method. The final score for each sentence is calculated by summing the individual score obtained from each method used.

Lam-Adesina and Jones also limited the optimal summary length from 15% of the original document length up to various maximum summary lengths (4, 6, or 9 sentences), empirically. Thus, we empirically set the optimal summary length, i.e. the measure of significant sentences (mss), to 7, for a medium sized document (i.e. 25 < number of sentences (NS) < 40) in our approach. For documents out of this range (NS < 25 or NS > 40), we compute the mss as follows:

$$mss = 7 + (0.1 * (NS - L))$$

Where NS is the number of sentences in a document and L is the limit (25 for NS < 25 and 40 for NS > 40). We intended to make this mss not only proportional to the document length, but also not too affected by the document length. Finally, we choose the highly ranked mss sentences as significant sentences. Then, we generate the summaries with these selected sentences for each given document [8]. If the NS is smaller than mss, we take all sentences in the document as significant sentences (mss = NS).

Then, we merge the selected important sentences that have overlapping terms among them inside a document. Therefore, we partition the selected significant sentences into several orthogonal feature vectors that do not share common terms.

Example 1. We represent a sentence v_i, as a vector of terms with their associated TF*IDF weight values in the document (e.g. see Figure 2). Thus, for each document, we can consider a set of vectors V = {v_1, v_2, v_3, v_4}. To build the feature vectors of the illustrated document d_1, we partition S as follows.

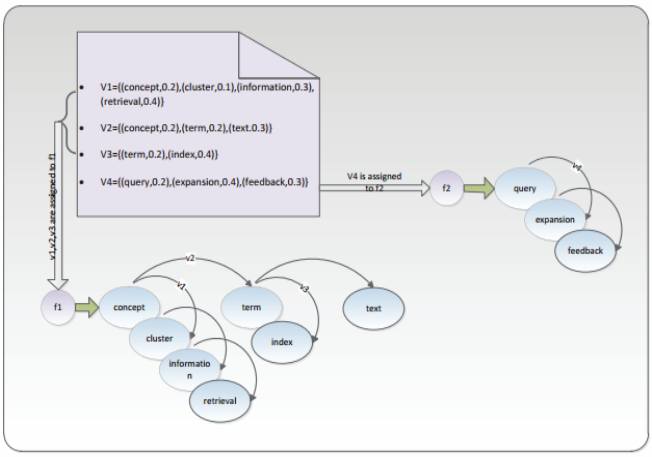


Figure 2. Merging the selected important sentences' features

First, v_1 is assigned to feature f_1. Since v_2 has a common term "concept" with v_1, v_2 is also assigned to f_1. On the other hand, v_4 is assigned to the new feature f_2 because it has no overlapping terms with the first two sentences. The sentence v_3 is merged into f_1 because of the shared term "term". This process thus results in two feature vectors that are orthogonal to each other. After the partitioning process, each feature vector, f_i, will be seen as a maximally connected component, as shown in Figure 2. We can see that the feature vectors, f_1 and f_2, have no overlapping terms in the document, i.e. they are orthogonal in a vector space. More formally, this idea of partitioning is devised by constructing maximally connected components in a graph representation. Each vector v_i is a subgraph, such that the vertices of the subgraph are the terms of the vector v_i, and the edges connect the terms, which are in the same sentence v_i.

In this process, we take only the significant terms included in the significant terms list, generated in the sentence selection process, to construct the feature vectors, and then partition these sentences.

3.3 Feature Vectors into Description Topics

Although the sentence selection allows us to find features inside a document according a query, the feature's table (formal context) is extremely sparse. To solve this question, we need a similarity measure-ng method to integrate the similar features into DTs from the retrieved collections in general, which we will use to create the DT's formal context.

Specifically, we calculate semantic similarity between features by considering the following relations in addition to standard stemming: (i) synonyms or hyponyms (such as war-battle); (ii) derivational morphological variations (such as decision-decide, Argentine-Argentina); and (iii) inflectional morphological variations (such as relations-relation). According to Lesks hypothesis [57], word senses that are related can be characterized by their shared words in their definition. In our work, the definition of a sense

of a feature consists of information from WordNet: its synset (a set of synonyms of the sense), gloss (the explanation of the sense with possibly specific examples), direct hypernyms (is-a relation) and meronyms (has-a relation). For two features, f_i and f_j, with sense definitions $FS_1^i, FS_2^i, ...FS_m^i$ and $FS_1^j, FS_2^j, ...FS_n^j$, respectively, their semantic similarity is:

$$Sim(f_i, f_j) = \max_{1 \leqslant x \leqslant m, 1 \leqslant y \leqslant n} \frac{\sum_{z=0}^{k} |Overlap^z(FS_x^i, FS_y^j)|^2}{|FS_x^i| \times |FS_y^j|} \quad (5)$$

Example 2. Consider Relation A in Figure 3. Relation A could reference the same feature. In other words, features f_1 and f_6 could denote similar meaning. Indeed, summarization methods discover only local patterns in a document.

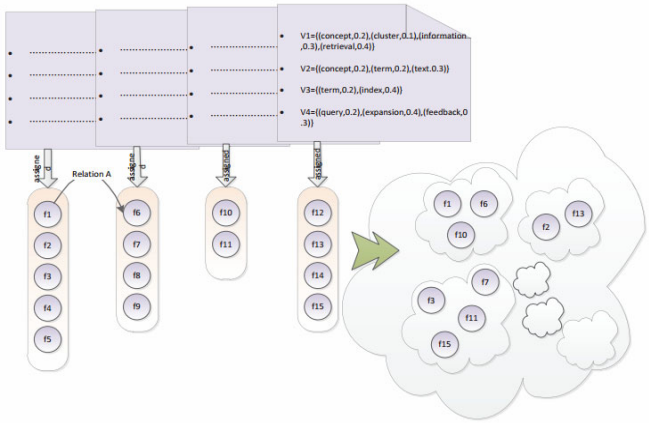


Figure 3. The procedure of getting description topics from feature vectors

Then we propose a clustering algorithm, which is suitable for our objective. This algorithm is similar to the single pass and reallocation method used in early work on cluster analysis [58]. In this algorithm, we divide the clustering procedure into two phases: (1) cluster the feature vectors ($f_1, f_2, \cdots f_n$) to centroid vectors ($c_1, c_2, \cdots c_m$) using the single pass method, and (2) reallocate the feature vectors ($f_1, f_2, \cdots f_n$) to the generated centroid vectors. In each phase, we apply the following algorithm to the feature vectors and generate the centroid vectors.

