Journal of King Saud University – Computer and Information Sciences (2014) **26**, 161–169

King Saud University

**Journal of King Saud University – Computer and Information Sciences**

www.ksu.edu.sa
www.sciencedirect.com

# Effective semantic search using thematic similarity

**Sharifullah Khan** *, **Jibran Mustafa**

*National University of Sciences and Technology (NUST), School of Electrical, Engineering & Computer Science H-12, Islamabad, Pakistan*

Received 12 October 2012; revised 18 April 2013; accepted 12 October 2013
Available online 22 October 2013

**KEYWORDS**

Semantic search;
Thematic similarity;
Semantic heterogeneity;
RDF triples;
Information retrieval

**Abstract** Most existing semantic search systems expand search keywords using domain ontology to deal with semantic heterogeneity. They focus on matching the semantic similarity of individual keywords in a multiple-keywords query; however, they ignore the semantic relationships that exist among the keywords of the query themselves. The systems return less relevant answers for these types of queries. More relevant documents for a multiple-keywords query can be retrieved if the systems know the relationships that exist among multiple keywords in the query. The proposed search methodology matches patterns of keywords for capturing the context of keywords, and then the relevant documents are ranked according to their pattern relevance score. A prototype system has been implemented to validate the proposed search methodology. The system has been compared with existing systems for evaluation. The results demonstrate improvement in precision and recall of search.



## 1. Introduction

Digital repositories facilitate users in archiving digital documents. However, semantic heterogeneity in their content causes difficulties in retrieving relevant documents (Alipanah et al., 2010; Rinaldi, 2009; Lee and Soo, 2005; Khan et al., 2004; Blasio et al., 2004). Semantic heterogeneity refers to similar data that are represented differently in a document, for example, the use of the word author versus the word writer. There are different semantic heterogeneity issues such as polysemy and synonymy (Yang et al., 2011; Fang et al., 2005; Lee and Soo, 2005; Rodriguez and Egenhofer, 2003; Uschold and Gruninger, 2004). A synonym refers to a word that has the same meaning as another word; e.g., movie is a synonym of film. Polysemy refers to a word or phrase with multiple related meanings; e.g., a bank can refer to a financial institute in one context and a river corner/edge in another context. The main concern in information retrieval (IR) is to effectively retrieve relevant information from repositories.

Domain ontology provides a conceptual framework for the structured representation of context, through a common vocabulary in a particular domain (Bonino et al., 2004; Fang et al., 2005). The vocabulary usually includes concepts, relationships between concepts, and definitions of these concepts and relationships. For example, in a statement "*Bilal works in HSBC*," *Bilal* and *HSBC* are concepts, and works is a relationship between these concepts. Moreover, ontology rules and

* Corresponding author. Tel.: +92 51 9085 2150; fax: +92 51 831 7363.
E-mail addresses: sharifullah.khan@seecs.edu.pk (S. Khan), jibran.mustafa@seecs.edu.pk (J. Mustafa).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier


http://dx.doi.org/10.1016/j.jksuci.2013.10.006

axioms are also defined to define new concepts that can be introduced in ontology and to apply logical inference (Ding et al., 2004). Semantic similarity refers to semantic closeness, proximity, or nearness. It indicates similarity between different concepts and their relationships. There are three types of semantic similarity: (a) surface, (b) structure, and (c) thematic similarity (Poole et al., 1995; Zhong et al., 2002; Zhu et al., 2002; Montes-Y-Gomez et al., 2000). Surface and structure similarity focus individually on concepts and relationships, respectively, whereas thematic similarity considers the pattern (i.e., combination) of concepts and the relationship that exists among them. The term "keyword" stands for either a concept or relationship of domain ontology alternatively in this paper.

Existing typical semantic search systems (Bonino et al., 2004; Fang et al., 2005; Varelas et al., 2005) expand individual keywords through domain ontology to deal with different semantic heterogeneity challenges such as synonymy. For example, a search for the concept *writer* can be expanded through domain ontology to the keywords *writer* and *author*. The search, looking only for a keyword *writer* may have fewer results than the search looking for *writer* and *author*. The existing systems focus on matching the semantic similarity of individual keywords (i.e., they apply either surface or structure similarity) and apply Boolean operators if multiple keywords are given in a query. They ignore the semantic relationships that exist among the multiple keywords themselves.

