Engineering Applications of Artificial Intelligence 69 (2018) 65–75

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Concept coupling learning for improving concept lattice-based document retrieval

Shufeng Hao [a], Chongyang Shi [a,*], Zhendong Niu [a], Longbing Cao [b]

[a] School of Computer Science, Beijing Institute of Technology, 100081, China
[b] Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia

ARTICLE INFO

Keywords:
Fuzzy formal concept analysis
Lattice-based document retrieval
Coupling relationship

ABSTRACT

The semantic information in any document collection is critical for query understanding in information retrieval. Existing concept lattice-based retrieval systems mainly rely on the partial order relation of formal concepts to index documents. However, the methods used by these systems often ignore the explicit semantic information between the formal concepts extracted from the collection. In this paper, a concept coupling relationship analysis model is proposed to learn and aggregate the intra- and inter-concept coupling relationships. The intra-concept coupling relationship employs the common terms of formal concepts to describe the explicit semantics of formal concepts. The inter-concept coupling relationship adopts the partial order relation of formal concepts to capture the implicit dependency of formal concepts. Based on the concept coupling relationship analysis model, we propose a concept lattice-based retrieval framework. This framework represents user queries and documents in a concept space based on fuzzy formal concept analysis, utilizes a concept lattice as a semantic index to organize documents, and ranks documents with respect to the learned concept coupling relationships. Experiments are performed on the text collections acquired from the SMART information retrieval system. Compared with classic concept lattice-based retrieval methods, our proposed method achieves at least 9%, 8% and 15% improvement in terms of average MAP, IAP@11 and P@10 respectively on all the collections.

1. Introduction

With the rapid growth of Web data, query understanding plays an essential role in obtaining information which is relevant to the user's needs. Classic information retrieval (IR) systems often rely on keyword-matching to index documents from the corpus, where queries and documents are represented by methods such as the Boolean Model, Vector Space Model and Probabilistic Model. In practice, however, existing retrieval systems often return inaccurate and incomplete results due to semantic challenges such as polysemy and synonymy. This is known as vocabulary or word mismatch (Furnas et al., 1987).

Various efforts have been made to address the word mismatch problem, such as query expansion techniques and concept lattice-based retrieval methods for query transformation. Query expansion generates a novel query by augmenting the original query with new features with similar meaning, where the features are additional terms extracted from a thesaurus, such as WordNet, explicit relevance feedback or pseudo relevance feedback (Carpineto and Romano, 2012). Rather than incorporating extra terms from other data sources to expand the original query, concept lattice-based retrieval methods can refine and expand the query and explore navigation search strategies using the specificity/generality relation of the concept lattice (Priss, 2000; Carpineto and Romano, 2005).

Concept lattice-based retrieval methods are based on formal concept analysis (FCA) (Ganter and Wille, 1999), a type of unsupervised classification that provides an intentional description for clusters, which contributes to better understanding. The concept lattice generated by FCA has demonstrated its usefulness in document indexing and navigation strategy in the IR domain (Priss, 2000; Carpineto and Romano, 2005; Codocedo and Napoli, 2015). For instance, the concept lattice can be used to drive the transformation between the representation of a query and the representation of each document and provide the navigation in a conceptual document space (Carpineto and Romano, 2000; Messai et al., 2010). Meanwhile, some methods have been proposed to obtain the semantic information between formal concepts (Formica, 2008; Codocedo et al., 2014). These approaches only consider whether terms occur in queries and documents, but regarding all terms equally may

* Corresponding author.
E-mail address: cy_shi@bit.edu.cn (C. Shi).

https://doi.org/10.1016/j.engappai.2017.12.007
Received 2 March 2017; Received in revised form 5 September 2017; Accepted 9 December 2017
0952-1976/

significantly reduce the quality of the retrieved outcomes since different terms may have different degrees of importance for those queries and documents. This type of problem can be tackled with fuzzy information (Formica, 2010).

To overcome the problem of uncertain, vague and implicit information in queries and documents for IR, fuzzy formal concept analysis (FFCA) can be adopted to model these characteristics by incorporating fuzzy logic into FCA (Bělohlávek et al., 2005). Several approaches using fuzzy concept lattices based on FFCA (Formica, 2012; Poelmans et al., 2014; Kumar et al., 2015) have been proposed to deal with this challenge. In these methods, queries and documents are represented by fuzzy formal concepts that consist of vague (non-crisp) extents and intents, i.e., crisply generated concepts (here, 'extent' refers to an object set in a concept, and 'intent' refers to an attribute set in a concept). They adopt the partial order relation of concepts to compute the relationship between concepts and return related documents for the given query. However, these methods neglect the explicit semantic information between concepts (the common objects and attributes of concepts). As a result, the coupling relationship between concepts, consisting of the common terms (objects and attributes) of concepts and the partial order relations of concepts, is neglected.

Learning coupling relationships, i.e. coupling learning (Cao, 2015), has demonstrated its significant value in improving existing analytical and learning tasks, e.g., similarity learning for clustering (Cheng et al., 2013), classification (Liu et al., 2014), recommendation systems (Li et al., 2013), keyword queries (Meng et al., 2014), and outlier detection (Pang et al., 2016). In this work, we propose a novel approach to measure the coupling relationship between concepts by capturing both the intra-concept coupling relation (explicit semantic similarity) and the inter-concept coupling relation (implicit semantic similarity) based on FFCA and the fuzzy concept lattice. The intra-concept coupling relation directly reveals the similarity between concepts by considering the common objects and attributes of concepts, and the inter-concept coupling relation reveals the dependency aggregation between concepts by exploring the topological distance between concepts based on the partial order relation of concepts in the concept lattice. Using this observation, the concept coupling relationship is used to generate a semantic similarity measure between the given query concept and other concepts. Lastly, we represent documents in a concept space and rank them based on the semantic similarity measure. The key contributions of this paper are as follows:

- The intra-concept coupling relation is learned to describe the explicit semantics of concepts by calculating the intersection of the intent and vague extent of concepts based on the Jaccard measure.
- The inter-concept coupling relation is analyzed to capture the implicit dependency of concepts by their topological distance based on the hierarchical structure of the lattice and the partial order relation of concepts.
- A novel concept lattice-based retrieval system based on the learned concept coupling relationships is proposed, which aggregates the intra- and inter-concept couplings, and we rank documents in a concept space using this system.

Substantial experiments are undertaken to test our method by comparing four currently used document retrieval techniques on text collections acquired from the SMART information retrieval system. The performance of our method is evaluated in terms of mean average precision, 11-point interpolated average precision and precision in the first 10 ranked documents. The results show that our approach achieves significant improvement over the baselines.

The rest of the paper is organized as follows. The preliminary work in this area is in Section 2. Section 3 introduces the framework of our proposed concept lattice and coupling learning-based retrieval system. Section 4 learns the concept coupling relationships, and a lattice-based retrieval system based on the concept coupling relationship is detailed in Section 5. The experimental results are presented in Section 6, followed by a summary of related work. Lastly, Section 8 concludes the paper and presents the prospective future work.

Table 1
Fuzzy formal context K for document representation using a threshold $T = 1/6$.

	DM	ML	TM	TR
d_1	0	2/3	0	1/3
d_2	0	0	1/2	1/2
d_3	0	0	0	1/3
d_4	1/4	0	0	0
d_5	1/2	1/3	1/6	0
d_6	0	0	1/2	0
d_7	2/3	1/3	0	0

2. Preliminary

In this section, the preliminary work consisting of fuzzy formal concept analysis and concept lattice-based retrieval is introduced in detail.

2.1. Fuzzy formal concept analysis

Definition 1 (*Fuzzy Formal Context*). A *fuzzy formal context* (fuzzy context for short) $K = (O, A, R = \phi(O \times A))$ consists of an object set O, an attribute set A, and a fuzzy relation R in $O \times A$. Each pair $(o, a) \in R$ has a membership value $\mu(o, a)$ ($\in [0, 1]$), meaning object o has attribute a with membership grade $\mu(o, a)$. The set $R = \phi(O \times A) = \{((o, a), \mu(o, a))|\forall o \in O, a \in A, \mu : O \times A \rightarrow [0, 1]\}$ is a fuzzy relation in $O \times A$.