Algorithm

1. Assign the first feature, f_1, as a representative for the description topic DP_1;

2. For each feature, f_i, compare the similarity, Sim, between the feature f_i and all generated DTs. The measure of similarity is given in (5).

3. Choose the maximum value of similarity Sim called Sim_{max}.

(a) If the similarity Sim_{max} is greater than a threshold value u (e.g. 0.8), we recompute the description topic DP_j.

(b) If the similarity Sim_{max} is less than a threshold value v (e.g. 0.2), we create a new DT using feature f_i.

(c) Otherwise, if the similarity Sim_{max} is between u and v, we ignore feature f_i.

4. Repeat Steps 2 and 3 until there is no more feature f_i.

All features are assigned according to the above algorithm. We obtain m initial DTs. In the reallocation step, we follow the procedure described above apart from Step 1. We reassign all features to the generated m clusters to obtain an improved partition. This reallocation process operates to select some initial partitions of the feature vectors and then to move these features from cluster to cluster, so as to obtain an improved partition. In all of our experiments, the parameters u and v were set experimentally to 80% and 20%, respectively. Finally, these generated clusters are designated as the DTs.

Then, we try to analyze and expand the user's query using these DTs and a concept lattice.

3.4 Organize Description Topics By Formal Concept Analysis

FCA is proposed by Wille R. in 1982 [10, 59], and it is a field of applied mathematics based on the mathematization of concept and conceptual hierarchy. As a type of very effective method for data analysis, FCA has been wildly applied to various fields, such as machine learning, information retrieval, software engineering, and knowledge discovery [11, 60-63]. It is suitable for exploration of symbolic knowledge (concepts) contained in a formal context, such as a corpus, a database, or an ontology [63].

Suppose we have a collection U of documents. Individual members of this collection (documents) are written with small letters like (d, d_1, d_2), while subsets are written in capitals (D, D_1, D_2). During the indexing process, DTs (attributes) are attached to documents. We write A to denote the set of all attributes (a, a_1, a_2) for individual attributes, and (A, A_1, A_2) for attribute sets (subsets of A). The result of the indexing process is reflected in the binary relation: we write aγd iff attribute a describes document d. The tuple (U, A, R) is called a context.

Remark 1. Let U and A be two finite and nonempty sets. Elements of U are documents, and elements of A are DTs. The relationships between documents and DTs are described by a binary relation R between U and A, which is a subset of the Cartesian product U ×A. For a pair of elements x ∈ U and y ∈ A, if (x, y) ∈ R, also written as xRy, we say that document x has the DT y, or the DT y is possessed by document x. In this paper, (x, a) ∈ R is denoted by 1, and (x, a) ∉ I is denoted by 0. Thus, a document collection's formal context can be represented by a table with only entries of 0 and 1.

A document x ∈ U has the set of DTs:

$$xR = \{y \in A \mid xRy\} \subset A \quad (6)$$

A property y ∈ V is possessed by the set of documents:

$$Ry = \{x \in U \mid xRy\} \subset U \quad (7)$$

Remark 2. For a (document-DT) formal context (U, A, R), for a set X ⊆ U of documents and a set B ⊆ A of DTs, we define a set-theoretic operator *[59]:

$$X^* = \{y \in A \mid \forall x \in X, (x, y) \in R\} \quad (8)$$

It associates a subset of DTs X^* to the subset of documents X. Similarly, for any subset of DTs B ⊆ A, we can associate a subset of documents B^*

$$B^* = \{x \in U \mid \forall y \in B, (x, y) \in R\} \quad (9)$$

X^* is the set of all the DTs shared by all the documents in X, and B^* is the set of all the documents that fulfill all the DTs in B.

Remark 3. Let (U, A, R) be a document collection's formal context. The pair (X, B) is called a formal concept, for short, a concept of (U, A, R), if and only if X ⊆ U, B ⊆ A, X^* = B and B^* = A. X is called the extension and B is called the intension of (X, B). The set of all concepts in (U, A, R) is denoted by L(U, A, R).

The operator * has the following DTs: for all X_1, X_2, X ⊆ U and all B_1, B_2, B ⊆ A.

$$X_1 \subseteq X_2 \Rightarrow X_2^* \subseteq X_1^*, B_1 \subseteq B_2 \Rightarrow B_2^* \subseteq B_1^* \quad (10)$$

$$X \subseteq X^{**}, B \subseteq B^{**} \quad (11)$$

$$X \subseteq X^{***}, B \subseteq B^{***} \quad (12)$$

$$X \subseteq B^* \Leftrightarrow B \subseteq X^* \quad (13)$$

$$(X_1 \cup X_2)^* = X_1^* \cap X_2^*, (X_1 \cup B_2)^* = B_1^* \cap B_2^* \quad (14)$$

$$(X_1 \cap X_2)^* \supseteq X_1^* \cup X_2^*, (B_1 \cap B_2)^* \supseteq B_1^* \cup B_2^* \quad (15)$$

Remark 4. Let (U, A, R) be a document collection's formal context, and (X_1, B_1) and (X_2, B_2) be the concepts of the context. Then, the concepts of a formal context (U, A, R) are ordered by:

$$(X_1, B_1) \leq (X_2, B_2) \Leftrightarrow X_1 \subseteq X_2 (\Leftrightarrow B_1 \supseteq B_2) \quad (16)$$

