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-

tional VSM by including taxonomic relationships (i.e., specialization and generalization) of keywords for query expansion. They expand a query vector through relationships and compute semantic similarity values between a document vector and the expanded query vector using the cosine of the angle between them.

In Fang et al. (2005), the authors employ *tf*/*idf* based on a traditional Vector Space Model (VSM) for keywords. They extend the model by considering semantic relationships (i.e., direct, strong, normal, weak and irrelevant) of keywords for query expansion. They define weights for these relationships (i.e., the weights of direct, strong, normal, weak and irrelevant are 1.0, 0.7, 0.4, 0.2 and 0.0, respectively). A user query is expanded through these relationships. The *tf*/*idf* values of query keywords in a document are adjusted by multiplying weights of the semantic relationships that exist between the query and document keywords. Then, documents are ranked according to the relevance score.

In Varelas et al. (2005), the authors use a Semantic Similarity Retrieval Model (SSRM) for keywords. They extend the model by including semantic relationships (i.e., synonyms, hyponyms and hypernyms) of keywords for query expansion and assign weights to relationships depending on the position of keywords in taxonomy. They expand a user query to a specific threshold weight. The *tf*/*idf* values of query keywords in a document are adjusted by multiplying the weights of the semantic relationships that exist between the query and document keywords. They rank documents according to the gained weights.

The system presented in Shah et al. (2002) computes the frequency of the RDF triples, instead of keywords, in a document. The system calculates the similarity on the basis of the RDF triples' frequency in the document and employs only inference rules for semantic searching. The system does not expand user queries through semantic relationships. Therefore, this system cannot resolve the semantic heterogeneity issue of polysemy where the context of keywords is needed to comprehend the accurate semantics of keywords required. Some systems (Khan et al., 2006; Zhong et al., 2002; Zhu et al., 2002) use the edge counting method (e.g., distance-based method) in conceptual graph (CG) for semantic search. The basic intuition in conceptual graph (CG) matching is to calculate semantic matching by comparing arcs. The comparison of CG arcs concentrates on the thematic behavior of concepts and relationships (i.e., keywords), which is a representative of a given context. We have borrowed their notion of semantic matching for RDF triples and extended their searching techniques by computing *tf*/*idf* of these triples, i.e., ranked-result.

In Khan et al. (2004), the authors developed a concept-based model that uses domain-dependent ontologies for responding to user requests. They apply an automatic query expansion technique on user queries that are expressed in natural language. This automatic expansion technique selects only relevant and controlled expansion. The AKTiveRank (Alani and Brewster, 2005) system ranks ontologies using graph analysis measures. The authors also apply Swoogle[1] to measure the semantic relatedness before applying their proposed technique. Both techniques cannot handle polysemous heterogeneity. For the Dynamic Semantic Engine (DySE) (Rinaldi, 2009), the authors designed a context-driven approach in which keywords are processed in the context of the information in which they are retrieved. Their query is a list of terms to retrieve and a domain of interest. Then, they apply a ranking to the retrieved results.

In Alipanah et al. (2010), the authors proposed a weighting mechanism to find the expansion of concepts from ontologies. They determined expanded terms/concepts in each ontology (i.e., document) on the basis of semantic similarity, density and betweenness to a user query. Then, they use the idea of co-occurrence across ontologies. Similar concepts are determined by their name and structural similarity in each ontology. At the end of expansion, the system generates a set of terms along with weights and ranks them according to weights. In Yang et al. (2011), the authors retrieve textual information via word meanings rather than lexical forms. The authors apply WorldNet for word sense disambiguation, and then, this semantic information is annotated in the documents in a RDF used for semantic searching. To the best of our knowledge, none of these techniques measure semantic relationships among keywords patterns (i.e., RDF triples) and then produce a ranked result to facilitate meeting the user's query request.

3. Proposed searching methodology

In this section, we discuss our searching methodology, which performs context-driven semantic search by matching a user query, given in RDF triples (i.e., user-given patterns), with RDF triples of a digital document.