Two derivation operators $(.)'$ for $E \subseteq O$, and $I \subseteq A$ in the fuzzy context $K = (O, A, R)$ with a confidence threshold T are defined as follows:

$$E' = \{a \in A |\, \mu(o, a) \geqslant T, \forall o \in E\} \tag{1}$$

$$I' = \{o \in O |\, \mu(o, a) \geqslant T, \forall a \in I\}. \tag{2}$$

Definition 2 (*Fuzzy Formal Concept*). A *fuzzy formal concept* (fuzzy concept for short) of a fuzzy context $K = (O, A, R = \phi(O \times A))$ with a threshold T is a pair $(\phi(E), I)$, where $E \subseteq O$ and $I \subseteq A$, $E' = I$, $I' = E$. Each object $o \in E$ has a membership value μ_o defined as $min_{a \in I}\, \mu(o, a)$, thus $\phi(E) = \{(o_1, \mu_0(o_1)), (o_2, \mu_0(o_2)), \ldots, (o_m, \mu_0(o_m))|o_i \in E\}$. The sets E and I are respectively called the *extent* and *intent* of the fuzzy concept.

The set $B(O, A, R)$, consisting of all fuzzy concepts from the fuzzy context K, is ordered by *inheritance relation* ($\leq$) as follows:

$$(\phi(E_1), I_1) \leq (\phi(E_2), I_2) \Leftrightarrow \phi(E_1) \subseteq \phi(E_2)\ or\ I_2 \subseteq I_1. \tag{3}$$

Thus $(\phi(E_1), I_1)$ is called a *sub-concept* of $(\phi(E_2), I_2)$ and $(\phi(E_2), I_2)$ is called a *super-concept* of $(\phi(E_1), I_1)$. The *fuzzy concept lattice* $\underline{B}(O, A, R)$ of the fuzzy context K is defined as $(B(O, A, R), \leq)$, where $B(O, A, R)$ is all the concepts from the fuzzy context K. In addition, the fuzzy concept lattice has *supremum* and *infimum*, grouping all the objects and attributes respectively of the fuzzy context. For instance, consider a fuzzy context using the bag of words representation for documents in Table 1 with a threshold $T = 1/6$. Suppose that object set $O = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7\}$, and attribute set $A = \{DM, ML, TM, TR\}$, where "DM", "ML", "TM", "TR" denote "data mining","machine learning", "text mining", "text retrieval" respectively. The corresponding fuzzy concepts and the fuzzy concept lattice are shown in Table 2 and Fig. 1 respectively. The membership value of d_5 in C_1 is 1/6. Concept C_{10} is the supremum of the lattice, and concept C_2 is the infimum of the lattice.

Table 2
Notations for fuzzy concepts from the fuzzy context K.

Notations	Fuzzy concepts for corresponding notations
C_1	$((d_5(1/6)), (DM, ML, TM))$
C_2	$((), (DM, ML, TM, TR))$
C_3	$((d_5(1/3), d_7(1/3)), (DM, ML))$
C_4	$((d_2(1/2), d_5(1/6), d_6(1/2)), (TM))$
C_5	$((d_1(1/3)), (ML, TR))$
C_6	$((d_4(1/4), d_5(1/2), d_7(2/3)), (DM))$
C_7	$((d_1(2/3), d_5(1/3), d_7(1/3)), (ML))$
C_8	$((d_2(1/2)), (TM, TR))$
C_9	$((d_1(1/3), d_2(1/2), d_3(1/3)), (TR))$
C_{10}	$((d_1(1), d_2(1), d_3(1), d_4(1), d_5(1), d_6(1), d_7(1)), ())$

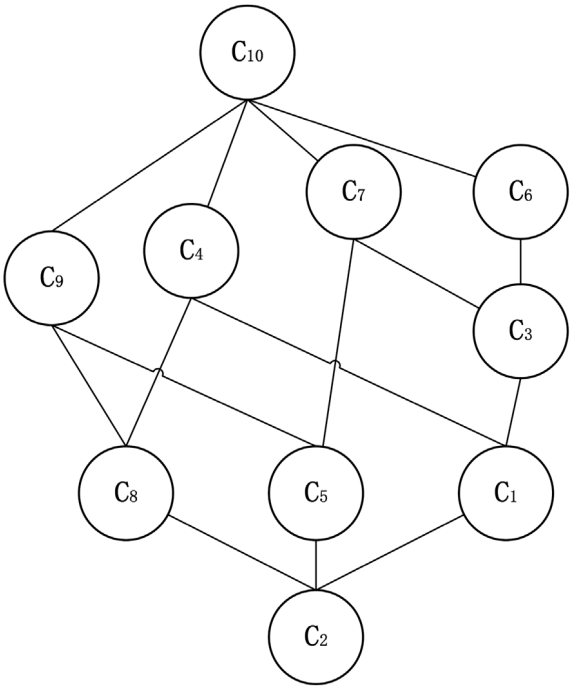

Fig. 1. The corresponding fuzzy concept lattice for the fuzzy context K.

2.2. Foundations of concept lattice-based retrieval

Concept lattice-based retrieval systems regard document searching as a transformation process from a query to each document in a concept space (Carpineto and Romano, 2000), based on a concept lattice that is a conceptual representation of a collection of documents. The concept lattice contains a set of concepts derived from the common terms found within that collection of documents. The intent of a concept provides a semantic "context" specific to the collection to describe the documents in the concept extent, following the assumption that if one term always appears jointly with other terms, the single terms do not refer to distinct concepts although their tuple does convey a useful meaning (Carpineto and Romano, 2000). To achieve a query-document transformation, a user query as a pseudo-document in a context is mapped into the concept lattice. In this manner, a query concept can be obtained when the intent of the concept is equal to the query description. For instance, concept C_1 in Fig. 1 is regarded as a query concept for the terms "data mining, machine learning, text mining". With the query concept as the starting point for IR, related concepts consisting of documents can be transformed from the query concept based on the generality/specificity relations in a concept lattice and type of navigation strategy.

Two main navigation strategies have been proposed in the literature. The neighborhood expansion (NE) strategy (Carpineto and Romano, 2000) transforms a query into each document in terms of the sequence of minimal refinements/enlargements determined by the concept lattice. The hierarchical exploration (HE) strategy (Messai et al., 2010) navigates the lattice by exploring the minimal enlargements that transform the query into each document. The ranked relevance of concepts determined by both navigation strategies can be computed in terms of the topological distance between both concepts, which is defined as the length of the shortest path between the two concepts. In Fig. 1, for instance, given the query concept C_1 for the query "data mining, machine learning, text mining", concepts C_2, C_3 and C_4 ranked at distance 1, concepts C_5, C_6, C_7, C_8 and C_{10} ranked at distance 2, and concepts C_9 ranked at distance 3 are obtained by using NE, while concepts C_3 and C_4 ranked at distance 1, and concepts C_6, C_7 and C_{10} ranked at distance 2 are obtained by using HE. For both strategies, the supremum C_{10} and the infimum C_2 of the lattice are omitted to compute the document ranking since the supremum has all the documents in the collection and the infimum has all the terms used to describe the documents in the collection. Thus NE provides all the documents, while HE provides the related documents containing d_1, d_2, d_4, d_5, d_6 and d_7. These observations illustrate that NE provides a larger quantity of relevant documents than HE, while the relevant documents HE provides are of better quality than the documents provided by NE. This has also been observed by Codocedo (Codocedo et al., 2014).

The navigation strategies discover the implicit semantic information between concepts based on the generality/specificity relations in a concept lattice. However, the strategies have several disadvantages. Some concepts obtain the same score as the query concept using both strategies. For example, concepts C_3 and C_4 obtain the same score as concept C_1. Meanwhile, both strategies ignore the explicit semantic information between concepts. The concept-based similarity measures integrate set-based similarity measures, such as the Jaccard index, to calculate the explicit semantic information between concepts. For instance, the similarity between concept $C_i = (\phi(E_i), I_i)$ and $C_j = (\phi(E_j), I_j)$ based on the Jaccard index is defined as follows:

$$S(C_i, C_j) = \frac{|E_i \cap E_j|}{|E_i \cup E_j|} * \lambda + \frac{|I_i \cap I_j|}{|I_i \cup I_j|} * (1 - \lambda) \tag{4}$$

where $|.|$ denotes the cardinality of the set, and λ is a parameter such that $\lambda \in [0, 1]$. The similarity measure considers the common objects and attributes, rather than the membership value between objects and attributes. To enrich the semantic information between concepts for IR, in this paper, we adopt a concept-based similarity measure that consists of the explicit and implicit semantic information between concepts. The explicit semantics of concepts is obtained by calculating the intersection of the intent and vague extent of concepts based on the Jaccard measure. The implicit semantics of concepts is computed by their topological distance based on the hierarchical structure of the lattice.