Where (X_1, B_1) and (X_2, B_2) are concepts. (X_1, B_1) is called a subconcept of (X_2, B_2), and (X_2, B_2) is called a super-concept of (X_1, B_1). The notation (X_1, B_1) < (X_2, B_2) denotes the fact that (X_1, B_1) ≤ (X_2, B_2) and (X_1, B_1) ≠ (X_2, B_2). If (X_1, B_1) < (X_2, B_2) and there does not exist a concept (Y, C) such that (X_1, B_1) < (Y, C) < (X_2, B_2), then (X_1, B_1) is called a child-concept (immediate sub-concept) of (X_2, B_2), and (X_2, B_2) is called a parent-concept (immediate super-concept) of (X_1, B_1). This is denoted by (X_1, B_1) < (X_2, B_2).

Remark 5. Let (U, A, R) be a document collection's formal context, then L(U, A, R) is a complete lattice. The infimum and supremum are given by:

$$(X_1, B_1) \wedge (X_2, B_2) = (X_1 \cap X_2, (B_1 \cup B_2)^{**}) \quad (17)$$

$$(X_1, B_1) \vee (X_2, B_2) = ((X_1 \cup X_2)^{**}, (B_1 \cap B_2)) \quad (18)$$

Remark 6. Let L(U, A_1, R_1) and L(U, A_2, R_2) be two concept lattices. If, for any (X, B) ∈ L(U, A_2, R_2), there exists (X', B') ∈ L(U, A_1, R_1) such that X' = X, then L(U, A_1, R_1) is said to be finer than L(U, A_2, R_2), denoted by:

$$L(U, A_1, R_1) \leq L(U, A_2, R_2) \quad (19)$$

If, in addition, L(U, A_1, R_1) ≤L(U, A_2, R_2), we say that the two concept lattices are isomorphic, denoted by:

$$L(U, A_1, R_1) \cong L(U, A_2, R_2) \quad (20)$$

Example 3. A document collection's formal context (U, A, R) is given in Table 1; where U = {d_1, d_2, ..., d_7} and A = {a, b, ..., f }. For the formal context given in Table 1, the corresponding concept lattice is given in Figure 4. For simplicity, a set is denoted by listing its elements. For example, the set d_2, d_3 is denoted by 23.

Table 1. A document collection's formal context example

	a	b	c	d	e	f
d_1	1	1	1	1	0	0
d_2	0	0	0	1	0	1
d_3	0	0	0	1	1	1
d_4	0	0	1	0	1	0
d_5	0	0	1	1	1	0
d_6	0	0	1	0	0	1
d_7	1	1	1	1	1	0

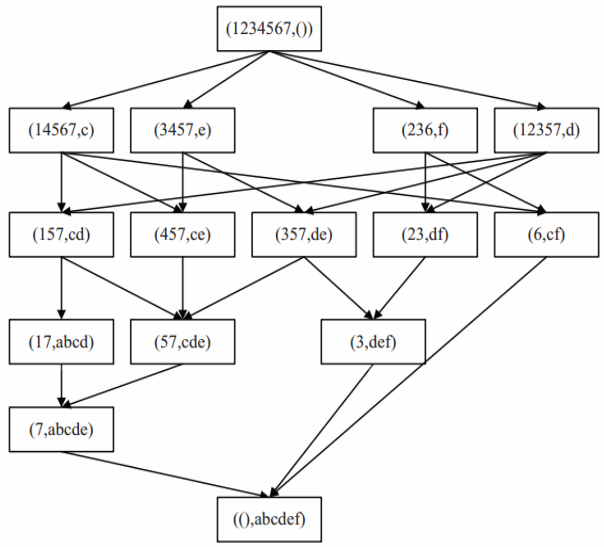


Figure 4. The concept lattice from Example 3

3.5 Generation of Query Concepts Using a Lattice of Description Topics

We assume that an initial query and/or documents contain several ideas or purposes. In other words, they are designed to represent several topics. Therefore, to enhance an initial query, we select its most similar DTs from the concept lattice and generate all its possible interpretations under a disjunctive normal forms (DNFs) form. The most probable DTs in the selected lattice node(s) are chosen to construct query concepts.

We propose to expand the initial queries by using the following methodology:

1. Build the DTs lattice of the retrieved document collection using the increment or batch lattice constructing algorithms.
2. Select first top N (here we choose N=1 or 2) DTs lattice nodes that are similar to the initial query q, using concept similarity method [64].
3. Select the DTs from the N nodes and extract the terms included in the topics.
4. Generate all the possible combinations using the extracted terms under a DNF form. These are called the candidates query concepts.
5. Choose the query concepts that are most similar to the initial query q, using the concept similarity method. The selected query concept is called QC.
6. Choose m high-frequency terms (usually between 5 and 10 terms) from the selected QC.
7. Construct the final expanded query qe = αq +βQC, where $0 \le \alpha \le 1$ and $\beta = 1-\alpha$. α and β are called weighting constants.

Figure 5 illustrates the process graphically. From all the possible DNFs, we select the one that is most similar to q_0 as the best query concept.