3.1. Query expansion

In the proposed system, we take into account the following semantic relationships: (a) synonyms, (b) semantic neighborhood and (c) hyponyms (i.e., Is-A relation-ship). Hyponyms are handled in the RDF by the built-in property of subclass (i.e., *rdfs*:*subClassOf*). We have designed two additional properties: *synOf* and *neighborOf* in the RDF for handling the remaining two relationships. A user query can be expanded by deducing inferences through rules from existing RDF triples. In the following subsections, we describe the properties and rules designed in this research.

3.1.1. SynOf property

The *synOf* property states that different individuals can be same (i.e., equivalence relationship). This property may be used to create a number of different names that refer to the same individual. It can also refer to acronyms and lexical variants. Fig. 1 shows the RDF graph of the *synOf* property. The following statement represents the RDF of *synOf* property in N3 notation:

uri : *synOfardfs* : *Property*

3.1.2. neighborOf property

The *neighborOf* property is used to explore the semantic neighborhood of a concept or relationship. The semantic neighborhood n of a concept c is the set C_i of concepts whose distance d to the concept c is an integer number r greater than zero, which is called the radius of the semantic neighborhood (Rodriguez and Egenhofer, 2003).

[1] http://swoogle.umbc.edu/.

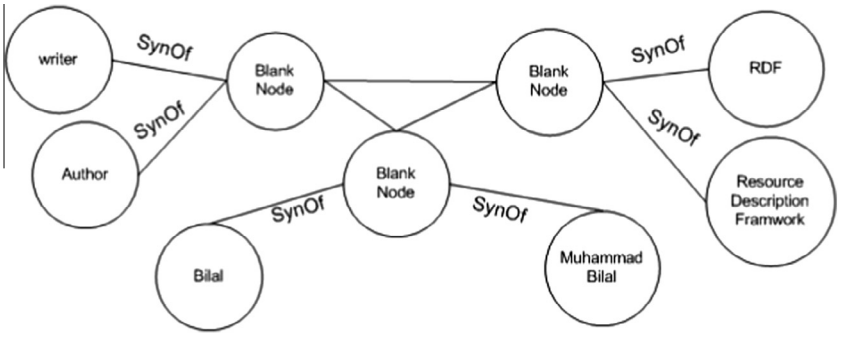

Figure 1 The RDF graph of the *synOf* property.

$$n(c,r) = \{C_i | \forall d(c,C_i) \leq r\} \quad (1)$$

A semantic neighborhood with $r = 1$ represents subclass, super-class and part-whole relationships. Fig. 2 shows the RDF graph of the *neighborOf* property. The below code snippet represents the RDF of the *neighborOf* property in N3 notation:

uri : *neighborOfardfs* : *Property*

3.1.3. Inference rules

A rule is an object that can be applied to deduce inferences from existing RDF triples (i.e., data). A user query can be expanded by deducing inferences through rules from existing RDF triples. A rule-base is an object that contains different rules. We have defined different rules in our rule-base, and it grows incrementally with the passage of time. Examples of the defined rules are *inverseOf* and *transitiveOf*, as shown in Table 1. The *inverseOf* rule defines the relationship taken *backwards*:c_1 is related to c_2 through a relationship R, then c_2 will be related to c_1 through $R-1$. The *transitiveOf* rule defines if c_1 is related to c_2 and c_2 is related to c_3 with a relationship R, then there exists a relationship R between c_1 and c_3.

3.2. Semantic similarity

After expanding a user query, the semantic similarity of the query with a document is computed. We focus on thematic similarity by matching RDF triples. The following subsections describe the details of the concepts and relationship similarities.

