Information Processing and Management 54 (2018) 116–128

Contents lists available at ScienceDirect

Information Processing and Management

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

Knowledge based collection selection for distributed information retrieval

Han Baoli, Chen Ling*, Tian Xiaoxue

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Keywords:
Collection selection
Distributed information retrieval
Knowledge base
Query expansion

ABSTRACT

Recent years have seen a great deal of work on collection selection. Most collection selection methods use *central sample index* (CSI) that consists of some documents sampled from each collection as collection description. The limitations of these methods are the usage of 'flat' meaning representations that ignore structure and relationships among words in CSI, and the calculation of query-collection similarity metric that ignore semantic distance between query words and indexed words. In this paper, we propose a knowledge based collection selection method (KBCS) to improve collection representation and query-collection similarity metric. KBCS models a collection as a weighted entity set and applies a novel query-collection similarity metric to select highly scored collections. Specifically, in the part of collection representation, context- and structure-based measures are employed to weight the semantic distance between two entities extracted from the sampled documents of a collection. In addition, the novel query-collection similarity metric takes the entity weight, collection size, and other factors into account. To enrich concepts contained in a query, DBpedia based query expansion is integrated. Finally, extensive experiments were conducted on a large webpage dataset, and DBpedia was chosen as the graph knowledge base. Experimental results demonstrate the effectiveness of KBCS.

1. Introduction

Distributed Information Retrieval (DIR), also known as Federated Search (FS) or Federated IR (FIR), concerns with aggregating multiple searchable sources of information under a single interface (Crestani & Markov, 2013). DIR consists of four main phases: collection (server/resource) description, collection selection, results merging, and results presentation. Given a query and a set of collection descriptions, collection selection ranks available collections based on their computed scores, then determines which collections to search (Callan, 2002). In a specific search circumstance, users are often interested in top-ranked search results. However, not all collections contain information that users need. If search engine only retrieve a small number of collections and get a similar effect to retrieve all collections, it would significantly enhance the efficiency of retrieval system. Collection selection plays an important role in reducing computational overhead and improving retrieval efficiency.

Recent years have seen a great deal of work on collection selection, which can be divided according to the mechanism to describe a collection: dictionary-based methods (Aly, Hiemstra, & Demeester, 2013, Callan, Lu, & Croft, 1995, Gravano & Garcia-Molina, 1995, Xu & Croft, 1999, Yuwono & Lee, 1997) and sampling-based methods (Baillie, Carman, & Crestani, 2011, Kulkarni, Tigelaar, Hiemstra, & Callan, 2012, Mendoza, Marín, Gil-Costa, & Ferrarotti, 2016, Paltoglou, Salampasis, & Satratzemi, 2011, Shokouhi, 2007, Shokouhi, Zobel, Tahaghoghi, & Scholer, 2007, Si & Callan, 2003, Thomas & Shokouhi, 2009, Wauer, Schuster, & Schill, 2011).

* Corresponding author.
E-mail addresses: litutor@zju.edu.cn (B. Han), lingchen@cs.zju.edu.cn (L. Chen), xxtian@zju.edu.cn (X. Tian).

http://dx.doi.org/10.1016/j.ipm.2017.10.002
Received 1 April 2017; Received in revised form 28 August 2017; Accepted 11 October 2017

Dictionary-based methods use the word statistics of all documents as collection description, and then exploit a scoring function to reflect the similarity between a collection and a query. However, it is unfeasible to acquire the word statistics of all collections in an uncollaborative distributed information retrieval environment. Another problem is that the scoring function based on word statistics loses a large amount of semantic information in calculating collection score, e.g., synonym, polysemy, and the order of words. These methods also have a low effectiveness in the environment of skewed collection sizes.

To overcome the limitation of usage in uncollaborative distributed information retrieval environment, sampling-based methods use some documents that are sampled from a collection as collection description. With the development of text representation technology, some sampling-based methods (Callan & Connell, 2001) begin to exploit ESA (Gabrilovich & Markovitch, 2007) and LDA (Blei, Ng, & Jordan, 2003) to represent collection description. Compared with dictionary-based methods, sampling-based methods can find entities or topics that reflect the semantic information of a collection and have higher accuracy of collection representation. However, they totally ignore the structure and relationships among entities or topics.

