Expert Systems with Applications 41 (2014) 8225–8233

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An architectural design for effective information retrieval in semantic web

M. Thangaraj[a], G. Sujatha[b,*]

[a] Madurai Kamaraj University, Madurai 625 021, India
[b] Sri Meenakshi Govt. Arts College for Women(A), Madurai 625 002, India

ARTICLE INFO

Article history:
Available online 19 July 2014

Keywords:
Information retrieval
Semantic web
Semantic search
Ontology
Semantic query

ABSTRACT

The current web IR system retrieves relevant information only based on the keywords which is inadequate for that vast amount of data. It provides limited capabilities to capture the concepts of the user needs and the relation between the keywords. These limitations lead to the idea of the user conceptual search which includes concepts and meanings. This study deals with the Semantic Based Information Retrieval System for a semantic web search and presented with an improved algorithm to retrieve the information in a more efficient way.

This architecture takes as input a list of plain keywords provided by the user and the query is converted into semantic query. This conversion is carried out with the help of the domain concepts of the pre-existing domain ontologies and a third party thesaurus and discover semantic relationship between them in runtime. The relevant information for the semantic query is retrieved and ranked according to the relevancy with the help of an improved algorithm. The performance analysis shows that the proposed system can improve the accuracy and effectiveness for retrieving relevant web documents compared to the existing systems.

1. Introduction

1.1. Motivation

The World Wide Web serves as a huge wide distributed global information center for many information services. By now the size of the web is billions of websites and is still growing rapidly. But to get an exact requirement a normal user often spends a lot of time. In order to present the relevant results from this voluminous data to the user, some new methods should be derived to filter the results. The current information technology from the web is mostly based on the keywords. It provides limited capabilities to capture the concept of the user requirement. To solve the limitations of the keyword based search the idea of semantic search is introduced in the field of information retrieval (IR). Information retrieval is the science of searching for documents, information within the documents as well as that of relational database and the World Wide Web. IR also deals with representing, storing and organizing the content.

Semantic search has been presented in the IR field since the early eighties (Croft, 1986). The use of ontologies with keyword based search is one of the motivations of the semantic web (SW). The semantic web “targets to build an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation” (Berners-Lee, Hendler, & Lassila, 2001). Fig. 1 shows the layers of the semantic web as suggested by Berners-Lee.

The bottom layer contains technologies that provide basics for the SW. *Uniform resource identifiers* (URIs) provide a standard way to refer to entities, while *Unicode* is a standard for exchanging symbols. The *Extensible Markup Language* (XML) fixes a notation for describing labelled trees, and XML Schema allows the definition of grammars for valid XML documents. XML documents can refer to different *namespaces* to make explicit the context (and therefore meaning) of different tags. The middle layer contains technologies to enable building SW applications. Resource Description Framework (RDF) is a framework for creating statements in the form of resources, properties and statements as triples. RDF schema provides a basic vocabulary of RDF. Web Ontology Language describes semantics of RDF statements. SPARQL is RDF Query language. Top layer contain just ideas that should be implemented in order to realize SW. Cryptography, Trust and Proof is to ensure that the

* Corresponding author.
E-mail addresses: thangarajmku@yahoo.com (M. Thangaraj), sujisekar05@rediffmail.com (G. Sujatha).

http://dx.doi.org/10.1016/j.eswa.2014.07.017
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

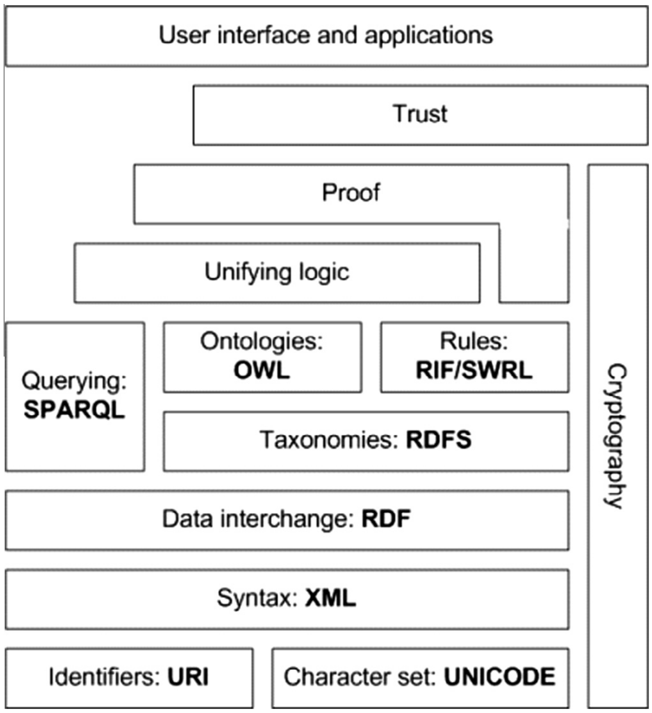

Fig. 1. Semantic web architecture.