3. Framework

In this section, we introduce the proposed framework which consists of a five-step processing approach to rank documents in the collection. The framework and working mechanism are illustrated in Fig. 2.

The first step is document representation of the document collection. Documents in text format are preprocessed by text segmentation, removing stop words and word stemming. They are then represented by the Vector Space Model. Each document is represented as a vector $d_i = (w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots, w_{in})^{\mathrm{T}}$, where n is the number of all the distinct words which are extracted from the collection. The weight w_{ij} of term t_j in document d_i is calculated by the probability of that term occurring in the document.

The second step is lattice generation. The full collection of documents and queries is analyzed to form a document-term matrix $D = (d_1, d_2, \ldots, d_i, \ldots, d_m)^{\mathrm{T}}$, where m is the number of documents and queries in the collection. In general, each query is inserted into matrix D as a pseudo-document. We convert matrix D to a fuzzy formal context K. Each document d_i in matrix D is regarded as an object $o_i \in O$ of context

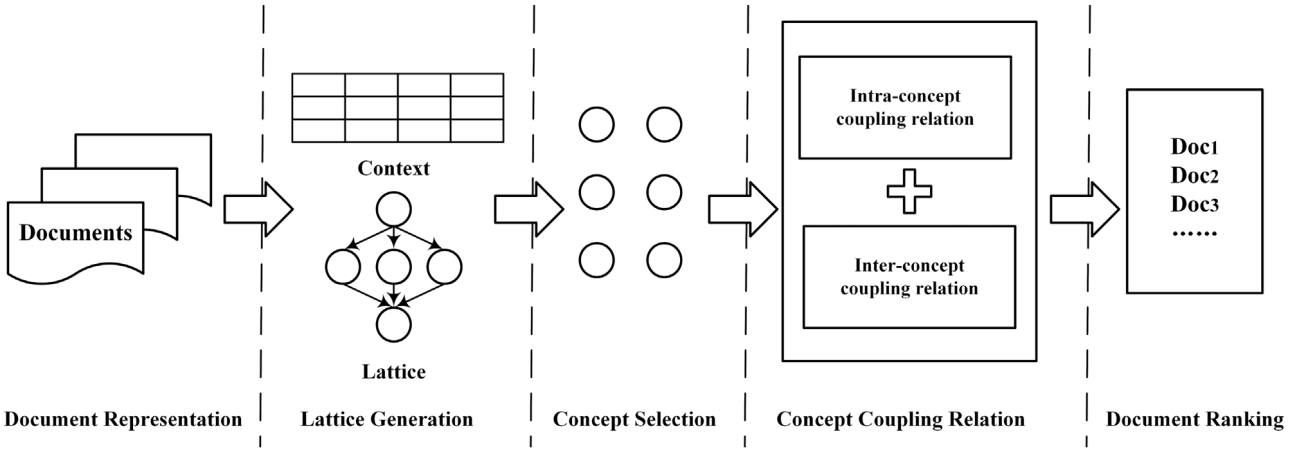

Fig. 2. The framework of lattice-based retrieval by integrating the concept coupling relationship.

K, and each term t_j in matrix D is regarded as an attribute $a_j \in A$ of context K. The membership value $\mu(o_i, a_j)$ of attribute a_j in object o_i is weight w_{ij} of term t_j in document d_i. Lastly, a concept lattice can be obtained from the fuzzy context K using lattice construction algorithms.

The third step is concept selection. A concept space can be constructed using the set of concepts, which is obtained by a lattice algorithm. Based on the concept space and the generality/specificity relation of concepts, query-document transformation is adopted for document ranking. To reduce the computing complexity for document ranking, important concepts need to be selected from the set of concepts in the constructed concept lattice. A concept is defined as an important concept when all the descriptive attributes of a document in a concept are equal to the intent of the concept.

The fourth step concerns the concept coupling relationship. Based on the important concepts and the concept lattice, the intra- and inter-concept coupling relations between concepts can be respectively calculated by leveraging the common objects and attributes of concepts with the Jaccard measure and the topological distance between concepts. The concept coupling relationship can then be characterized by a linear combination of the intra- and inter-concept coupling between concepts.

The final step is to rank documents in a concept space. Each document in the collection is represented by a concept in a concept space. A similarity measure between concepts is obtained based on concept coupling relationship analysis, and documents are ranked by the similarity measure.

4. Concept coupling relationship analysis

Motivated by the coupled nominal similarity in unsupervised learning (Wang et al., 2011) and the term coupling analysis in document analysis (Cheng et al., 2013), we propose concept coupling relationship analysis for IR. Intra- and inter-concept coupling relations are detailed in this section. Here concept coupling relationship analysis based on a concept lattice is presented.

4.1. Intra-concept coupling relation

A number of approaches have been proposed to measure the similarity between concepts and capture the relation between concepts in FFCA using set theory analysis (Formica, 2008, 2010). These approaches suppose that concepts are relational if they have common objects or attributes. For instance, fuzzy concepts "C_6" and "C_7" in Table 2 are similar since they have common documents, i.e., "d_5" and "d_7". Accordingly, the explicit relation between concepts in the fuzzy context is estimated by detecting the common objects and attributes. We employ the explicit relation between concepts to describe the intra-concept coupling relation between concepts in a concept space.

In most existing approaches, the explicit relation between concepts is simply estimated by the cardinality of the intersection of extents and intents, i.e., the number of elements in the set of the intersection of extents and intents. These methods ignore the weights of objects of extents (the weights of objects of extents refer to the membership values of objects of the extents). In the proposed method, the Jaccard measure (Bollegala et al., 2007) is applied to compute the explicit relation between concepts by integrating the weights of the extent objects. For the calculation of the explicit relation, the Jaccard measure can be replaced by other similarity measures, such as cosine similarity and matching coefficient, which will be explored in our future work.

Definition 3. Concepts $C_i = (\phi(E_i), I_i)$ and $C_j = (\phi(E_j), I_j)$ are related if they have common objects and attributes in the fuzzy context K. The *explicit relation* between C_i and C_j is quantified as:

$$ExR(C_i, C_j) = ER(E_i, E_j) * \lambda + \frac{|I_i \cap I_j|}{|I_i \cup I_j|} * (1 - \lambda) \tag{5}$$

$$ER(E_i, E_j) = \frac{1}{|H|} * \sum_{o \in H} \frac{\mu_{oi}\mu_{oj}}{\mu_{oi} + \mu_{oj} - \mu_{oi}\mu_{oj}} \tag{6}$$

where $|.|$ denotes the cardinality of the set, λ is a parameter such that $\lambda \in [0, 1]$; μ_{oi} and μ_{oj} represent the membership values of the object o of C_i and C_j; $ER(E_i, E_j)$ is the extent relationship between extent E_i and extent E_j respectively, and $|H|$ denotes the number of the elements in $H = \{o | o \in E_i \cap E_j\}$. If $H = \emptyset$, $ER(E_i, E_j) = 0$.

The intra-concept coupling relation is defined as a conditional probability manner by normalizing the relation $ExR(C_i, C_j)$ between C_i and C_j with respect to the total number of relations between concept C_i and other concepts. This is because our task is to rank documents for a given query. In the situation, concept C_i is regarded as a query concept. With normalization, the relation $ExR(C_i, C_j)$ is scaled into [0,1] to compose the concept coupling relationship with the inter-concept coupling relation (introduced below). The intra-concept coupling relation is defined as follows.

Definition 4. The *intra-concept coupling relation* between concepts $C_i = (\phi(E_i), I_i)$ and $C_j = (\phi(E_j), I_j)$ is defined as:

$$IaCR(C_i, C_j) = \begin{cases} 1 & i = j \\ \dfrac{ExR(C_i, C_j)}{\sum_{j=1, j \neq i}^{n} ExR(C_i, C_j)} & i \neq j \end{cases} \tag{7}$$

where n is the total number of all distinct concepts in the concept lattice.

From the above observations, we have $IaCR(C_i, C_j) \geq 0$ and $\sum_{j=1, j \neq i}^{n} IaCR(C_i, C_j) = 1$ for all the concepts $C_j (i \neq j)$. Note that

Table 3
The intra-concept coupling relation between C_1 and other concepts.

	C_1	C_2	C_3	C_4	C_5
C_1	1.000	0.210	0.222	0.119	0.070
	C_6	C_7	C_8	C_9	C_{10}
C_1	0.133	0.133	0.070	0.000	0.046

$IaCR(C_i, C_j) \neq IaCR(C_j, C_i)$, due to the dominators of $IaCR(C_i, C_j)$ and $IaCR(C_j, C_i)$ that capture the different relations with other concepts. For instance, given concept C_1 and parameter $\lambda = 0.5$, the intra-concept coupling relation between C_1 and the other concepts from Table 2 is obtained by considering the weights of the objects of the extents and the occurrence of the intents of concepts in Table 3.