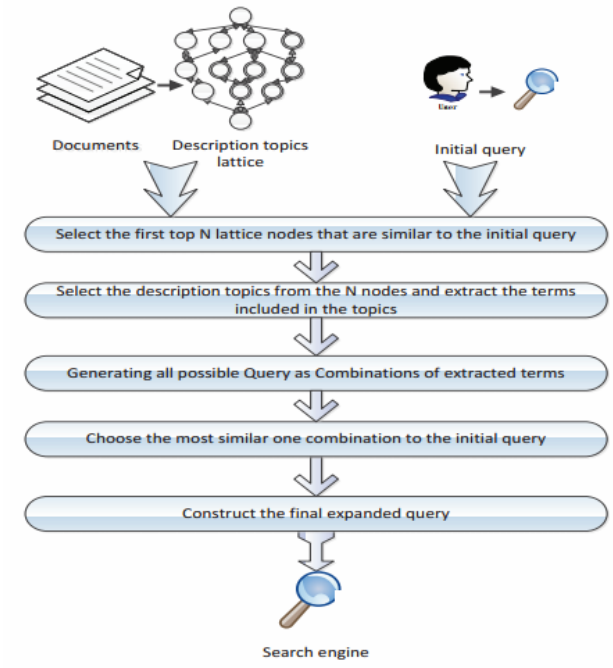


Figure 5. The procedure of generating query concepts in query expansion

4 Experiment

4.1 Experiment Setup

In order to evaluate the effectiveness of the proposed model, we conducted experiments using the TREC-7 and TREC-8 information retrieval test collections, as well as the TREC AP89 collection (TIPSTER disk 1). The TREC-7 and TREC-8 test collections consist of 50 topics (queries), respectively, and 528,155 documents from several sources: the Financial Times (FT), the Federal Register (FR94), the Foreign Broadcast Information Service (FBIS), and the LA Times. The TREC AP89 test collection contains 84,678 documents in which we use topics 151-200 as queries [65-66]. Each topic consists of three sections, the "Title", "Description" and "Narrative". Table 2 shows statistics of the document collections. For the query, we use the title and the description only. Note that in the TREC-7 and TREC-8 collections, the description section contains all the terms in the title section.

Table 2. Test document collection statistics

Source	Size (MB)	No. of docs	Words/doc. (mean)
TREC-7 and TREC-8			
FT	564	210,158	412.7
FR94	395	55,630	644.7
FBIS	470	130,471	543.6
LA Times	475	131,896	526.5
TIPSTER			
AP89	198	84,678	252

4.2 Baselines

We compare our method to:

(1) Traditional query expansion using the relevance feedback technique, in which 30 documents, among the documents retrieved in the initial retrieval, are used for feedback. We use the Rocchio formula for term reweighting as follows:

$$Q_{new} = \alpha \bullet Q_{old} + \beta \sum_{r=1}^{n_{rel}} \frac{D_r}{n_{rel}} - \gamma \sum_{n=1}^{n_{nonrel}} \frac{D_n}{n_{nonrel}} \qquad (21)$$

where α, β and γ are constants, D_r is the vector of a relevant document d_r, D_n is the vector of an irrelevant document d_n, n_{rel} is the number of relevant documents retrieved, and n_{nonrel} is the number of irrelevant documents. We set $\alpha = 8$, $\beta = 16$, and $\gamma = 4$ for this experiment [67].

(2) A light ontology method by Dragoni et al. [44], named "OntoConcept" in experiments, in which a vector space model approach is designed to represent documents and queries based on concepts instead of terms, as well as using WordNet as a light ontology.

(3) The two systems att98atdc and tno8dc are presented at the TREC-7 and TREC-8 ad-hoc track.

These two systems will be tested on TREC-7 and TREC-8, respectively.

4.3 Experiment Results and Analysis

The DT method reformulates the initial query with the query concepts generated by the process described in Figure 5. For a given query, using the title and the description only, we initially retrieve all the documents that contain the query terms.

We perform these processes on each query, and then calculate the average precision of the queries on three test document collections. Figure 6 to Figure 8 shows the precisions on deferent recall and Figure 9 to Figure 11 shows the precisions on deferent top n documents. In order to examine the impact of sampling a subset of the top-ranked documents, we restrict the set of returned documents to only the top 5000(L5000), 1000(L1000), 500(L500), and 100(L100) retrieved documents in our method, respectively.

