Expert Systems with Applications 39 (2012) 4111–4121

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Searching and ranking method of relevant resources by user intention on the Semantic Web

Myungjin Lee [a], Wooju Kim [a,*], Sangun Park [b,1]

[a] Dept. of Information and Industrial Engineering, Yonsei University, 134 Shinchon-dong, Seodaemoon-ku, Seoul 120-749, Republic of Korea
[b] Division of Business Administration, Kyonggi University, 94-6 Yiui-dong, Yeongtong-gu, Kyonggi 443-760, Republic of Korea

ARTICLE INFO

Keywords:
Semantic search
Ontology retrieval
Semantic Web
Semantic associations
Spreading activation

ABSTRACT

As the information on the Internet dramatically increases, more and more limitations in information searching are revealed, because web pages are designed for human use by mixing content with presentation. In order to overcome these limitations, the Semantic Web, based on ontology, was introduced by W3C to bring about significant advancement in web searching. To accomplish this, the Semantic Web must provide search methods based on the different relationships between resources.

In this paper, we propose a semantic association search methodology that consists of the evaluation of resources and relationships between resources, as well as the identification of relevant information based on ontology, a semantic network of resources and properties. The proposed semantic search method is based on an extended spreading activation technique. In order to evaluate the importance of a query result, we propose weighting methods for measuring properties and resources based on their specificity and generality. From this work, users can search semantically associated resources for their query, confident that the information is valuable and important. The experimental results show that our method is valid and efficient for searching and ranking semantic search results.

1. Introduction

Information retrieval (Singhal, 2001) is the science of searching for relevant documents, information within documents, or metadata about documents. Most information retrieval systems compute a numeric score based on how well each object matches with the user's query and rank the objects according to their scores. After that, the top ranking objects are displayed for the users. The first generation of automated information retrieval systems was introduced in the 1950s and 1960s. These systems highlighted the need for research into information retrieval technology, after which Tim Berners-Lee suggested a hypertext project, called the World Wide Web (Lee & Cailliau, 1990), in 1989. From the onset of the World Wide Web, continued progress in network technologies and data storage techniques has enabled the digitalization of huge numbers of documents. Consequently, search engines became very common and may be the best instantiation of information retrieval models in the huge hyperlink network. However, the existing web makes it more and more difficult for users to find relevant information, as the available information continues to dramatically increase. One of the problems is that web pages are designed by mixing content with presentation. Hyper Text Markup Language (HTML) (World Wide Web Consortium, 1992) focuses on describing the structures of web pages. HTML provides users with the means to display online documents with headings, text, tables, lists, and photos and to retrieve online information via hypertext links. Another problem is that a web search engine is typically based on keyword search techniques. Generally web search engines gather information from different sources via a web crawler by collecting, parsing, and storing keyword data from the texts of web pages to facilitate fast and accurate information retrieval. When users enter query keywords into a search engine, the web search engine examines its index of resources and provides a list of best-matching web pages according to its page ranking criteria based only on the query keywords. In order for users to locate information accurately, web page semantics should be separate from the syntax of HTML, and the web search engines need to be able to search for information based on semantics, rather than only on keywords. This need for semantics has lead to the creation of the Semantic Web.

The Semantic Web is an evolutionary progression of the World Wide Web in which the semantics of information and services are defined, making it possible for the web to satisfy user requests for web content (Lee, Hendler, & Lassila, 2001). This new system is

* Corresponding author. Tel.: +82 2 2123 7754; fax: +82 2 364 7807.
E-mail addresses: xml@yonsei.ac.kr (M. Lee), wkim@yonsei.ac.kr (W. Kim), supark@kgu.ac.kr (S. Park).
[1] Tel.: +82 31 249 9459.

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.09.127

based on the idea of providing information with explicit and formal machine-accessible descriptions of meaning. In order to make and exchange the semantics of information, the ontology that defines a common information-sharing vocabulary is generally used. An ontology, which is a formal explicit description of concepts or classes in a domain of discourse (Gruber, 1993), can be used to annotate data using metadata and interrelations. An ontology in this context consists of resources on the World Wide Web and their relationships – the network structure. The current web is a network structure that consists of web pages, with only one relationship denoting the hyperlink. However, an ontology produces a more complex network structure because it includes descriptions of concepts and represents various kinds of user-defined relationships between concepts; we call this a semantic network (Sowa, 1992). The applications in the Semantic Web can obtain an increased accuracy when processing information, providing the potential to improve the way in which search engines perform. Therefore, a different search method from that of the traditional keyword search is needed to identify relevant information in a user query. One of the core differences between the semantic search and the keyword search is the utilization of interrelationships among data, which is a resource in the Semantic Web.