The intra-concept coupling relation reflects the explicit semantic similarity between concepts, however, it ignores the implicit semantic information between concepts. Next, the topological distance between concepts based on the hierarchical structure of a concept lattice is used to define the implicit semantic information and specify the inter-concept coupling relation between them.

4.2. Inter-concept coupling relation

Concepts generated using FFCA are organized as a concept lattice by the partial order relation, i.e., inheritance relation in the concept lattice. The partial order relation reflects the implicit relatedness between concepts rather than considering the explicit relatedness between concepts by using the intra-concept coupling relation. The topological distance between concepts is used to characterize the partial order relation, where the topological distance is defined by the hierarchical distance between concepts. The smaller the distance between concepts, the more related the concepts.

Definition 5. Concepts $C_i = (\phi(E_i), I_i)$ and $C_j = (\phi(E_j), I_j)$ are interrelated if they are a parent–child relation (refer to the sub-concept and super-concept relation described in Section 2) in a concept lattice. The *implicit relation* between C_i and C_j is formalized as:

$$ImR(C_i, C_j) = \begin{cases} 0 & \text{otherwise} \\ \dfrac{1}{1+e^x} & \text{if } C_i \text{ and } C_j \text{ are inter-related} \end{cases} \tag{8}$$

where x represents the hierarchical distance between concepts, i.e., the linked edge number of the shortest path between C_i and C_j. For example, the hierarchical distance between C_1 and C_6 is 2 through the link "$C_1 \rightarrow C_3 \rightarrow C_6$", while the hierarchical distance between C_1 and C_9 is 0 since there are no parent–child relations between C_1 and C_9 in Fig. 1.

Similar to the intra-concept coupling relation, the inter-concept coupling relation is defined as a conditional probability manner by normalizing the implicit relation $ImR(C_i, C_j)$ between C_i and C_j with respect to the total number of relations between concept C_i and other concepts.

Definition 6. The *inter-concept coupling relation* between concepts $C_i = (\phi(E_i), I_i)$ and $C_j = (\phi(E_j), I_j)$ is defined as:

$$IrCR(C_i, C_j) = \begin{cases} 1 & i = j \\ \dfrac{ImR(C_i, C_j)}{\sum_{j=1, j\neq i}^{n} ImR(C_i, C_j)} & i \neq j \end{cases} \tag{9}$$

where n is the total number of all distinct concepts in the concept lattice. We determine that $IrCR(C_i, C_j) = 0$, when $\sum_{j=1, j\neq i}^{n} ImR(C_i, C_j) = 0$.

Note that $IrCR(C_i, C_j)$ falls in [0,1]. When C_i has no parent–child relation with any other distinct concept C_j, we determine that $IrCR(C_i, C_j) = 0$. The definition reflects that the smaller the topological distance between concepts, the more similar the concepts with respect to the underlying relation. For instance, given the concept C_1 from Table 2,

Table 4
The inter-concept coupling relation between C_1 and other concepts.

	C_1	C_2	C_3	C_4	C_5
C_1	1.000	0.326	0.326	0.326	0.000
	C_6	C_7	C_8	C_9	C_{10}
C_1	0.144	0.144	0.000	0.000	0.058

Table 5
The concept coupling relationship between C_1 and other concepts.

	C_1	C_2	C_3	C_4	C_5
C_1	1.000	0.268	0.274	0.223	0.035
	C_6	C_7	C_8	C_9	C_{10}
C_1	0.139	0.136	0.035	0.00	0.052

we obtain the inter-concept coupling relation between C_1 and other concepts in Table 4. The underlying (implicit) relation by the topological distance between concepts makes the related concepts more alike, which facilitates the document ranking for IR.

4.3. Concept coupling relationship

The concept coupling relationship captures the comprehensive semantic relation between concepts by aggregating the intra-concept coupling relation and inter-concept coupling relation. Based on Eqs. (7) and (9), the concept coupling relationship is defined as follows.

Definition 7. The *concept coupling relationship* between concepts $C_i = (\phi(E_i), I_i)$ and $C_j = (\phi(E_j), I_j)$ is defined as:

$$CCR(C_i, C_j) = \begin{cases} 1 & i = j \\ \beta * IaCR(C_i, C_j) + (1-\beta) * IrCR(C_i, C_j) & i \neq j \end{cases} \tag{10}$$

where $\beta \in [0, 1]$ is the parameter that determines the weight of the intra-concept coupling relation and inter-concept coupling relation, i.e., $IaCR(C_i, C_j)$ and $IrCR(C_i, C_j)$ respectively. The concept coupling relationship not only captures the explicit semantic relation by the intra-concept coupling relation, but also obtains the implicit semantic relation by inter-concept coupling relation. We observe that the higher the concept coupling relationship, the more related the concepts. For instance, given concept C_1 from Table 2 and parameter $\beta = 0.5$, the concept coupling relationship between C_1 and other concepts in Table 5 is obtained. Compared with the ranking concepts using NE and HE for query concept C_1, the concept coupling relationship has the better ability to distinguish the correlation between concepts. Each concept (excluding concept C_5 and C_8) obtains different scores to concept C_1 by the concept coupling relationship (CCR), while many concepts obtain the same scores as concept C_1 by NE and HE. The concept relation computed by CCR is more reasonable since CCR concerns the weight of the extent objects in the concepts.

5. Concept lattice-based retrieval system

In this section, a concept lattice-based retrieval system called coupled concept lattice-based retrieval (CCLR) which is based on concept coupling relationships is introduced. The system consists of three core parts: lattice generation, concept selection and document ranking with concept coupling relationship analysis. The details are introduced below.

5.1. Lattice generation

In this work, a concept lattice is generated by using FFCA as the classification technique for a document collection and consists of all the

concepts and relations between concepts, which are composed of the concept space for document ranking.

To obtain the concept lattice, a fuzzy context $K = (O, A, R = \phi(O \times A))$ as defined in Section 2.1 is obtained by a document-term matrix $D = (d_1, d_2, \ldots, d_i, \ldots, d_m)^{\mathrm{T}}$, where m is the number of documents and queries in the collection. Each document d_i in matrix D is regarded as an object $o_i \in O$ of context K, and each term t_j in matrix D is regarded as an attribute $a_j \in A$ of context K. The membership value $\mu(o_i, a_j)$ of an object $o_i \in O$ and an attribute $a_j \in A$, $(o_i, a_j) \in R$, is the weight w_{ij} of term t_j in document d_i, represented by the probability of term t_j occurring in document d_i, which is defined as follows:

$$\mu(o_i, a_j) = w_{ij} = \frac{tf(t_j)}{N_{d_i}} \tag{11}$$

where $tf(t_j)$ is the frequency of term t_j in document d_i, and N_{d_i} is the total number of terms in document d_i.

To achieve better retrieval performance, the most informative terms need to be selected as attributes of the fuzzy context using the signal-noise ratio as a measure to compute the weight of terms in the document collection. For a collection of n documents, noise N_{t_j} of term t_j is defined as:

$$N_{t_j} = \sum_{i=1}^{n} \frac{tf_{it_j}}{F_{t_j}} log \frac{F_{t_j}}{tf_{it_j}} \tag{12}$$

and signal S_{t_j} is:

$$S_{t_j} = log F_{t_j} - N_{t_j} \tag{13}$$

where F_{t_j} is the frequency of term t_j in the collection, tf_{it_j} is the frequency of term t_j in the ith document. The weight of term t_j in the ith document is $(tf_{it_j}/F_{t_j})(S_{t_j}/N_{t_j})$. For our approach, the top J terms in each document are chosen as the attributes of the fuzzy context. To obtain better retrieval results, the words in the queries are incorporated as the attributes of the fuzzy context.

Depending on the fuzzy context K in the above, a concept lattice can be constructed by using any lattice construction algorithm. Incremental lattice algorithms, adopted in the work of Carpineto (Carpineto and Romano, 2000), have better performance for lattice-based retrieval, however, the lattice structure computed by these methods is not immediately available. In this paper, the LATTICE algorithm (Lindig, 2000) is applied to compute both concepts and the explicit lattice structure, which are used for the intra- and inter-concept coupling relations between concepts.