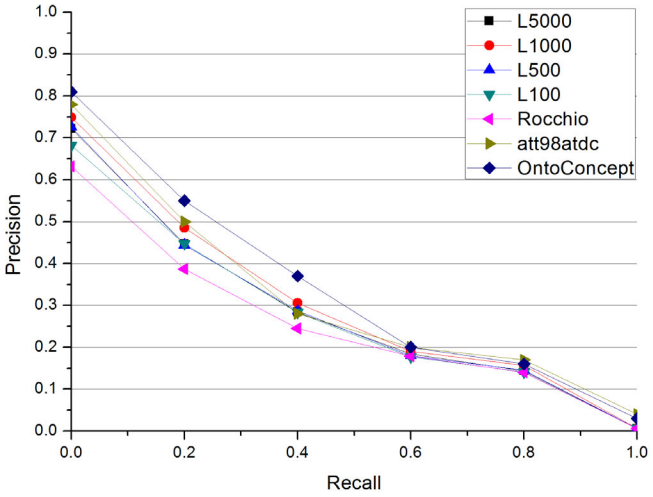


Figure 6. Precision on deferent recall on TREC-7 topics

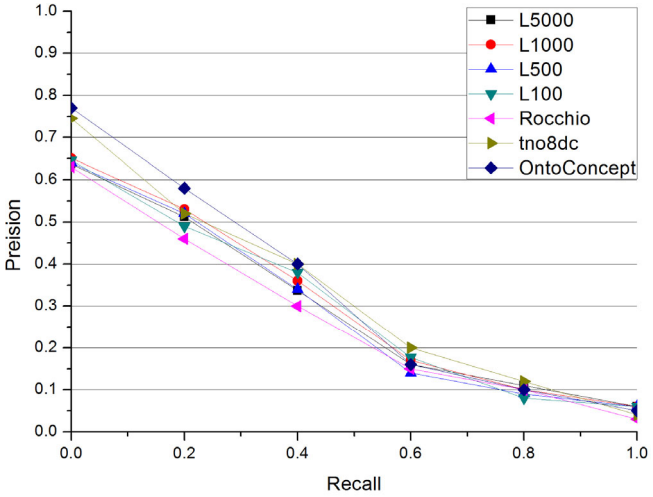


Figure 7. Precision on deferent recall on TREC-8 topics

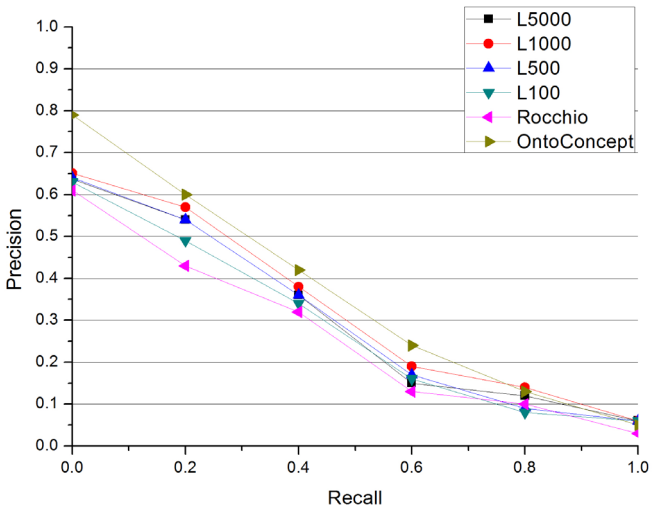


Figure 8. Precision on deferent recall on AP89 topics

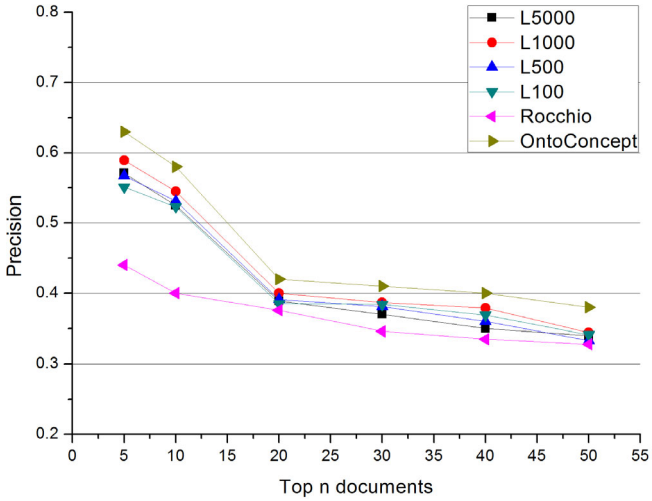


Figure 9. Precision on deferent top n documents on TREC- 7 topics

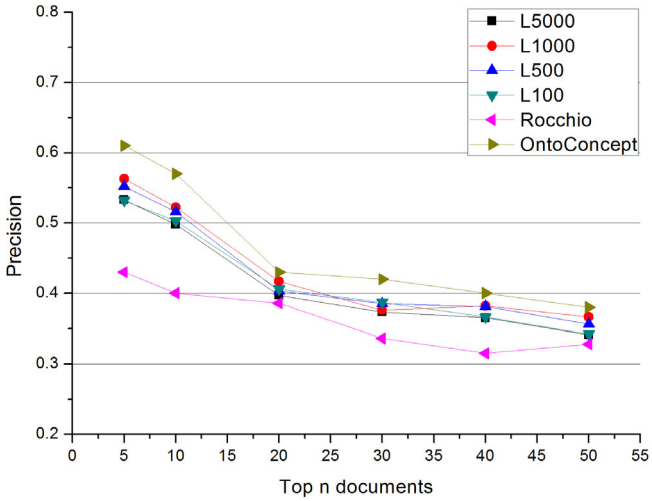


Figure 10. Precision on deferent top n documents on TREC-8 topics

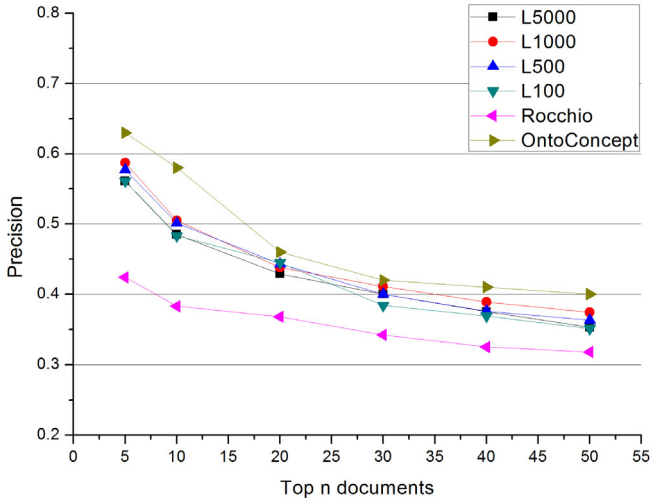


Figure 11. Precision on deferent top n documents on AP89 topics

From all these results, we may infer that:

(1) Precisions on recall demonstrate that our method is better than the traditional relevance feedback model in Figure 6 to Figure 8, and as good as the att98atdc and tno8dc systems in Figure 6 to Figure 7. The possible explanation is that we use FCA to construct query concepts, which could express some semantic meanings of queries and documents than traditional methods.

(2) Precisions on recall and top n documents show that the light ontology-based method is better than ours in Figure 6 to Figure 11. Our method, and the light ontology-based method, both use the vector space model. The light ontology-based method deals with the whole dataset (against retrieved partial documents) based on concepts instead of terms as well as using WordNet as a light ontology, which could capture and express more accurate semantic meanings of query and documents.

(3) Precisions on top documents from 5 to 20 on TREC-7, TREC-8 and AP89 in Figure 9 to Figure 11 have a rapid decline, and from 10 to 50, have relatively stable performance. The top 5 documents retrieved gain more precision than other parts.

(4) The results do not act well when the returned documents are 5000 in our method. But when the returned documents are 500 and 1000, the precisions are higher than the baseline. It means that more returned documents may not perform as well, and it is important to choose moderate returned documents.