SW statements are from trusted source. User interface is the final layer that will enable humans to use SW applications.

Some of the main goals of the semantic web are Semantic Query Processing and discovering the semantic information available in the unstructured web information with the help of domain ontologies. The IR from the semantic web combines the fast-developing research areas information retrieval, semantic web, and Web content Mining.

The various problems associated with the unstructured web pages are identified as follows. (i) Web pages are far complex than that of any traditional document collection. (ii) The web is a highly dynamic information source. (iii) The web serves a broad diversity of user communities. (iv) Only a small portion of the information on the web is truly relevant or useful. These challenges have promoted research into effective and efficient discovery and uses of resources on the internet. There are extensive research activities on the construction and use of semantic web which is nothing but the structure of semantic meaning of the content of the web pages. Web document classification by web mining will help in building the ontology for the semantic web with the automatic extraction of the semantic meaning of web pages.

In order to address this issues, the Semantic Based Information Retrieval System (SBIRS) mechanism for SW is proposed. This architecture handles the semantic indexing, extraction, extensions of query and matching of content semantics to achieve the following objectives. (i) Analyze and determine the semantic feature of the content by means of semantic annotation. (ii) Analyze the user's query and extend it to semantic query using link extraction, ontology and thesaurus. (iii) Match the semantic query with the semantic content using a semantic indexing structure. (iv) Arrange the retrieved results in the order of their relevancy to the query using proposed dynamic ranking algorithm. This architecture eliminates the problems of traditional keyword search and enables the user to retrieve the concept oriented relevant results for any domain.

The significance of the framework is to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data space. SBIRS also has a few aspects that distinguish it from other related work. Unlike typical search algorithms this framework is based on keyword-to-concept mapping with an improved semantic indexing structure and searching technique. The proposed dynamic ranking algorithm presents the results in the order of their relevance for the expanded semantic query.

1.2. Research contributions

Contributions of this research fall into the following categories.

- Clear knowledge of the semantic search, possibilities of semantic enhancements in the IR models.
- Definition and implementation of a semantic retrieval model with generic domain ontology.
- Creation of an improved semantic indexing structure.
- Implementation of a dynamic ranking algorithm
- Investigation of the feasibility of semantic retrieval in cloud environment.
- Checking the feasibility of semantic image retrieval.

1.3. Structure of the paper

The rest of the paper is organized as follows. An overview of related work is given in Section 2. In Section 3 the working mechanism of the proposed architecture is explained. Section 4 elucidates the retrieval and the ranking algorithms. The performance evaluation is given in Section 5. In Section 6 the main achievements and the tasks that remain is discussed.

2. Related work

The unsolved problems of current search engines have led to the development of semantic web search system (Yi Jin & Hongwei Lin., 2008). Conceptual search has been the motivation of a large body of research in the IR field long before the semantic web vision emerged (Jo rvelin, Kekalainen, & Niemi, 2001). ''SemSearch'' (Yuangui Lei & Enrico Motta, 2006) is a layered architecture that separates end users from the back-end heterogeneous semantic data repositories. ''SemSearch'' accepts keywords as input and delivers results which are closely relevant to the user keywords in terms of semantic relations. The SBIRS compliments SemSearch with a ranking algorithm designed specifically for an ontology-based information retrieval model with a semantic indexing structure based on annotation weighing techniques.

The inherited relationships between the keywords are analyzed in terms of concepts in ''Ontolook'' (Li, Wang, & Huang, 2007). From this concepts and relations a concept-relation graph is formed which is used to eliminate the less ranked arcs. It also creates a property-keyword candidate set and sent it to the web page database to get a retrieved result set for the users. The efficiency of this approach is limited by lack of ranking technology. This motivates a relation based page ranking algorithm for semantic web search (Lamberti, Sanna, & Demartini, 2009). The ranking technology is based on the estimate of the probability that keywords/concepts within an annotated page are linked one with another in a way that is the same to the one in the user's mind at the time of submitting the query. The probability is measured using a graph-based description of ontology, user query and the annotated page. In these approaches further efforts are requested for future semantic web repositories based on multiple ontologies and better ranking. By building upon a dynamic ontology our model supports multiple domains with semantic dynamic ranking.