5.2. Concept selection

In existing work, a concept space can be constructed using the set of concepts, which is obtained by a concept lattice algorithm. Based on the concept space and the generality/specificity relation of concepts, we can transform a query concept into each concept. Each document in the collection can be represented by a concept the intent of which is equal to the description of the document. The concept is defined as the important concept. Such a concept exists, and its extent contains at least that document. For instance, concept $C_3 = ((d_5(1/3), d_7(1/3)), (DM, ML))$ in Table 2 is important since the intent of C_3 is equal to the description of d_7 in Table 1. To reduce the computing complexity of document ranking and obtain high quality related documents, the important concepts need to be chosen from the set of concepts. A search algorithm is designed to traverse the concept lattice and obtain all important concepts from the concept lattice in Algorithm 1, where the *upper_neighbors* function of a node *temp* refers to the closest super-concepts of *temp*, implying that there is an edge between the node *temp* and the closest super-concepts of *temp*, and there is no other intermediate concept between both concepts in the lattice. For example, C_1, C_3, C_4, C_5, C_6, C_8 and C_9 are important concepts in Table 2.

Algorithm 1 Important concept search

Input: Lattice $H = \underline{B}(O, A, R)$
Output: Important concepts *result*

```
queue ← Queue()
queue.addQueue(H.infimum)
result = set()
label = set()
while queue.isNotEmpty() do
    temp ← queue.pop()
    label.add(temp)
    for obj in temp.extent do
        att = obj.description
        if att == temp.intent then
            result.add(temp)
            break
        end if
    end for
    for term in temp.upper_neighbors do
        if not label.has(term) then
            queue.addQueue(term)
        end if
    end for
end while
result.delete(H.infimum)
result.delete(H.supremum)
```

5.3. Document ranking

The documents in the collection can be represented by the important concepts selected in the above step. To rank these documents for a given query, we propose a novel method, namely concept coupling relationship analysis, to measure the similarity between concepts. The concept coupling relationship between concepts aggregates the intra- and inter-concept coupling relation between concepts. A similarity measure between the given query concept and other concepts can be generated using Eq. (10). In particular, when the two documents d_i and d_j obtain the same score as a result of the similarity of the corresponding concepts and the query concept, these documents are ranked based on the cosine similarity of the descriptive vector of documents. The cosine similarity between two documents is given as

$$sim(d_i, d_j) = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2} \times \sqrt{\sum_{k=1}^{n} w_{jk}^2}} \tag{14}$$

where n is the number of keywords and w_{ik} and w_{jk} are the weight of the kth keyword of document d_i and d_j, respectively. For instance, documents d_5, d_7, d_6, d_1, d_4, d_2 and d_3 can be respectively represented by the important concepts C_1, C_3, C_4, C_5, C_6, C_8 and C_9 in Table 2. Given the query "data mining, machine learning, text mining", corresponding to the query concept C_1, we can obtain a document ranking list, i.e., d_7, d_6, d_4, d_1, d_2 and d_3 for the query based on the concept coupling relationship between C_1 and other concepts in Table 5.

6. Experiments and evaluation

6.1. Experimental settings

Experiments are conducted on data collections from the SMART Information Retrieval System:[1] CACM (3204 document abstracts extracted from the Association of Computing Machinery, 45 queries), CISI (1460 document abstracts in library science and related areas extracted from Social Science Citation Index by Institute for Scientific Information,

[1] http://ir.dcs.gla.ac.uk/resources/test_collections/.

Table 6
Data and lattice description. Note: object number (ON), attribute number (AN), threshold (TS), the density of the context (DC [%]), concept number (CN), time spent on lattice construction (TLC, in minutes).

	ON	AN	TS	DC	CN	TLC
CACM	3204	4075	0.03	0.18	24 945	2492
CISI	1460	4604	0.03	0.17	11 078	192
CRAN	1400	2787	0.03	0.30	13 355	145
MED	1033	4770	0.02	0.34	21 594	171

35 queries), CRAN (1400 document abstracts extracted from publications of aeronautic reviews, 35 queries), MED (1033 document abstracts extracted from the National Library of Medicine, 30 queries), where queries are keyword sequences in text.

Textual data is first pre-processed for queries and documents by text segmentation, word stemming and removing stop words. The parameters in our experiments are listed in Table 6. We set $\lambda = 0.5$ and $\beta = 0.5$ for the framework. Lattice constructions are undertaken once offline. All experiments are conducted on a Think Centre M6400T desktop computer with Intel Core i5 CPU and 4G RAM.

In our approach, a query is regarded as an object in the fuzzy context. A corresponding query concept in the lattice is obtained for the query when the description of the query is equal to or closest to the intent of the concept. To compare the results of our CCLR approach, we have implemented four classic retrieval methods, namely lattice-based ranking using the HE (LRHE) strategy (Messai et al., 2010), lattice-based ranking using the NE (LRNE) strategy (Carpineto and Romano, 2000), the query likelihood (QL) model, and the exact matching (EM) method. LRHE and LRNE are classic lattice-based retrieval methods. The QL model is a language model constructed for each document in the collection and ranks each document by the probability of specific documents given a query. The EM method is a Boolean retrieval which searches the collection for documents with at least one keyword provided in a query.

6.2. Evaluation measures

Precision and recall are measures used to assess the relevance of documents returned by a retrieval system. They are defined as follows:

$$precision = \frac{|positive|}{|retrieved|} \tag{15}$$

$$recall = \frac{|positive|}{|relevant|} \tag{16}$$

where $relevant$ represents the document set in which documents are relevant to a query, $retrieved$ represents the document set in which documents are retrieved by retrieval systems, and $positive$ is defined as $relevant \bigcap retrieved$. Precision is a measure of exactness or quality, whereas recall is a measure of completeness or quantity. Literatures show that a good retrieval system should have a good quality/quantity balance of the answers, where precision and recall are considered as a trade-off between quality and quantity of related documents in document retrieval (Codocedo et al., 2014). To achieve balance between quality and quantity in our work, we consider the coupled concept relation for document ranking in a concept space, where the common objects and attributes of concepts are used to capture the explicit concept relation and the partial order relation of concepts is used to capture the implicit concept relation.

To evaluate CCLR with other systems, more robust evaluation measures are needed, i.e., *eleven-point interpolated average precision* ($IAP@11$) and *mean average precision* (MAP), since precision and recall only consider unordered documents in the answer of a retrieval system. To formalize the measures, $L = \{d_1, d_2, \ldots, d_n\}$ denotes the retrieval results ranked by a retrieval system for a query, and $L_j \in L$ denotes the sub-list of L which contains from $d_1 \in L$ to $d_j \in L$. Interpolated precision (ip) is calculated in a given interval defined by the edges r_1 and r_2 as follows:

$$ip(r_1, r_2) = argmax_{\forall d_j \in L}\{precision(L_j) \Leftrightarrow recall(L_j)\} \tag{17}$$

where $recall(L_j) \in [r_1, r_2]$. Then $IAP@11$ and MAP are defined as follows:

$$IAP@11 = \frac{\sum_{i=0}^{10} ip(\frac{i}{10}, \frac{i+1}{10})}{11} \tag{18}$$

$$MAP = \frac{\sum_{\forall j \in |positive|} precision(L_j)}{|positive|}. \tag{19}$$

6.3. Experimental results

For all data collections, we compare LRHE, LRNE, EM and QL with our CCLR approach by *mean average precision* (MAP) in Table 7, *eleven-point interpolated average precision* (IAP@11) in Table 8 and *precision in the first 10 ranked documents* (P@10) in Table 9. We observe that LRNE augments the performance of LRHE on all the datasets, excluding P@10 on the CISI collection. LRNE achieves around 51%, 41% and 25% improvement in terms of average MAP, IAP@11 and P@10 respectively. LRNE achieves better performance over LRHE, since LRNE can obtain more important concepts to represent documents than LRHE. This demonstrates that the number of important concepts is critical for document ranking in classic concept lattice-based retrieval methods.

Compared to LRHE and LRNE, CCLR further improves the performance over LRHE and LRNE and achieves the best scores on all the datasets. Compared to LRHE, CCLR achieves 63%, 50% and 44% improvement in terms of average MAP, IAP@11 and P@10 respectively. Compared to LRNE, CCLR achieves 9%, 8% and 15% improvement in terms of average MAP, IAP@11 and P@10 respectively. These results indicate that CCLR achieves better performance than classic lattice-based retrieval methods. First, CCLR has the advantage of LRNE, which can obtain more important concepts to represent documents. Second, CCLR can capture the explicit and implicit semantic information between concepts through concept coupling relationship analysis, while LRHE and LRNE can only obtain the implicit semantic information between concepts by the partial order relation of the concept lattice.