5 Conclusions

In this paper, we have discussed query expansion using concept lattices from a document collection. The query concepts, which are meant to precisely denote the user's information needs, are based on the extraction of DTs from the retrieved document collection and demonstrated that we could obtain better results or performance.

The proposed approach has many potential applications, in addition to query expansion. Firstly, the methods, e.g. summarization, clustering, and classification were used to construct approximate DTs. However, if we develop more appropriate methods for constructing the lattice, a greater improvement in retrieval performance can be achieved. Secondly, since our approach is able to construct lattice of DTs directly from a document collection, it could be used as a fundamental step in the building of ontology for a given collection.

Another interesting application of the proposed methodology would be the prediction of the quality of the initial query, which is still an open research question. If we can predict the quality of the initial query, we could use it to expand an appropriately enhanced query. Similarly, we are considering other methods, not only for the purpose of constructing a lattice of DTs from document spaces more effectively, but also for the generation of a probable query concept.

Acknowledgements

This work was supported by the National Key Research and Development Program of China [No. 2018YFB1003903]; and National Natural Science Foundation of China [No. 61502033, 61472034, 61772071, 61272361, and 61672098].

References

[1] I.-C. Wu, G.-W. Chen, J.-L. Hsu, C.-Y. Lin, An Entropy-Based Query Expansion Approach for Learning Researchers' Dynamic Information Needs, *Knowledge-Based Systems*, Vol. 52, pp. 133-146, November, 2013.

[2] G. W. Furnas, T. K. Landauer, L. M. Gomez, S. T. Dumais, The Vocabulary Problem in Human-system Communication, *Communications of the ACM*, Vol. 30, No. 11, pp. 964-971, November, 1987.

[3] H. Imran, A. Sharan, A Framework for Automatic Query Expansion, *Proceedings of the 2010 International Conference on Web Information Systems and Mining*, Sanya, China, 2010, pp. 386-393.

[4] R. Mandala, T. Tokunaga, H. Tanaka, Query Expansi-on Using Heterogeneous Thesauri, *Information Processing & Management*, Vol. 36, No. 3, pp. 361-378, May, 2000.

[5] K. Riaz, Concept Search in Urdu, *Proceedings of the 2nd PhD Workshop on Information and Knowledge Management*, New York, NY, 2008, pp. 33-40.

[6] G. J. Hahm, M. Y. Yi, J. H. Lee, H. W. Suh, A Personalized Query Expansion Approach for Engineering Document Retrieval, *Advanced Engineering Informatics*, Vol. 28, No. 4, pp. 344-359, April, 2014.

[7] D. Zhou, S. Lawless, V. Wade, Improving Search via Personalized Query expansion Using Social Media,

Information Retrieval, Vol. 15, No. 3-4, June, 2012, pp. 218-242.

[8] Y. Chang, I. Choi, J. Choi, M. Kim, V. V. Raghavan, Conceptual Retrieval Based on Feature Clustering of Documents, *Proceedings of ACM SIGIR Workshop on Mathematics/ Formal Methods in Information Retrieval*, Tampere, Finland, 2002, pp. 89-104.

[9] R. Wille, Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts, in: S. Ferré, S. Rudolph (Eds.), *Formal Concept Analysis. ICFCA 2009. Lecture Notes in Computer Science*, Vol. 5548, Spring, 2009, pp. 314-339.

[10] C. Carpineto, G. Romano, A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval, *Machine Learning*, Vol. 24, No. 2, pp. 95-122, August, 1996.

[11] A. Formica, Semantic Web Search Based on Rough Sets and Fuzzy Formal Concept Analysis, *Knowledge-Based Systems*, Vol. 26, No. 9, pp. 40-47, February, 2012.

[12] C. De Maio, G. Fenza, V. Loia, S. Senatore, Hierarchical Web Resources Retrieval by Exploiting Fuzzy Formal Concept Analysis, *Information Processing & Management*, Vol. 48, No. 3, pp. 399-418, May, 2012.

[13] A. Qadi, D. Aboutajedine, Y. Ennouary, Formal Concept Analysis for Information Retrieval, *International Journal of Computer Science & Information Security*, Vol. 7, No. 2, pp. 117-121, February, 2010.

[14] K. Mihye, C. Paul, A Hybrid Browsing Mechanism Using Conceptual Scales, in: A. Hoffmann, B. Kang, D. Richards, D. Tsumoto (Eds.), *Advances in Knowledge Acquistion and Management PKAW 2006. Lecture Notes in Computer Science*, Vol. 4303, Springer, 2006, pp. 132-143.

[15] Z. Gong, C. W. Cheang, L. Hou U, Multi-term Web Query Expansion Using Wordnet, *Proceedings of the 17th International Conference on Database and Expert Systems Applications*, Krakow, Poland, 2006, pp. 379-388.

[16] P. Goyal, L. Behera, T. M. McGinnity, Query Repres-entation through Lexical Association for Information Re-trieval, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 12, pp. 2260-2273, December, 2012.

[17] J. Xu, W. B. Croft, Query Expansion Using Local and Global Document Analysis, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in information retrieval*, Zurich, Switzerland, 1996, pp. 4-11.

[18] M. M. Rahman, S. K. Antani, G. R. Thoma, A Query Expansion Framework in Image Retrieval Domain Based on Local and Global Analysis, *Information Processing & Management*, Vol. 47, No. 5, pp. 676-691, September, 2011.

[19] F. Colace, M. D. Santo, L. Greco, P. Napoletano, Weighted Word Pairs for Query Expansion, *Information Processing & Management*, Vol. 51, No. 1, pp. 179-193, July, 2015.

[20] A. Shiri, C. Revie, Query Expansion Behavior within a Thesaurus-enhanced Search Environment: A User-centered Evaluation, *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 4, pp. 462-478, February, 2006.