3.2.1. Concepts similarity

Concepts similarity is measured by calculating the distance between the concepts (Khan et al., 2006; Varelas et al., 2005). Distance is calculated between different concepts from a concept position in the hierarchy. The position of a concept, milestone (n), in the hierarchy is defined in Khan et al. (2006) as follows:

$$\text{milestone}(n) = \frac{\frac{1}{2}}{k^{l(n)}} \quad (2)$$

where k is a predefined factor and larger than one and indicates the rate at which the value decreases along the hierarchy, and $l(n)$ is the depth of the keyword n in hierarchy. For the root of a hierarchy, $l(\text{root}) = 0$. We used $k = 2$ to construct hierarchy milestone values as a multiple of 2 (i.e., binary number system). Any two concepts in the hierarchy are assumed to have a closest common parent (*ccp*). The distance between two concepts c_1, c_2 and their *ccp* will be determined by their closest common parent as follows:

$$d_c(c_1,c_2) = d_c(c_1,ccp) + d_c(c_2,ccp) \quad (3)$$

$$d_c(c,ccp) = \text{milestone}(ccp) - \text{milestone}(c) \quad (4)$$

Thus, the similarity between two concepts c_1 and c_2 is calculated as follows:

$$\text{sim}_c(c_1,c_2) = 1 - d_c(c_1,c_2) \quad (5)$$

There are some exceptions that if the concept c_1 and concept c_2 are synonyms or acronyms of each other, the distance will be set to zero, i.e., the similarity between these two concepts will be one. We assume synonym and acronym relations between concepts at the same level.

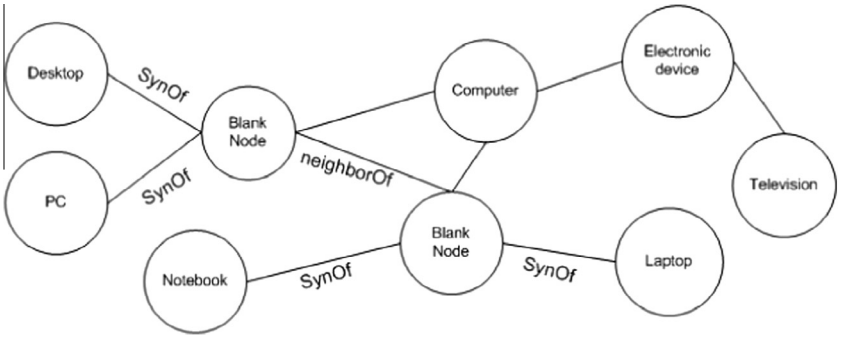

Figure 2 RDF graph of the *neighborOf* property.

Table 1 Description of different proposed rules.

Rule name	Rule description
inverseSynOf	(?x synOf ?y) → (?y synOf ?x)
inverseNeighborOf	(?x neighborOf ?y) → (?y neighborOf ?x)
transitiveSynOf	(?x synOf ?y) (?y synOf ?z) → (?x synOf ?z)
transitiveNeighborOf	(?x neighborOf ?y) (?y neighborOf ?z) → (?x neighborOf ?z)
neighborOf	(?x neighborOf ?y) (?x synOf ?w) (?y synOf ?u) → (?w neighborOf ?u)
parentOf	(?x subClassOf?y) → (?y parentOf ?x)

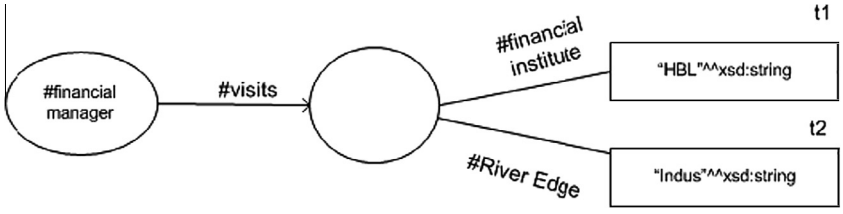

Figure 3 Content S1 metadata triples in an RDF graph.

3.2.2. Relationship similarity

Likewise, the similarity between two relationships is defined as follows:

$$\mathrm{sim}_r(r_1, r_2) = 1 - d_r(r_1, r_2) \tag{6}$$

The distance between two relations is also calculated by their respective positions in the relation hierarchy. The only difference is that a relation hierarchy was constructed manually by us. There are some exceptions that if relations r_1 and r_2 are synonyms or acronyms of each other, then the distance will be set to zero and consequently, the similarity between these two relations will be one.