The search method proposed in this paper allows for the identification of all concepts which are related to a user query even if the concepts do not explicitly include any query string. This ability is based on the spreading activation method (Crestani, 1997). Traditional search methods determine whether at least one of the query keywords appears within the documents and, if so, provide the documents to the user as a search result. For example, if a user inputs a query with the keywords "Metaweb Technology," then the search engine locates documents that include the query keywords and provides the documents to the user for review. However, our method based on spreading activation provides semantically related concepts (e.g., persons, companies, etc.) to the query "Metaweb Technology" as the search result. Therefore, the properties which relate resources in the Semantic Web are very important in semantic searching because they show why and how each resource is related to the query. Web pages in the current web are connected by only hyperlink relations, but resources in the Semantic Web are connected by one or more properties. This means, however, that different properties can imply different importances for the connected resources. The traditional keyword search generally shows just a ranked list of the keyword similarities between the user query and the identified documents. However, the search results of a semantic search can be sorted by the weights of the properties and resources, as each has its own individual importance. In order to rank the search results, the weights of properties and resources are assigned based on *specificity* and *generality*. A very interesting outcome has resulted from these weighting methods. The questions of where to start and how to span and explore the semantic network are the main issues in our research. Moreover, the visual presentation of the search results is also an important issue.

In this paper, we propose a semantic search method based on the spreading activation method and used to locate relevant results which are most semantically related to a user query. The approach is to retrieve all concepts that are related to a given keyword even if the keyword does not appear within the document. Moreover, we discuss the assignment of the weights of properties and resources in order to support semantic searching, to provide users with properly ranked search results. In other words, the processes of assigning weights and spreading to other resources on the semantic network are examined. Finally we evaluate the proposed search method over real-world data to compare our approach with another semantic search method and to test the effects of weighting.

The paper is organized as follows. Section 2 discusses related works. In Section 3 we define the data model and describe the weighting method for properties and resources. In Section 4 we propose a semantic search method. Section 5 discusses our semantic search system, and we present experimental results in Section 6. Finally we conclude the paper and our future work in Section 7.

2. Related works

Recently, a number of semantic search approaches have been published, and their application areas are diverse. However, they are based on a common set of ideas, presented and connected by Mangold (2007). Mangold presented a categorization scheme that is used to classify different approaches for semantic searches along several dimensions. In particular, he introduced categories for the following criteria: architecture, coupling, transparency, user context, query modification, ontology structure, and ontology technology. He selected ten different semantic document retrieval systems, i.e., Simple HTML Ontology Extensions (SHOE) (Heflin & Hendler, 2000), Inquirus2 (Glover, Lawrence, Gordon, Birmingham, & Giles, 2001), TAP (Guha, McCool, & Miller, 2003), etc. He compared the systems by means of the classification criteria and discussed issues that are open to further research and application development. According to his research, our system can be classified as a tight coupling between web pages and the ontology, meaning that the metadata of documents refer explicitly to concepts of a specific ontology. Therefore, our approach is classified as a graph-based approach that perceives both ontological concepts and documents as the nodes of a graph.

The MultimediaN E-Culture project (Schreiber et al., 2008), one of the semantic search systems, demonstrates how the novel Semantic Web and presentation technologies can be deployed to provide better indexing and search support within large virtual collections of cultural heritage resources. To search semantic paths, this system checks all RDF literals in the repository for matches to the given keyword and traverses the RDF graph until a resource of interest is found. Finally, the results are clustered based on the paths from the matching literals to their result. This research has some similarity with our approach, but it lacks the ability to assign weights to properties and resources and the ability to identify the search-ending threshold. These limitations are the most important issues with this system, because they are two of the determinants for expanding the semantic network. Moreover, information travels in one direction only in this limited system: always from the query object in the triple to the corresponding subject.

Some ranking techniques for the Semantic Web have been proposed. SemRank (Anyanwu, Maduko, & Sheth, 2005), which ranks results based on their predictabilities, is one of them. SemRank is based on a relevance model which is a rich blend of semantic and heuristic-including information-theoretic techniques that support the novel idea of modulative searches, in which users may vary their search modes to effect changes in results ordering depending on their need. To rank results of semantic associations, a model is constructed for measuring the information content of a semantic association by considering the occurrence of an edge as an event and RDF properties as its outcomes. In other words, SemRank proposes a measure of a property's uniqueness relative to those of other properties in the description base. However, to apply the spreading activation method in a semantic search, available resources must also be considered. Accordingly, we expand this property measurement and propose a new measure that takes into consideration the uniqueness of the resources in the semantic network.

In associative retrieval, relationships among information items are often represented as a network, where information items are represented by nodes, and associations are shown as the links