[21] J. Zhang, C. Shi, B. Deng, X. Li, Using Wordnet in Conceptual Query Expansion, in: T. Kim, W. C. Fang, C. Lee, K. P. Arnett (Eds.), *Advances in Software Engineering, ASEA 2008. Communications in Computer and Information Science*, Vol. 30, Springer, 2009, pp. 210-218.

[22] J. Zhang, B. Deng, X. Li, Concept Based Query Expansion Using Wordnet, *Proceedings of the 2009 International e-Conference on Advanced Science and Technology*, Washington, DC, 2009, pp. 52-55.

[23] J. Lei, W. Li, F. Wang, H. Deng, A Survey on Query Expansion Based on Local Analysis, *Proceedings of the 2011 4th International Conference on Intelligent Networks and Intelligent Systems*, Washington, DC, 2011, pp. 1-4.

[24] Z. Liu, S. Natarajan, Y. Chen, Query Expansion Based on Clustered Results, *Proceedings of the Vldb Endowment*, Vol. 4, No. 6, pp. 350-361, March, 2011.

[25] Q. Huang, D. Song, A Latent Variable Model for Query Expansion Using the Hidden Markov Model, *Proceedings of the 17th ACM Conference On Information and Knowledge Management*, New York, NY, 2008, pp. 1417-1418.

[26] J. Bai, J.-Y. Nie, Adapting Information Retrieval to Query Contexts, *Information Processing & Management*, Vol. 44, No. 6, pp. 1901-1922, July, 2008.

[27] B. M. Fonseca, P. Golgher, B. Possas, B. Ribeiro-Neto, N. Ziviani, Concept-based Interactive Query Expansion, *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, New York, NY, 2005, pp. 696-703.

[28] Y. Qiu, H.-P. Frei, Concept Based Query Expansion, *Proceedings of the 16th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, 1993, pp. 160-169.

[29] A. P. Natsev, A. Haubold, J. Tesic, L. Xie, R. Yan, Semantic Concept-based Query Expansion and Re-ranking for Multimedia Retrieval, *Proceedings of the 15th International Conference on Multimedia*, New York, NY, 2007, pp. 991-1000.

[30] G. Cao, J. Gao, J.-Y. Nie, J. Bai, Extending Query Translation to Cross-language Query Expansion with Markov Chain Models, *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, New York, NY, 2007, pp. 351-360.

[31] R. Yan, A. Hauprmann, Query Expansion Using Probabilistic Local Feedback with Application to Multimedia Retrieval, *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, New York, NY, 2007, pp. 361-370.

[32] J. Xu, W. B. Croft, Improving the Effectiveness of Information Retrieval with Local Context Analysis, *ACM Transactions on Information Systems*, Vol. 18, No. 1, pp. 79-112, January, 2000.

[33] C. Carpineto, R. de Mori, G. Romano, B. Bigi, An Information-theoretic Approach to Automatic Query Expansion, *ACM Transactions on Information Systems*, Vol. 19, No. 1, pp. 1-27, January, 2001.

[34] C. Carpineto, G. Romano, A Survey of Automatic Query Expansion in Information Retrieval, *ACM Computing Surveys*,

Vol. 44, No. 1, pp. 1-50, January, 2012.

[35] M.-A. Cartright, J. Allan, V. Lavrenko, A. McGregor, Fast Query Expansion Using Approximations of Relevance Models, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Toronto, Canada, 2010, pp. 1573-1576.

[36] C. Zhai, J. Lafferty, Model-based Feedback in the Language Modeling Approach to Information Retrieval, *Proceedings of the Tenth International Conference on Information and Knowledge Management*, New York, NY, 2001, pp. 403-410.

[37] D. Metzler, W. B. Croft, Latent Concept Expansion Using Markov Random Fields, *Proceedings of the 30th annual international ACMSIGIR conference on Research and development in information retrieval*, New York, NY, 2007, pp. 311-318.

[38] H. Lang, D. Metzler, B. Wang, J.-T. Li, Improved Latent Concept Expansion Using Hierarchical Markov Random Fields, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, New York, NY, 2010, pp. 249-258.

[39] K. Collins-Thompson, J. Callan, Estimation and Use of Uncertainty in Pseudo-relevance Feedback, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, 2007, pp. 303-310.

[40] G. Cao, J.-Y. Nie, J. Gao, S. Robertson, Selecting Good Expansion Terms for Pseudo-relevance Feedback, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, 2008, pp. 243-250.

[41] N Stojanovic, An Approach for Defining Relevance in the Ontology-based Information Retrieval, *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, 2005, pp. 359-365.

[42] M. Baziz, M. Boughanem, G. Pasi, H. Prade, An Information Retrieval Driven by Ontology from Query to Document Expansion, *LE Centre De Hautes Etudes Internationals D'informatique Documentaire*, Paris, France, 2007, pp. 301-313.

[43] M. Dragoni, C. D. C. Pereira, A. G. B. Tettamanzi, An Ontological Representation of Documents and Queries for Information Retrieval Systems, *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Berlin, Germany, 2010, pp. 555-564.

[44] M. Dragoni, C. D. C. Pereira, A. G. B. Tettamanzi, A Conceptual Representation of Documents and Queries for Information Retrieval Systems by Using Light Ontologies, *Expert Systems with Applications*, Vol. 39, No. 12, pp. 10376-10388, January, 2012.

[45] J. Waitelonis, C. Exeler, H. Sack, Linked Data Enabled Generalized Vector Space Model to Improve Document Retrieval, *Proceedings of NLP & DBpedia 2015 Workshop in Conjunction with 14th International Semantic Web Conferenc*, Bethlehem, PA, 2015, pp. 1-12.

[46] F. Corcoglioniti, M. Dragoni, M. Rospocher, A. P. Aprosio, Knowledge Extraction for Information Retrieval, *International Semantic Web Conference. Springer International Publishing*, Monterey, CA, 2016, pp. 317-333.