3.2.3. RDF triples similarity

A user query and data source RDF triples are matched to find their similarity. The final triple similarity matching formula based on combining Eq. (5) (for concepts similarity) and Eq. (6) (for relations similarity) is as follows:

$$\mathrm{sim}(q,s) = \prod_{i=0}^{n}\prod_{j=0}^{m} \begin{matrix} \mathrm{sim}_{\mathrm{sub}}(q^i_{\mathrm{sub}}, s^j_{\mathrm{sub}}) \\ \mathrm{sim}_r(q^i_r, s^j_r) \\ \mathrm{sim}_{\mathrm{obj}}(q^i_{\mathrm{obj}}, s^j_{\mathrm{obj}}) \end{matrix} \tag{7}$$

where q_{sub}, q_{obj} and s_{sub}, s_{obj} are matched concepts, whereas q_r and s_r are matched relations of the RDF triple query q, and the RDF triple source s, respectively. sim (q, s) is the overall similarity between the query q and source s RDF triples. Here, i and j represent ith and jth subject or object or relation of the query and source RDF triples, respectively.

3.3. Documents ranking (R(d))

Identified relevant documents are ranked according to their relevance to a user query. The relevance of a document is computed by extending a *tf.idf* weighting scheme (Zhong et al., 2002) for triples instead of keywords. Let N be the total number of documents and n_i the number of documents in which the triple t_i appears.

$$tf_{ij} = \frac{\mathrm{freq}_{ij}}{\max_i(\mathrm{freq}_{ij})} \tag{8}$$

Let freq_{ij} be the frequency of the triple t_i in the document d_j. Then, the normalized frequency tf_{ij} of the triple t_i in d_j is the ratio of the term frequency of t_i to the maximum term frequency of any triple in the document d_j.idf_i is the inverse document frequency for t_i given by:

$$idf_i = \log\left(\frac{N}{n_i}\right) \tag{9}$$

The final *tf*:*idf* weight of ith triple to jth document is calculated as follows:

$$W_{ij} = tf_{ij} \times idf_i \tag{10}$$

The ranking algorithm combines two factors: (i) the RDF triple score using Eq. (7) and (ii) its relevance to a document indicated by W_{ij} using Eq. (10). The documents relevance, $R(d)$, can be calculated as follows:

$$R(d) = \sum_{i=0}^{n} \mathrm{sim}(q_i, s_i) \times (W_{ij} + \lambda) \tag{11}$$

where d represents a document, n is the total number of triples in a document and λ is used to normalize the effect of partial and imprecise RDF triples. We used 1 (one) as the default value for λ. The documents are ranked according to their relevance score and returned to a user.

4. Walk-through example

To demonstrate our proposed methodology, we consider an example. We employed RDF to represent keyword patterns of data sources and queries and represented their RDF triples with graph notations in this paper. Figs. 3–5 show the metadata triples of documents 1, 2 and 3 (i.e., s_1, s_2 and s_3), respectively. Table 2 shows the frequency of triples in the given documents. Fig. 6 illustrates the concepts hierarchy and their milestone value with respect to their position in the taxonomy. This modified segment is adopted from the WordNet[2] ontology.

[2] http://wordnet.princeton.edu/.

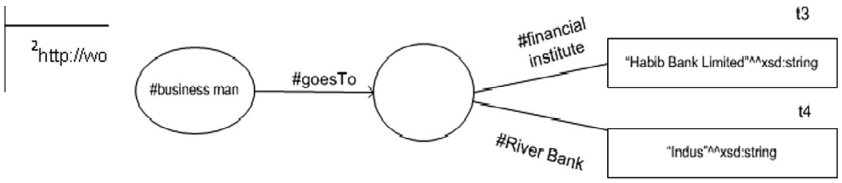

Figure 4 Content S2 metadata triples in an RDF graph.

