Information Processing and Management 50 (2014) 132–155

Contents lists available at ScienceDirect

Information Processing and Management

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

Effective ranking and search techniques for Web resources considering semantic relationships ☆

Jihyun Lee [a], Jun-Ki Min [b], Alice Oh [a], Chin-Wan Chung [a,*]

[a] Korea Advanced Institute of Science and Technology, Yuseong-gu, Guseong-dong, Daejeon 305-701, Republic of Korea
[b] Korea University of Technology and Education, Byeongcheon-myeon, Chungnam 330-708, Republic of Korea

ARTICLE INFO

Article history:
Received 3 July 2010
Received in revised form 27 August 2013
Accepted 28 August 2013

Keywords:
Semantic search
Ranking
Semantic relationship
Ontology
Semantic Web

ABSTRACT

On the Semantic Web, the types of resources and the semantic relationships between resources are defined in an ontology. By using that information, the accuracy of information retrieval can be improved.

In this paper, we present effective ranking and search techniques considering the semantic relationships in an ontology. Our technique retrieves *top-k* resources which are the most relevant to query keywords through the semantic relationships. To do this, we propose a weighting measure for the semantic relationship. Based on this measure, we propose a novel ranking method which considers the number of meaningful semantic relationships between a resource and keywords as well as the coverage and discriminating power of keywords. In order to improve the efficiency of the search, we prune the unnecessary search space using the length and weight thresholds of the semantic relationship path. In addition, we exploit Threshold Algorithm based on an extended inverted index to answer top-*k* results efficiently. The experimental results using real data sets demonstrate that our retrieval method using the semantic information generates accurate results efficiently compared to the traditional methods.

1. Introduction

With the massive growth of the Web, we have been confronted with a flood of information, and hence search engines have become one of the most helpful tools for obtaining desired information from the Web. The keyword-based search method has been the most popular search method in the search engines since it provides simple and user friendly interface. The keyword-based search method requires users to input several keywords describing the search target and returns the search results containing the keywords.

In general, the keyword-based search determines the relevance of resources for query keywords mainly based on the occurrence of the keywords in their textual descriptions (e.g., title, body, anchor text, and so on). It cannot ensure that the returned results preserve the semantic relationships among the keywords which users have intended when submitting the keywords (Li et al., 2007). Because of this reason, current search engines sometimes miss highly relevant results and return some irrelevant results for user requests. For example, consider a query looking for laboratories in Europe researching on Semantic Web. You may input the following keywords: ‘Laboratory’, ‘Semantic Web’, and ‘Europe’. For this query, one of the most popular search engines returns about 2,360,000 pages. However, on the top 20 pages, only four pages are related to

☆ A short preliminary version of this paper was published in the proceeding of WWW 2009 as a two page poster paper.

* Corresponding author. Tel.: +82 42 350 3537; fax: +82 42 350 7737.
E-mail addresses: jihyun.s.lee@gmail.com (J. Lee), jkmin@kut.ac.kr (J.-K. Min), alice.oh@cs.kaist.ac.kr (A. Oh), chungcw@kaist.edu (C.-W. Chung).

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.08.007

the desired laboratories and other irrelevant pages just contain some of the three keywords, such as 'W3C Semantic Web FAQ' and 'Semantic Web Europe'. In the irrelevant pages, the relationship *'research on'* between 'Laboratory' and 'Semantic Web' and the relationship *'located on'* between 'Laboratory' and 'Europe' are not preserved. Furthermore, the diverse implicit meanings in the relationships among keywords are ignored in the search. For example, *'research on'* implies some indirect relationships via 'published papers' about Semantic Web and 'researchers' studying on Semantic Web. As a result, many relevant pages have been pushed down in the ranked list.

We propose a semantic search framework to overcome such limitations of the traditional keyword-based search by enriching the search process with an ontology, which is one of the purposes of the Semantic Web. An ontology is a formal knowledge description of concepts and their relationships. The semantic relationships between resources and keywords could be extracted by traversing the ontology. The extracted semantic relationships can complement the keyword-based search method. Our semantic search framework extends keyword-based search by using ontologies, with the aim of finding resources relevant to query keywords through the semantic relationships. The semantic search makes hidden relationships between the words of desired resources and keywords explicit by using diverse semantic relationships defined in the ontology, thereby it can effectively access the relevant resources and rank them. Consequently, we expect that the precision and recall of the search would be improved. Recently, Google has started to support a primitive semantic search based on the knowledge graph which is similar to the ontology. Google enhances its search service by augmenting the search results with sets of associated facts based on the knowledge graph. The usefulness and feasibility of our semantic search could be confirmed by this attempt.

This semantic search would be useful when a sufficient ontology associated with the search domain is prepared in advance. As more resources and their relationships in the domain become well defined, more comprehensive semantic search would become feasible. In addition, in order to effectively limit the search scope, we assume that each query contains the type (i.e., a class in an ontology) of the desired resource such as *Publication* or *Professor*.

Fig. 1 shows the overall search process in our framework. Before searching documents (e.g., biography or web page), the ontology generator constructs an ontology describing the contents in the collected documents. For this work, we assume that the ontology has already been constructed. Note that there has been active research on the ontology construction (Kiryakov et al., 2003; Ceravolo and Damiani, 2007; Suchanek et al., 2007). Given a user query containing a type **T** and a set of keywords $\{k_1, k_2, \ldots, k_n\}$, the semantic search engine finds the relevant resources through the exploration of the ontology and returns a ranked list of the URIs of the resources in the order of their relevance. Finally, the document retriever retrieves the documents corresponding to the returned URIs.

Since we consider the semantic relationships between the resources and the query keywords in the search process, the ranking of the set of results should reflect how well each semantic relationship discriminates a result from the other results in the set. Thus, we devise a novel weighting measure for the semantic relationships, such that it assigns a higher weight to semantic relationships with less ambiguity in identifying the target resources. Further, we design a novel ranking model for resources by considering the following three major relevance criteria: the number of important relationships between resources and query keywords, the coverage of the keywords, and the discriminating power of the keywords. If there are many kinds of semantic relationships from resources to query keywords in an ontology, it would take much time to examine the entire set of relationships in the semantic search. In order to improve the efficiency of our semantic search, we prune the search space using the length and importance of the semantic relationship. Our semantic search may generate many results that are related to the query keywords. Therefore, we adapt the Threshold Algorithm (Fagin et al., 2003) to efficiently retrieve the *top-k* results without examination of the entire result set.

The main contributions of our work are summarized below:

- *A weighting method for semantic relationships:* We propose a weighting measure for semantic relationships which assigns a higher weight to significant relationships having a higher level of contribution to discriminate the answer. In our work, the weight is automatically computed without the intervention of domain experts. Thus, the weighting method could be applied to large and complex ontologies.

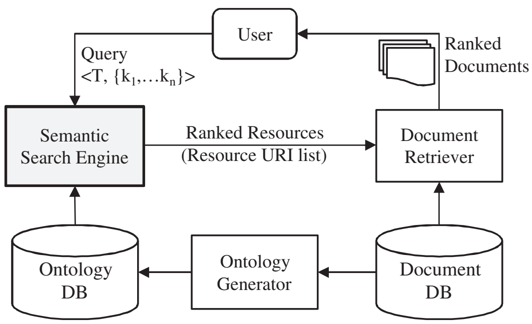

Fig. 1. Document retrieval process based on ontologies.