Compared to QL and EM, CCLR surpasses EM with a score of CCLR: 7, EM: 6 (the tie for P@10 in the MED dataset is considered as a point for CCLR and EM), and exceeds QL with a score of CCLR:8, QL:4 in terms of MAP, IAP@11 and P@10. This demonstrates that CCLR obtains relatively better performance than EM and QL on all the datasets. First, CCLR employs formal concepts to represent documents in a concept space, while QL and EM apply the vector space model to represent document in a term space. Second, CCLR ranks documents for a query using the intra- and inter-concept coupling relation. The intra-concept coupling relation employs common attributes and objects to obtain the explicit semantic information between concepts and the inter-concept coupling relation applies the partial order relation of the concept lattice to capture the implicit semantic information between concepts. QL and EM can only employ the common terms between documents and a query to capture the explicit semantic information between documents and the query. Rather than the above reasons, the characteristic of each collection may affect the performance of each method. We observe that CCLR obtains the best performance on the CACM collection, QL obtains the best performance on the CISI collection and EM obtains the best performance on the CRAN and MED collections. The results show that the CACM dataset may contain fewer common terms between documents and queries than the other three datasets, namely the CISI, CRAN and MED datasets.

To further prove the performance of CCLR, experiments are conducted with the interpolated precision at 11 different recall levels on the four collections CACM, CISI, CRAN and MED. The experiment results are shown in Fig. 3. In practice, ordinary users want to find related documents in the first web pages, not to find all related documents.

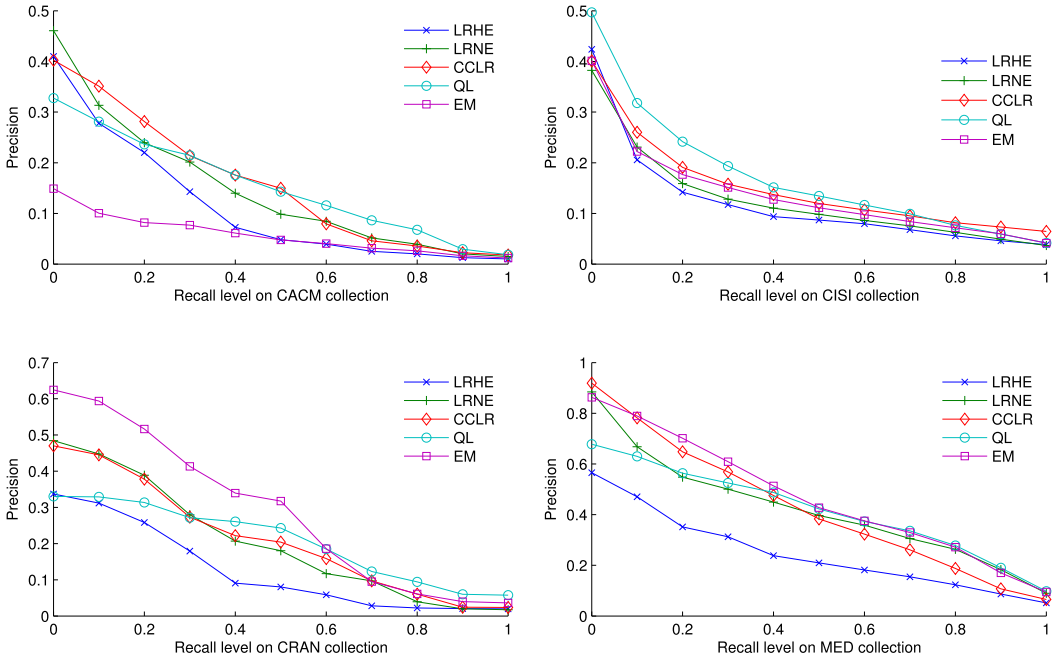

Fig. 3. The interpolated precision at 11 different recall levels on four collections.

Table 7
MAP comparison between our approach (CCLR) and the other models (LRHE, LRNE, EM, QL).

	LRHE	*LRNE*	*EM*	*QL*	*CCLR*
CACM	0.097	0.135	0.051	0.139	0.143
CISI	0.099	0.113	0.124	0.156	0.137
CRAN	0.111	0.186	0.270	0.185	0.198
MED	0.216	0.399	0.451	0.403	0.410

Table 8
IAP@11 comparison between our approach (CCLR) and the other models (LRHE, LRNE, EM, QL).

	LRHE	*LRNE*	*EM*	*QL*	*CCLR*
CACM	0.116	0.151	0.058	0.154	0.161
CISI	0.123	0.129	0.140	0.175	0.153
CRAN	0.128	0.207	0.293	0.206	0.214
MED	0.250	0.422	0.467	0.416	0.429

Table 9
P@10 comparison between our approach (CCLR) and the other models (LRHE, LRNE, EM, QL).

	LRHE	*LRNE*	*EM*	*QL*	*CCLR*
CACM	0.144	0.180	0.064	0.186	0.204
CISI	0.189	0.180	0.188	0.251	0.203
CRAN	0.123	0.168	0.214	0.200	0.191
MED	0.323	0.463	0.550	0.476	0.550

Thus, we focus on the low recall points, such as the 0.1, 0.2, 0.3, 0.4 recall point, at which there are a subset of related documents returned by a retrieval system. We observe that CCLR achieves stable good precision on all the collections at the low recall points, while the performance of EM and QL, fluctuates at the low recall points on all the collections. For instance, at the low recall points in Fig. 3, CCLR obtains the third best performance on the CRAN collection and the second best performance on the other collections; QL obtains the best performance on the CISI collection and the fourth best performance on the other collections; EM obtains the best performance on the CRAN collection and the worst performance on the CACM collection. These results show that CCLR obtains high precision at the low recall points, though it does not always obtain the best performance compared to the other approaches.

In all, CCLR can obtain better and more stable performance because concept coupling relationship analysis can capture more explicit and implicit semantic information between concepts to rank documents in CCLR.

6.4. Sensitivity of thresholds

In this section, we analyze the effect of threshold selection on retrieval performance, i.e., retrieval efficiency and effectiveness, for CCLR on the CISI collection. The threshold values are set as 0.025, 0.030, 0.035, 0.040, 0.045 and 0.050, and we fix all the other parameters of CCLR. The results are shown in Fig. 4.

The time complexity of CCLR is $O(|M| \times |Doc|^2 \times |V|)$, where $|.|$ is the cardinality of the set, M includes all the concepts of the lattice, Doc refers to the documents in the collection, and V is the vocabulary dictionary of the collection. When the threshold becomes smaller, the corresponding lattice size M in the same context grows exponentially, which has been proved by Lin (Lindig, 2000). Hence, the retrieval efficiency (refers to the time complexity) of CCLR grows exponentially.

CCLR achieves two peaks with MAP, IAP@11 and P@10 respectively when the threshold is in the interval [0.025, 0.05]. Meanwhile, CCLR achieves the best performance at 0.03. This shows that CCLR may obtain better performance using the smaller threshold. However, the retrieval efficiency will be larger using the smaller threshold. To balance retrieval effectiveness and efficiency, the threshold of CCLR is set as 0.03 on the CISI collection in our experiments.

6.5. Query analysis

In this section, a brief analysis of some queries of the CISI collection is presented in Tables 10 and 11 to illustrate how CCLR obtains better performance and further improvements to the approach are described. In Tables 10 and 11, the top 10 returned documents (Doc), the corresponding concepts ($Intent$ and $Extent\ support$) and the corresponding similarity (Sim) between concepts are presented for Query 6 and 7. A document is represented by a concept, where the concept intent is equal to the descriptions of the document and the concept extent contains the document.