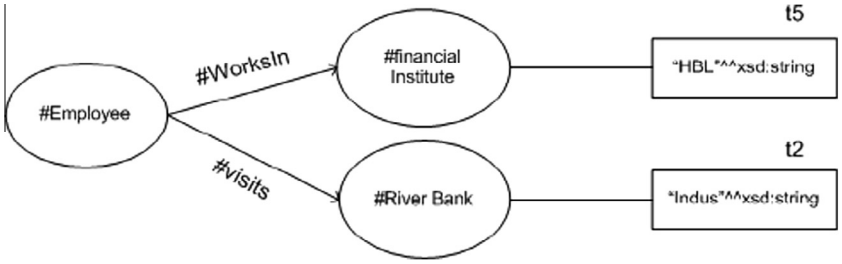

Figure 5 Content S3 metadata triples in an RDF graph.

Table 2 Triple frequency in documents.

Doc.	t_1	t_2	t_3	t_4	t_5
s_1	20	5	0	0	0
s_2	0	0	11	4	0
s_3	0	15	0	0	10

Suppose a user enters a query: "*Find all worker(s) who has visited the bank*: *HBL on certain date*." This query can be represented in a RDF triple as follows:

(?worker: visits: bank^^HBL) where date like '?date'

We expand the concepts and relationships (i.e., terms) and compute the similarity of terms between the query RDF triples and the documents' RDF triples. Their similarity scores, using a distance-based approach, are shown in Table 3. Based on the similarity scores for concepts and relations, the RDF triples similarity for three relevant triples is calculated and shown in Table 4. After identifying the relevant triples, we calculate a *tf.idf* weight according to formulas given above using the values given in Table 2. The ranks of documents are computed by combining the triple similarity score and documents *tf.idf* score, as shown in Table 4.

5. Evaluation

Traditional information retrieval systems employ a trade-off between the precision and recall to quantitatively measure the performance of information retrieval. Precision is the ratio of relevant retrieved documents to the number of retrieved documents and recall is the ratio of relevant retrieved documents to the all relevant documents (Baeza-Yates et al., 1999; Kobayashi and Takeda, 2000). A prototype system has been implemented to validate/evaluate our proposed methodology. To evaluate the research, the prototype system has been compared with the existing systems.

The experiments were performed with a collection of 100 documents that includes master thesis and conference papers from the computer science domain. We manually constructed on, an average, 37 RDF triples for each document. To perform the evaluation, we extended the WordNet[3] research ontology and ACM topic Hierarchy[4] to create our own extended domain ontology for the selected documents. To compare the proposed system, we selected two semantic search techniques, i.e., RDF-based VSM (Shah et al., 2002) and the IR framework proposed in Fang et al. (2005) because our search approach is similar to them. They both use semantic similarity and ranking for searching purposes, as we used in our approach. However, they maintain statistics of concepts in a document, and we maintain statistics of triples in a document, and there is a difference in the semantic similarity techniques used for searching. They do not apply thematic semantic similarity. Our focus is on thematic similarity. The aim of selecting these approaches for evaluation was to evaluate how much semantic search results can be improved if a thematic similarity approach is used.

Thirty test queries were formulated and run on all the three systems. Two test queries out of the set are (i) *show all IEEE conference paper written by Brown* and (ii) *Find papers about the use of ontologies in data integration in year 2005*. These are shown in Table 5. Q_1 is quite simple, and the precision of all the systems on this query is quite high, whereas Q_2 is not simple, so the precision of VSM is quite low. The precision and recall of the proposed system are better than the RDF-based VSM (Shah et al., 2002) and IR framework (Fang et al., 2005), as shown in Fig. 7. The results of the experiments have revealed that the proposed system has improved precision by 42% and 27% and recall by 19% and 16% compared to the RDF-based VSM and IR framework, respectively, as shown in Fig. 8.

Fig. 9 shows a comparison graph of the f-measure of the proposed and the existing systems. F-measure is the weighted mean of precision and recall. The f-measure of the RDF-based VSM is 0.59 as the number of documents is lower, but it de-

[3] http://www.w3.org/2001/sw/BestPractices/WNET/wordnet.rdf [July 23, 2008].
[4] http://www.acm.org/class/1998/ccs98.html [July 21, 2008].

