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 $\mu_o$ 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.