Query 6, as the best case, is represented by the query concept with the intent *word, human, communication, verbal, possibility, computer* and

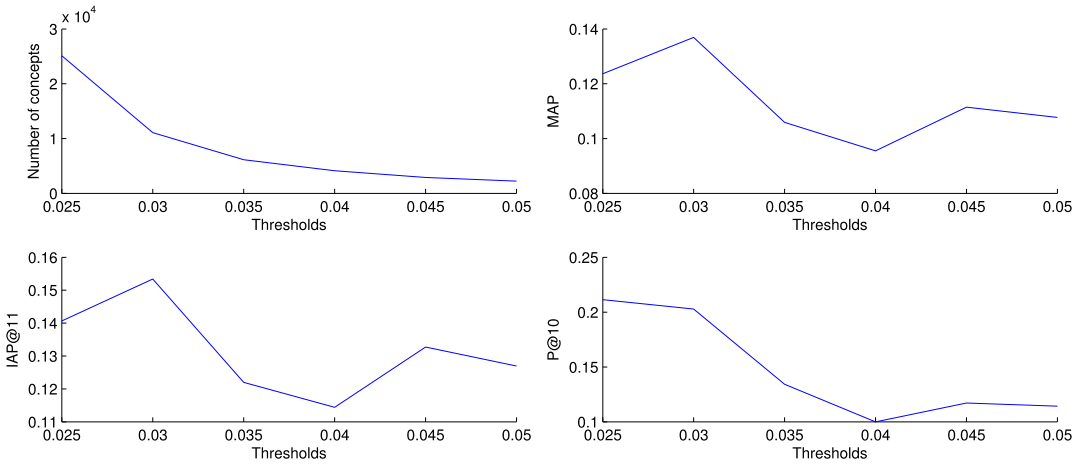

Fig. 4. The effect of threshold selection on CISI collection.

Table 10
Query analysis for query 6 in the CISI collection.

Doc	*Intent*	*Extent support*	*Sim*
query 6	Word, human, communication, verbal, possibility, computer	1	1
doc 400	Telephone, computer	1	0.0052
doc 967	Communication, information	28	0.0052
doc 1357	Feedback, communication, improvement, investigation	1	0.0042
doc 602	Network, communication, information, scientist	1	0.0042
doc 156	Result, information, search, computer	1	0.0042
doc 398	Communication, information, channel, formal	1	0.0042
doc 694	Generation, program, develop, computer	1	0.0042
doc 529	Provide, retrospect, search, computer	1	0.0042
doc 680	Word,program, full, swift	1	0.0042
doc 228	Communication, theory, selection, message	1	0.0042

Table 11
Query analysis for query 7 in the CISI collection.

Doc	*Intent*	*Extent support*	*Sim*
query 7	Describe, work, original, save, form, data, article, paper, publish, print, computer, plan, byproduct, code, retrieve, process, system, present	1	1
doc 1349	Work	55	0.1794
doc 357	Data	102	0.1794
doc 457	Paper	51	0.1794
doc 1004	Describe, standard, Henriette, bibliography, form, international, readable, avaram, remain, catalogue, machine, interchange, system, progress, discuss, record, problem, work, paper,	1	0.0018
doc 689	Grease, input, code, retrieval, system, integration	1	0.0017
doc 917	Work,field, process, system, library	1	0.0017
doc 703	Inform, chemic, small, search, computer, retrieval, system	1	0.0016
doc 1136	Specific, data, inform, retrieval , system, discuss, problem	1	0.0016
doc 1307	Original, easier, extent, make, measure, inform, term, review, contrast, propos, stew, compare, retrieval, common, system, present	1	0.0016
doc 1207	Technical, project, inform, computer, facility, system	1	0.0015

has only one relevant document in the CISI collection. Through concept coupling relationship analysis, doc 400 achieves the highest similarity to the query, where doc 400 is mapped into the concept with the intent *telephone, computer*. Though the concept with the intent *communication, information* has the same score as the concept with the intent *telephone, computer*, doc 400 has a higher cousin similarity with the query than doc 976. In this case, a query modification can be obtained by refining the terms *word, human, communication, verbal, possibility* with the term *telephone*. Thus, CCLR is able to find the unique relevant document by concept coupling relationship analysis, demonstrating the capabilities of our approach.

Query 7, as the inferior case, is represented by the query concept with the intent *describe, work, original* et al. and has eight relevant documents in the CISI collection. However, no relevant document occurs in the top 10 document set for query 7. This problem may be due to the fact that there are less common attributes between concept intents and the attributes of concept intents are equally important. To overcome this issue, a semantic measure, such as the information content approach proposed by Formica (2008), and the weight model, such as the TF-IDF model, of terms need to be adopted to improve the approach.

6.6. Scalability issue

Our experiments are conducted on the smaller datasets due to the limitation of the computation of a concept lattice. With the state-of-the-art FCA algorithms, it is feasible to apply CCLR for smaller datasets, such as personal book or email collections. It is difficult to apply CCLR directly to large-scale data, such as the entire Web and TREC collections. For large-scale data, lattice-based retrieval, such as CCLR, can be used to refine and represent the top results returned by an effective retrieval algorithm from language models or probability models. This procedure may be more efficient than constructing the lattice of the entire document collection to rank documents.

7. Related work

Concept lattices based on formal concept analysis (FCA) (Ganter and Wille, 1999) are important techniques for representing the conceptual hierarchies used in information retrieval (Priss, 2000; Carpineto and Romano, 2005; Poelmans et al., 2013; Codocedo and Napoli, 2015). They integrate the discovery, dependencies and reasoning with general/specific relationships between concepts. For instance, an algorithm mining association rule between user query keywords (UQWs) and non-user query keywords from the concept lattice of the low-adjacence set, defined as the webpages that include UQWs, has been proposed to obtain the semantic information between user queries and web pages (Du and Li, 2010), and a lattice-based approach for a mathematical search has been proposed in which math expressions are converted into the corresponding MathML representation (Nguyen et al., 2012), similar to the work of Peng Tang, which employs lattices for chemical structural retrieval (Tang et al., 2015). FCA provides the capabilities of query representation and transformation, document browsing, visualization, and navigation in the standard IR models. For example, CREDO (Carpineto et al., 2004) proposed by Carpineto et al. is a system that allows the users to query Web documents and see the retrieval results organized in a browsable concept lattice. Different to the work of Carpineto, there is other work on FCA-based retrieval systems that embed the users in the IR process to improve the performance of the systems. FooCA employs search engine retrievals to represent a context, and establishes a concept lattice to visualize the data with the refining context based on the search strategies and the preferences users have chosen (Koester, 2006). CreChainDo adopts the user feedback approach to lattice navigation in which the feedback is converted into a reduction or an extension of the context of the lattice (Nauer and Toussaint, 2009).

A number of approaches employ different navigation strategies based on concept lattices to find the related documents for a given user query. In general, these approaches first search for a query concept that best represents the user query in the given concept lattice. The neighborhood expansion strategy (Carpineto and Romano, 2000) searches concepts in an "expanding ring" order from the query concept, where there may be super- and sub-concepts of the query concept in the ring. The hierarchical exploration strategy (Messai et al., 2005) searches concepts by exploring the super-concepts of the query concept. The above methods regard document ranking as query-document transformation driven by conceptual representations of the whole document collection to obtain the implicit semantic information between documents, rather than considering the explicit similarity metrics between documents or concepts. A concept lattice exploration strategy based on the notion of "cousin concepts" (Codocedo et al., 2014) has been proposed to obtain the explicit semantic information between documents. The strategy ranks documents based on the semantic measure, which consists of the occurrence of the concept extents and the semantic information of the concept intents using an external lexical hierarchy (Formica, 2008). The method ignores the implicit semantic information between documents. In our work, we leverage the intra- and inter-concept coupling relation to obtain the explicit and implicit semantic information between documents. An exploring strategy has been proposed in this paper to compute the inter-concept coupling relation between concepts by the super- and sub-concepts of the query concept in the hierarchical structure of lattices.

Apart from the above navigation strategies of concept lattices, fuzzy concept lattices based on FFCA, incorporating fuzzy logic into FCA, can be adopted to model the uncertain, vague and implicit information in queries and documents (Georgescu and Popescu, 2002). For example, rough set theory is employed in combination with FFCA to perform a Semantic Web search and to discover information on the Web, and FFCA is used to support the construction of formal ontologies in the presence of uncertain data for the development of the Semantic Web (Formica, 2012). The fuzzy extension of FFCA, as a mathematical model, is exploited to automatically build ontologies, which is regarded as a formal and reusable model for the knowledge representation (De Maio et al., 2012). Regarding the similarity between concepts based on FFCA, Formica proposed to combine the similarity of concept extents (fuzzy sets) and concept intents (Formica, 2010, 2013). In particular, concept intents are compared according to the information content approach. Similar to the work of Formica, we employ FFCA in this paper to model the uncertain information of queries and documents, and integrate the similarity of concept extents (fuzzy sets) and concept intents to compute the explicit semantic information between concepts, i.e., the intra-concept coupling relation.