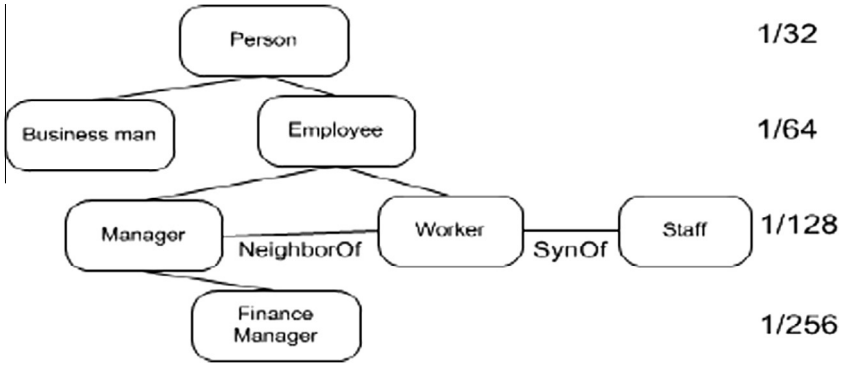

Figure 6 Ontology segment.

Table 3 Terms similarity scores.

Similarity	Score	Remarks
sim_c (worker; finance manager)	0.98047	
sim_c (worker; businessman)	0.09609	
sim_c (worker; employee)	1	Isa
sim_r (visits; visits)	1	Same
sim_r (visits; goesto)	1	Synonym
sim_r (visits; worksin)	1	Synonym
sim_c (HBL; Habib Bank Limited)	1	Acronym
sim_c (HBL; HBL)	1	Same
sim_c (HBL; Indus)	0	Not related

Table 4 Document ranking.

s_j $\{t_i\}$	sim (q, s_j)	tf_{ti}	idf_{ti}	$W_{ti,sj}$	R (s_j)
s_1 $\{t_1\}$	0.9804875	1	0.447	0.447	1.41877
s_2 $\{t_3\}$	0.9609375	1	0.447	0.447	1.39048
s_3 $\{t_5\}$	1	0.667	0.447	0.298	1.298

Table 5 Sample test RDF queries.

Q_1	Q_2
(?p: written By: Brown)	(?p: has Content: ontologies)
(?p: has Type: conference)	(?p: is About: Data Integration)
(?p: has Publication Org: IEEE)	(?p: has Publication Year: 2005)

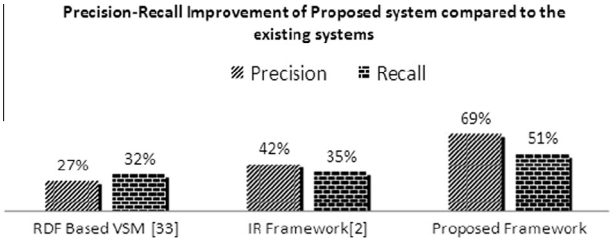

Figure 8 Precision-recall improvement of proposed system compared with two existing systems.

creases to 0.09 as the number of documents increases. These results demonstrate the inconsistency of the system. The upper bound of the f-measure for the IR framework is 0.76, and 0.26 is the lower bound. The upper bound of the f-measure of the proposed system is 0.85, and 0.48 is the lower bound. The proposed system displays better consistency between precision and recall compared with the other two systems when the number of documents is increased.