The growing popularity of graph knowledge bases, e.g., DBpedia (Bizer et al., 2009), Yago (Hoffart, Suchanek, Berberich, & Weikum, 2013), and Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008), has inspired the emergency of new text representation methods. To represent documents with more fine-grained semantic information, knowledge based text representation methods (Ni et al., 2016, Schuhmacher & Ponzetto, 2014) exploit the relational knowledge and network structure encoded in graph knowledge bases. These methods have achieved better performance than ESA and LDA.

Query expansion is also important for collection selection, which can help reduce the vocabulary mismatch problem by adding related terms to a query and find potential related documents and collections (Carpineto & Romano, 2012). If a collection is modeled as a weighted entity set, without query expansion, it would be hard to compute its score for a query containing few entities. There is a vast amount of research work on knowledge based query expansion (Arguello, Elsas, Callan, & Carbonell, 2008, Dang & Croft, 2010, Eiron & McCurley, 2003, Guisado-Gámez, Dominguez-Sal, & Larriba-Pey, 2014), which can obtain expansion terms from external knowledge sources, e.g., Wikipedia and DBpedia. Arguello et al. (2008), Dang and Croft (2010) and Eiron and McCurley (2003) only focus on adding entities directly connected to original query entities in a graph knowledge base. Due to the limitation in connected entities, these entities that are actually related to original query entities but not connected in a knowledge base are excluded in expansion terms. Guisado-Gámez et al. (2014) not only find the semantic relevant terms but also the connected entities in a knowledge base, and thus has a better performance.

As mentioned above, collection representation, query-collection similarity metric, and query expansion are three key points in collection selection research. In this paper, we propose *knowledge based collection selection* (KBCS), which covers the three key points. KBCS explicitly models a collection as a weighted entity set. Entities are firstly extracted from the sampled documents of a collection and the semantic distance of each pair of entities is weighted by employing semantic relation measures, including context- and structure-based measures based on knowledge bases. Then, to identify the important entities, entities are assigned with different weights by calculating semantic distances between an entity and all the other entities in a collection. In addition, we integrate a query expansion method (Arguello et al., 2008) to find more related entities. Specifically, to distinguish entities in a query and avoid query drift, the way to weight query entities is adjusted in our method. Finally, we present a novel query-collection similarity metric based on the knowledge representation that takes sampling factor (the proportion of sampled documents in a collection), collection entity frequency, collection entity weight, and query entity weight into account.

To summarize, our contributions are as follows:

1. Propose KBCS, which exploits the relational knowledge and network structure encoded in knowledge bases to solve the ‘flat’ meaning representation problem, and integrates a query expansion method to solve the vocabulary mismatch problem.
2. Present a query-collection similarity metric, which aggregates both context- and structure-based semantic distances between query entities and collection entities. To improve its precision, the sampling factor, collection entity frequency, collection entity weight, and query entity weight are incorporated into the calculation formula.
3. Conduct extensive experiments to verify the effectiveness of KBCS. On one hand, evaluation on KBCS and baselines has demonstrated that KBCS can get a high precision while chooses few collections. On the other hand, evaluation on query expansion has demonstrated that enriching entities found in query terms can significantly improve KBCS's precision.

The rest of this paper is structured as follows: the next section describes related work on collection selection. Section 3 firstly gives some definitions about collection selection and a brief introduction to KBCS's architecture, and then elaborates on KBCS's algorithm. Section 4 describes experimental details and discusses experimental results. We conclude our work in Section 6.

2. Related work

There has been considerable research on collection selection. Collection selection algorithms can be divided into three classes: dictionary-based methods, sampling-based methods, and classification-based methods which combine the above methods and a number of other query- and corpus-based features in a machine learning framework. In this paper, we focus on dictionary- and sampling-based methods, which are most related prior research work.

2.1. Dictionary-based methods

Dictionary-based collection selection methods use the word statistics of all documents in a collection as collection description.