If a user inputs a multiple keywords query, for example, "pipe in computer science domain," conventional IR systems retrieve thousands of documents where pipe might be used as (a) a tube of any kind, (b) a device for smoking, (c) a musical instrument or (d) a portion of memory that can be used by one process to pass information to another process in computer. Sometimes none of search results may be relevant to a user requirement. The systems return less relevant answers for multiple keywords queries although they expand individual keywords in a query with different semantic relationships.

More relevant documents for a multiple keywords query can be retrieved if systems know the meanings and relationships that exist among the multiple keywords themselves in the query. By keywords pattern, we mean a combination of at least two concepts and their relationship that exists in the domain ontology. A pattern can represent the context/theme, that is, circumstances in which something happens or should be considered. Therefore, the existing systems (Bonino et al., 2004; Fang et al., 2005; Varelas et al., 2005; Rinaldi, 2009; Alipanah et al., 2010; Yang et al., 2011) cannot resolve the semantic heterogeneity issue of polysemy because it requires identification of the context of keywords to comprehend their actual semantics. Moreover, the existing systems also ignore other important relationships, such as semantic neighborhoods (Rodriguez and Egenhofer, 2003), that can also contribute to useful search results.

To overcome the limitations of existing semantic searching systems, we need to represent the context of keywords through keyword patterns for effective searching using thematic similarity (Khan et al., 2006; Poole et al., 1995). The proposed system concentrates on searching keyword patterns and not on the individual keywords. We employed Resource Description Framework (RDF) triples to describe the keyword patterns of document metadata and search queries. We have developed a prototype system for the validation of the proposed solution. The system was compared with existing systems (Fang et al., 2005; Shah et al., 2002) for evaluation, and the results demonstrate improvement in precision and recall of semantic searching.

The remainder of this paper is structured as follows: Section 2 reviews the current approaches to semantic search techniques and their proposed systems. Section 3 explains our proposed searching methodology in detail. Section 4 illustrates a walk-through example for demonstrating the proposed methodology. Section 5 discusses the evaluation of the prototype system, and Section 6 concludes the paper.

## 2. Related work

Several methods for determining semantic similarity between keywords, i.e., either concepts or relationships, have been proposed in the literature. These methods are classified into three main categories (Varelas et al., 2005). We discuss first the methods in this section and then describe existing systems that have applied the methods.

### 2.1. Semantic similarity methods

#### 2.1.1. Edge counting methods

These methods measure semantic similarity between two keywords as a function of length of the path (i.e., distance) linking keywords and their position in their respective hierarchy (Rodriguez and Egenhofer, 2003; Varelas et al., 2005). This similarity calculation simply relies on counting the number of edges separating two keywords by an 'Is-A' relation in ontology (Rada et al., 1989). This technique assumes that the semantic difference between upper-level keywords in a hierarchy is greater than the semantic difference between lower-level keywords. In other words, general concepts are less similar than two specialized concepts. Because the specialized concepts may appear more similar than general ones, depth is taken into account by calculating either the maximum depth in the hierarchy (Leacock et al., 1998) or the depth of the most specific concept, while subsuming the two compared concepts/relationships (Hirst et al., 1998; Wu et al., 1994). Semantic similarity between concepts is calculated with reference to its closest common parent (ccp).

#### 2.1.2. Information content methods

These methods measure the difference in information of two concepts as a function of their probability of occurrence in a corpus. They are also known as term frequency (*tf*)/inverse document frequency (*idf*). In these methods, two concepts are similar to an extent to which they share information in common. Therefore, the information content value for each concept in the hierarchy is calculated using its frequency in the corpus (Resnik, 1999).

#### 2.1.3. Feature-based methods

These methods measure similarity between two concepts either as a function of their properties or characteristics. These methods assume two concepts are similar if they have more common characteristics than non-common characteristics (Tversky, 1977).

### 2.2. Existing systems

DOSE (Bonino et al., 2004) uses *tf*/*idf* based on a Vector Space Model (VSM) for keywords. This system extended the tradi-