The coupling relationship has been proposed recently to represent the complex relation between terms, which consists of explicit and implicit relationships. The coupled object similarity measure, consisting of both attribute value frequency distribution (intra-coupling) and feature dependency aggregation (inter-coupling), has been proposed to measure attribute value similarity for unsupervised learning of nominal data (Wang et al., 2011). Coupled term-term relation analysis has been designed for clustering by integrating the intra-relation (i.e. co-occurrence of terms) and inter-relation (i.e. dependency of terms via link terms) between a pair of terms (Cheng et al., 2013). Keyword and query coupling relationship analysis has been developed to select semantically related queries based on intra- and inter-keyword couplings (Meng et al., 2014). Similar to the above work, we propose concept coupling relationship analysis to capture the intra-concept coupling relation (explicit similarity) and inter-concept coupling relation (implicit similarity) between concepts. The intra-concept coupling relation is represented by the similarity of fuzzy concepts, and the inter-concept coupling relation is represented by the topological distance between fuzzy concepts. In this way, concept coupling relationship analysis can be used to rank documents in a concept space.

8. Conclusions and future work

To address the challenges of learning semantic information in query and document representation for information retrieval, especially uncertain, vague and fuzzy semantic information, this work proposes a concept lattice-based retrieval framework based on concept coupling relationship analysis. The proposed framework employs formal concepts extracted by fuzzy formal concept analysis to represent queries and documents in a concept space. Then concept coupling relationship analysis, consisting of the intra- and inter-concept coupling relation, is applied to rank documents represented by formal concepts. The intra-concept coupling relation is computed by the common objects and attributes of concepts and used to capture the explicit semantic information between concepts. The inter-concept coupling relation is calculated by the partial order relation of concepts and used to capture the implicit semantic information between concepts. Substantial experiments are conducted on four classical datasets used in information retrieval which show that our approach significantly outperforms the baselines. We discuss the importance of balancing effectiveness and efficiency in the threshold selection and provide a method to address the deficiencies of lattice-based retrieval, whereby lattice-based retrieval refines the top results returned by other retrieval models.

Our further work will focus on investigating (i) the effect of introducing weighted models to enrich the fuzzy context and rationally represent documents, (ii) the improvement of lattice construction using parallel algorithms, and (iii) how to combine the proposed method with other models, such as linguistic models, for semantic retrieval.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61502033, 61472034, 61772071, 61370137, 61272169, 61672098) and JiLin Natural Science Foundation (No. 20160101251JC).

References

Bělohlávek, R., Sklenář, V., Zacpal, J., 2005. Crisply generated fuzzy concepts. In: Proceedings of the 3rd International Conference of Formal Concept Analysis, Lens, France, pp. 269–284.

Bollegala, D., Ishizuka, M., Matsuo, Y., 2007. Measuring semantic similarity between words using web search engines. In: Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, pp. 757–766.

Cao, L., 2015. Coupling learning of complex interactions. Inf. Process. Manage. 51 (2), 167–186.

Carpineto, C., Romano, G., 2000. Order-theoretical ranking. J. Am. Soc. Inf. Sci. 51, 587–601.

Carpineto, C., Romano, G., 2005. Using concept lattices for text retrieval and mining. In: Formal Concept Analysis: Foundations and Applications. Springer Berlin Heidelberg, pp. 161–179.

Carpineto, C., Romano, G., 2012. A survey of automatic query expansion in information retrieval. ACM Comput. Surv. 44, 159–170.

Carpineto, C., Romano, G., Bordoni, F.U., 2004. Exploiting the potential of concept lattices for information retrieval with CREDO. J. UCS 10, 985–1013.

Cheng, X., Miao, D., Wang, C., Cao, L., 2013. Coupled term-term relation analysis for document clustering. In: Proceedings of the International Joint Conference on Neural Networks, Dallas, TX, USA, pp. 1–8.

Codocedo, V., Lykourentzou, I., Napoli, A., 2014. A semantic approach to concept lattice-based information retrieval. Ann. Math. Artif. Intell. 72, 169–195.

Codocedo, V., Napoli, A., 2015. Formal concept analysis and information retrieval–a survey. In: Proceedings of 13rd International Conference of Formal Concept Analysis, Nerja, Spain, pp. 61–77.

De Maio, C., Fenza, G., Loia, V., Senatore, S., 2012. Hierarchical web resources retrieval by exploiting fuzzy formal concept analysis. Inf. Process. Manage. 48, 399–418.

Du, Y., Li, H., 2010. Strategy for mining association rules for web pages based on formal concept analysis. Appl. Soft Comput. 10, 772–783.

Formica, A., 2008. Concept similarity in formal concept analysis: An information content approach. Knowl.-Based Syst. 21, 80–87.

Formica, A., 2010. Concept similarity in fuzzy formal concept analysis for semantic web. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 18, 153–167.

Formica, A., 2012. Semantic web search based on rough sets and fuzzy formal concept analysis. Knowl.-Based Syst. 26, 40–47.

Formica, A., 2013. Similarity reasoning for the semantic web based on fuzzy concept lattices: An informal approach. Inf. Syst. Front. 15, 511–520.

Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T., 1987. The vocabulary problem in human-system communication. Commun. ACM 30, 964–971.

Ganter, B., Wille, R., 1999. Formal Concept Analysis: Mathematical Foundations, first ed.. Springer-Verlag New York, Inc., Secaucus.

Georgescu, G., Popescu, A., 2002. Concept lattices and similarity in non-commutative fuzzy logic. Fund. Inform. 53, 23–54.

Koester, B., 2006. Conceptual knowledge retrieval with fooca: Improving web search engine results with contexts and concept hierarchies. In: Proceedings of the 6th Industrial Conference on Data Mining, Leipzig, Germany, pp. 176–190.

Kumar, C.A., Mouliswaran, S.C., Amriteya, P., Arun, S., 2015. Fuzzy formal concept analysis approach for information retrieval. In: Proceedings of the 5th International Conference on Fuzzy and Neuro Computing, Hyderabad, India, pp. 255–271.

Li, F., Xu, G., Cao, L., Fan, X., Niu, Z., 2013. CGMF: Coupled group-based matrix factorization for recommender system. In: Proceedings of the 14th International Conference on Web Information Systems Engineering, Nanjing, China, pp. 189–198.

Lindig, C., 2000. Fast concept analysis. Working with Conceptual Structures-Contributions To ICCS 2000, pp. 152–161.

Liu, C., Cao, L., Yu, P.S., 2014. Coupled fuzzy k-nearest neighbors classification of imbalanced non-IID categorical data. In: Proceedings of the International Joint Conference on Neural Networks, Beijing, China, pp. 1122–1129.

Meng, X., Shao, J., et al., 2014. Semantic approximate keyword query based on keyword and query coupling relationship analysis. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, Shanghai, China, pp. 529–538.

Messai, N., Devignes, M.-D., Napoli, A., Smaïl-Tabbone, M., 2005. Querying a bioinformatic data sources registry with concept lattices. In: Proceedings of 13rd International Conference on Conceptual Structures: Common Semantics for Sharing Knowledge, Kassel, Germany, pp. 323–336.

Messai, N., Devignes, M.-D., Napoli, A., Smaïl-Tabbone, M., 2010. Using domain knowledge to guide lattice-based complex data exploration. In: Proceedings of 19th European Conference on Artificial Intelligence, Lisbon, Portugal, pp. 847–852.

Nauer, E., Toussaint, Y., 2009. CreChainDo: An iterative and interactive Web information retrieval system based on lattices. Int. J. Gen. Syst. 38, 363–378.

Nguyen, T.T., Hui, S.C., Chang, K., 2012. A lattice-based approach for mathematical search using formal concept analysis. Expert Syst. Appl. 39, 5820–5828.

Pang, G., Cao, L., Chen, L., 2016. Outlier Detection in Complex Categorical Data by Modeling the Feature Value Couplings. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, USA, pp. 1902–1908.

Poelmans, J., Ignatov, D.I., Kuznetsov, S.O., Dedene, G., 2013. Formal concept analysis in knowledge processing: A survey on applications. Expert Syst. Appl. 40, 6538–6560.

Poelmans, J., Ignatov, D.I., Kuznetsov, S.O., Dedene, G., 2014. Fuzzy and rough formal concept analysis: A survey. Int. J. Gen. Syst. 43, 105–134.

Priss, U., 2000. Lattice-based information retrieval. Knowl. Organ. 27, 132–142.

Tang, P., Hui, S.C., Fong, A.C., 2015. A lattice-based approach for chemical structural retrieval. Eng. Appl. Artif. Intell. 39, 215–222.

Wang, C., Cao, L., Wang, M., Li, J., Wei, W., Ou, Y., 2011. Coupled nominal similarity in unsupervised learning. In: Proceedings of the 20th ACM Conference on Information and Knowledge Management, Glasgow, United Kingdom, pp. 973–978.