6. Conclusion

In this paper, we proposed a semantic search methodology using thematic similarity to resolve semantic heterogeneity issues involved in retrieving information. We proposed a triple-centric technique for maintaining source(s) metadata to

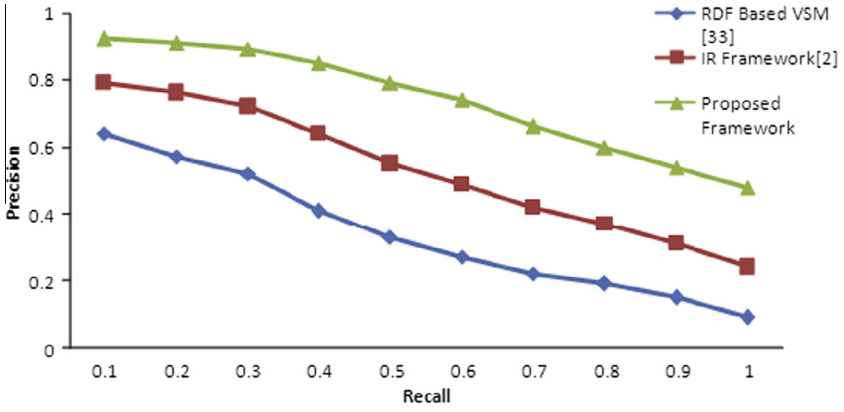

Figure 7 Precision-recall of RDF-based VSM, IR framework and proposed system.

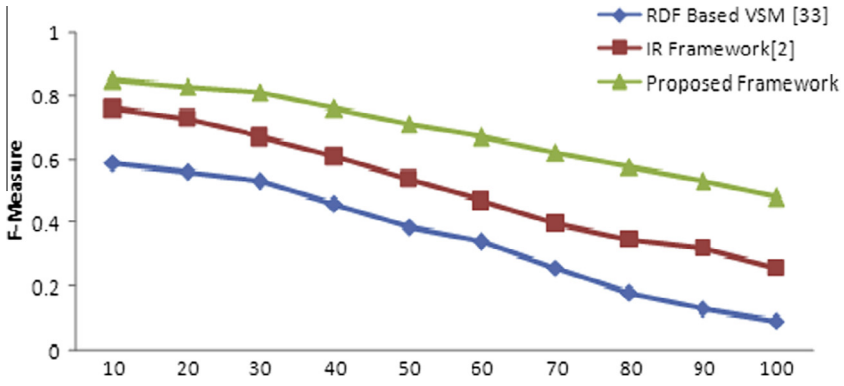

Figure 9 F-measure of RDF based VSM, IR framework and the proposed framework.

capture the context of keywords. The thematic similarity approach has been used for information retrieval to capture the context of concepts. A user submits an RDF triples query. The query is expanded through synonyms and a semantic neighborhood using distance based-approaches with the help of domain ontology. The relevance between the RDF triples of a document and a user query is measured, and the relevant documents are identified. The documents are ranked according to their importance. The contribution of this research work is to combine existing measures and design a novel semantic search methodology for thematic similarity to handle semantic heterogeneity, particularly polysemy.

The results of the experiments performed on the system indicate improvements in precision and recall and encourage new efforts in this direction. The proposed search methodology can be easily extended using recent measures for semantic relatedness and ranking methods. Moreover, we intend to automate the process of generating RDF triples from documents, as we generated them manually in this research to evaluate the prototype system. Lastly, we urge augmentation of the system for other heterogeneities i.e., incomplete and incompatible RDF triples. In the current system, we do not consider partially (i.e., incomplete) matched RDF triples that may contain important information.

References

Alani, H., Brewster, C., 2005. Ontology ranking based on the analysis of concept structures. In: Proceedings of the 3rd International Conference on Knowledge Capture. ACM, pp. 51–58.

Alipanah, N., Parveen, P., Menezes, S., Khan, L., Seida, S.B., Thuraisingham, B., 2010. Ontology-driven query expansion methods to facilitate federated queries. In: Proceedings of 2010 IEEE International Conference on Service-Oriented Computing and Applications (SOCA). IEEE, pp. 1–8.

Baeza-Yates, R., Ribeiro-Neto, B., et al, 1999. In: Modern Information Retrieval, vol. 82. Addison-Wesley, New York.

Blasio, J.D., Kawamura, T., Hasegawa, T., 2004. Catalog search engine: Semantics applied to products search. In: Proceedings of 4th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2004), vol. 184, pp. 11–20. Available from: <http://ceur-ws.org/Vol-184/>.