[47] M. A. Hearst, J. O. Pedersen, Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, *Proceedings of the 19th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, 1996, pp. 76-84.

[48] A. Leuski, Evaluating Document Clustering for Interactive Information Retrieval, *Proceedings of the Tenth International Conference on Information and Knowledge Management*, New York, NY, 2001, pp. 33-40.

[49] J. Kupiec, J. Pedersen, F. Chen, A Trainable Document summarizer, *Proceedings of the 18th Annual International ACM SI- GIR Conference on Research and Development in Information Retrieval*, New York, NY, 1995, pp. 68-73.

[50] J. K. Tarus, Z. D. Niu, A. Yousif, A Hybrid Knowledge-based Recommender System for e-learning Based on Ontology and Sequential Pattern Mining, *Future Generation Computer Systems*, Vol. 72, pp. 37-48, July, 2017.

[51] J. K. Tarus, Z. D. Niu, G. Mustafa, Knowledge-based Recommendation: A Review of Ontology-based Recommender Systems for e-learning, *Artificial Intelligence Review*, Vol. 50, No. 1, pp. 21-48, June, 2018.

[52] J. K. Tarus, Z. D. Niu, D. Kalui, A Hybrid Recommender System for e-learning Based on Context Awareness and Sequential Pattern Mining, *Soft Computer*, Vol. 22, No. 8, pp. 2449-2461, April, 2018.

[53] S. S. Wan, Z. D. Niu: An e-learning Recommendation Approach Based on the Self-organization of Learning Resource, *Knowl.-Based Systems*, Vol. 160, pp. 71-87, November, 2018.

[54] W. Chen, Z. D. Niu, X. Y. Zhao, Y. Li, A Hybrid Recommendation Algorithm Adapted in e-learning Environments, *World Wide Web*, Vol. 17, No. 2, pp. 271-284, March, 2014.

[55] K. J. Mock, Hybrid Hill-climbing and Knowledge-based Methods for Intelligent News Filtering, *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Portland, Oregon, 1996, pp. 48-53.

[56] P. Achananuparp, *Similarity Measures and Diversity Rankings for Query-focused Sentence Extraction*, Ph.D. Thesis, Philadelphia, PA, 2010.

[57] M. Lesk, Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, *Proceedings of the 5th Annual International Conference on Systems Documentation*, New York, NY, 1986, pp. 24-26.

[58] W. B. Frakes, R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, 1992.

[59] B. Ganter, R. Wille, *Formal Concept Analysis: Mathematical Foundations*, Springer-Verlag, 1997.

[60] G. Mineau, R. Godin, Automatic Structuring of Knowledge Bases by Conceptual Clustering, *Knowledge and Data Engineering*, Vol. 7, No. 5, pp. 824-829, October, 1995.

[61] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, L. Lakhal, Computing Iceberg Concept Lattices with Titanic, *Data & Knowledge Engineering*, Vol. 42, No. 2, pp. 189-222, August,

2002.

[62] R. Wille, Knowledge Acquisition by Methods of Formal Concept Analysis, in: E. Diday (Ed.), *Data Analysis, Learning Symbolic and Numeric Knowledge*, Nova Science, 1989, pp. 365-380.

[63] K. X. S. de Souza, J. Davis, Aligning Ontologies and Evaluating Concept Similarities, in: R. Meersman, Z. Tari (Eds.), *Lecture Notes in Computer Science*, Vol. 3291, Springer, 2004, pp. 1012-1029.

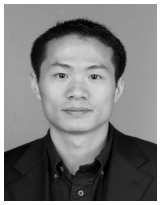
[64] C. Shi, Z. Niu, A Novel Similarity Evaluating Model Based on RFCA and ICS, *Fifth International Conference on Digital Information Management*, Thunder Bay, Canada, 2010, pp. 114-119.

[65] E. M. Voorhees, D. Harman, Overview of the Seventh Text Retrieval Conference, *NIST Special Publication 500-242: The Seventh Text Retrieval Conference (TREC-7)*, NIST, 1996.

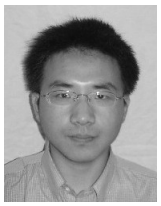
[66] D. Hawking, E. Voorhees, N. Craswell, P. Bailey, *Overview of the Trec-8 Web Track*, Vol. 33, pp. 131-148, December, 2009.

[67] C. Marton, Salton and Buckleys Landmark Research in Experimental Text Information retrieval, *Evidence Based Library and Information Practice*, Vol. 6, No. 4, pp. 169-176, December, 2011.

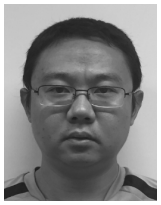
Biographies



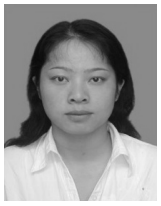
Haibin Yu received his master's degree from school of computing, Beijing Institute of Technology, China, in 2006. He is currently an University Staff in Beijing Institute of Technology. His research interests include knowledge management, machine learning, etc.



Chongyang Shi is a lecturer in School of Computer Science, Beijing Institute of Technology. He obtained his PhD degree from Beijing Institute of Technology in 2010, all in Computer Science. His research areas focus on Information Retrieval, Knowledge Engineering, Personalized Service, Sentiment Analysis etc.



Yu Bai received his B.E. and M.E. degree in computer science from Tianjin University, China. Currently, he is a senior research assistant at the Centre for Quantum Computation and Intelligent Systems (QCIS), University of Technology Sydney, Australia. His research focus on data mining and machine learning.



Chunxia Zhang received her Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2005. She is currently an associate professor in School of Software, Beijing Institute of Technology. Her research interests include information extraction, knowledge management, machine learning, etc.

Ryan Hearne is a Computer Science Masters student at Beijing Institute of Technology. His area of research is data mining and sentiment analysis using Python. He graduated with a First Class Honours degree in Software Engineering from Waterford Institute of Technology in 2015. His research focuses on mobile and web development.