Bonino, D., Corno, F., Farinetti, L., Bosca, A., 2004. Ontology driven semantic search. WSEAS Transaction on Information Science and Application 1 (6), 1597–1605.

Ding, L., Finin, T.W., Joshi, A., et al., 2004. Swoogle: a search and metadata engine for the semantic web. In: Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management. Washington, DC, USA, pp. 652–659.

Fang, W.D., Zhang, L., Wang, Y.X., Dong, S.B., 2005. Toward a semantic search engine based on ontologies. In: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol. 3. IEEE, pp. 1913–1918.

Hirst, G., St-Onge, D., 1998. WordNet: an electronic lexical database. In: Lexical chains as representations of context for the detection and correction of malapropisms. The MIT Press, Cambridge, MA, pp. 305–332.

Khan, L., McLeod, D., Hovy, E., 2004. Retrieval effectiveness of an ontology-based model for information selection. The VLDB Journal 13 (1), 71–85.

Khan, S., Marvon, F., 2006. Identifying relevant sources in query reformulation. In: Proceedings of the 8th International Conference on Information Integration and Web-based Applications & Services (iiWAS). Yogyakarta, Indonesia, pp. 99–130.

Kobayashi, M., Takeda, K., 2000. Information retrieval on the web. ACM Computing Surveys (CSUR) 32 (2), 144–173.

Leacock, C., Miller, G.A., Chodorow, M., 1998. Using corpus statistics and wordnet relations for sense identification. Computational Linguistics 24 (1), 147–165.

Lee, C.Y., Soo, V.W., 2005. Ontology-based information retrieval and extraction. In: Proceedings of 3rd International Conference on Information Technology: Research and Education (ITRE). IEEE, pp. 265–269.

Montes-Y-Gomez, M., Lopez-Lopez, A., Gelbukh, A.F., 2000. Information retrieval with conceptual graph matching. In: Proceedings of 11th International Conference on Database and Expert Systems Applications (DEXA). London, UK, pp. 312–321.

Poole, J., Campbell, J.A., 1995. A novel algorithm for matching conceptual and related graphs. In: Proceedings of the 3rd International Conference on Conceptual Structures: Applications, Implementation and Theory, vol. 954, Lecture Notes in Computer Science. Springer-Verlag, pp. 293–307.

Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19 (1), 17–30.

Resnik, P., 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130.

Rinaldi, A.M., 2009. An ontology-driven approach for semantic information retrieval on the web. ACM Transactions on Internet Technology (TOIT) 9 (3), 10.

Rodriguez, M.A., Egenhofer, M.J., 2003. Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering 15 (2), 442–456.

Shah, U., Finin, T.W., Joshi, A., 2002. Information retrieval on the semantic web. In: Proceedings of the ACM International Confer-

ence on Information and Knowledge Management (CIKM). McLean, VA, USA, pp. 461–468, November.

Tversky, A., 1977. Features of similarity. Psychological Review 84 (4), 327–352.

Uschold, M., Gruninger, M., 2004. Ontologies and semantics for seamless connectivity. ACM SIGMod Record 33 (4), 58–64.

Varelas, G., Voutsakis, E., Raftopoulou, P., et al, 2005. Semantic similarity methods in wordnet and their application to information retrieval on the web. In: Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management. ACM, pp. 10–16.

Wu, Z., Palmer, M., 1994. Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. ACL, pp. 133–138.

Yang, Che-Yu, Wu, Shih-Jung, 2011. A wordnet based information retrieval on the semantic web. In: Networked Computing and Advanced Information Management (NCM), 7th International Conference. IEEE, pp. 324–328.

Zhong, J., Zhu, H., Li, J., Yu, Y., 2002. Conceptual graph matching for se-mantic search. In: Proceedings of 10th International Conference on Conceptual Structures: Integration and Interfaces (ICCS). Borovets, Bulgaria, pp. 92–196, July.

Zhu, H., Zhong, J., Li, J., Yu, Y., 2002. An approach for semantic search by matching RDF graphs. In: Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference. AAAI Press, pp. 450–454.