Query and Answering on Computer Science Documents base on Ontology

Thanyaporn Boonyoung  
Faculty of Information Technology  
King Mongkut’s University of Technology North Bangkok  
Bangkok  
Bangkok, Thailand  
thanyaporn.b@rmut.ac.th

Anirach Mingkhwan  
Faculty of Industrial and Technology Management  
King Mongkut’s University of Technology North Bangkok  
Bangkok  
Prachinburi, Thailand  
anirach@ieee.org

Abstract—The limitations of information retrieval systems traditionally are not enable us to understand the meaning of user’s query. Most existing assignment based on the information retrieval term frequency (tf) that appear in the document. In this paper, we have presented the answering for Computer Science documents using Computer Science Ontology that significantly improve the accuracy of recall. The results can answer about popular researching of Computer Science Knowledge area in country and can shows document similarity score with Computer Science areas. For instance, the most similarity score of the document #115 is Software Engineering and Social and Professional Issue (46.512%).

Keywords—Document Similarity; Query and Answer; Ontology;

I. INTRODUCTION

The rapidly growth of information technology are improved to brought new challenges for research paper retrieval. While most of existing research paper databases use matching search technology based on exactly keyword and purely statistical techniques, this retrieval processing does not consider any semantic information. The limitations of pure keyword based search are not enable us to understand the meaning of keyword and the ability to distinguish between relevant and irrelevant keywords that became important for user’s query. Semantic search can solved the limitations of keyword-based search by semantic measuring among words, concepts or ontologies and became methods to understanding keyword meaning. The nature of keyword meaning is what it refer to and reference to things (entities, relations).

In some situations in which semantic search can help, such as When user can’t remember exactly what they was looking for or can’t remember how to spell the search term. For example, when user interested research document about “decision making”, keywords similarity or keywords related cannot considered e.g., decision tree, decision making and supporting decision making that similarity keywords are not able to include in user’s query. Another problem is the difference in the word usage, for example, United Kingdom, UK and England may refer to the same country. Semantic search uses language processing to assess the “meaning” of content to return more relevant results. Semantic search technique, as an application of Semantic Web in the field of information retrieval, improve performance of retrieval.

Ontology engineering is involved with making representational choices that capture the relevant distinctions of a domain at the highest level of abstraction while still being as clear as possible about the meanings of terms. The importance of ontology operation is computing semantic similarity between terms such as mapping aligning, and integrating. The goal of ontology is to catch the knowledge in the correlation domain; provide common understanding about knowledge and determine the approval glossary in this domain; and produce the clear definition about these terminologies and the relations between the glossaries from the formalized pattern in different levels [1].

Several researches propose about ontology construction based on semantic similarity ontology matching and semantic indexing referring to semantic closeness or nearness. It computes similarity between different search terms. OWL (Web Ontology Language) defines any instance of a term in RDF (Resource Description Framework) triples instead of keywords. The form of RDF triples is submitted in the query which is expanded through synonyms and semantic neighborhood using the distance based approach [2]-[3].

In this paper, we adopt structure of research paper ontology and apply to semantic search using Computer Science ontology [4]-[6]. In addition, we compute document similarity with knowledge of Computer Science and proposed the process for finding popular keywords in country. The result is presented in infographics.

The rest of this paper is organized as follows: A review related work on document similarity measure in Section 2; Computer Science document ontology, ontology design, RDF triple and Computer Science knowledge taxonomy used in this work, Section 3; a query and answering process in Section 4; and the experiment results of our approach following by Conclusion in Section 5.
II. RELATED WORK

The recently-emerging ontology and the information retrieval are reviewed in this section.

A. Ontology

Boamergerg Aleman-Meza proposed SwetoDblp ontology of Computer Science publications in RDF from an XML document. The following guidelines for the creation of SwetoDblp are creation of URIs that can be easily recognized and/or reused on other applications or datasets; usage of existing vocabulary whenever is possible and integration of relationships and entities from additional data sources [7].

Furthermore, Tim Berners-Lee[8] recommended four principles in order to link data around the world below.

1. Use URIs as name for things (resources) in order to identify them.

2. Use HTTP URIs, so that people and machines can look up those URIs.

3. When someone looks up an URI, they provide useful information about the URI by using the standards such as RDF and SPARQL.

4. Include links to other resources in order to enable the discovery more data.

Sung Shun Weng designed ontology construction system architecture using information retrieval terms, such as term parsing, and calculating weight of related term[9].

Xiaomei Xu proposed query expansion and ontology construction algorithm query system for digital libraries. It has been used widely in natural language processing (NLP) and information retrieval applications. The core concept in WordNet was synsets, which was a set of synonyms [10].

The OWL, Web Ontology Language extends the RDFS vocabulary with additional resources that can be used to build more expressive ontologies for the Web. OWL introduces added restrictions regarding the structure and contents of RDF documents in order to make processing and reasoning more computationally decidable. OWL uses the RDF and RDFS, XML Schema datatypes, and OWL namespaces. The system provides a searching module to match between keyword and user’s query based on the search semantic algorithm, including applying the SPARQL[11], which is a query language to exact RDF.

B. Information Retrieval

Term-frequency (tf) Weight [12] in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term frequency regardless of the actual importance of that term in the document) to give a measure of the importance of the term t within the particular document can be calculated by formula (1). The matching scores cannot show what we want because the relevance document does not increase with more term frequency. So we use log frequency weighting (2) instead.

\[ tf_{ij} = f_{ij} / (\max f_{ij}) \]  
\[ w_{tr} = \begin{cases} 1 + \log tf_{ij}, & \text{if } tf_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \]

III. COMPUTER SCIENCE DOCUMENT ONTOLOGY

We set up structure of the computer science research paper ontology using a set of relations and a set of terms. In addition, we collected documents in field of Computer Science from Scopus database.

A. Ontology Design

We designed computer science research ontology. Ontologies are specifications of vocabularies and relations

1) Ontology Structure

The structure of the Computer Science paper ontology has been designed from six important parts on research document: Title, Author, Affiliation, Keyword, Conference Name, and Year. The structure is shown on Figure 1.

2) Define Vocabularies and relations

We made the usage of existing vocabulary or namespace whenever is possible such as dublin core, FOAF. For namespaces are shown on Table 1. Figure 2. An ontology fragment representing computer science research Entity terms and their relationships. In this graphical notation, ellipses denote terms, arrows denote relationships and cardinality.

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Namespace</th>
<th>Namespace URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>dc:Title</td>
<td><a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a></td>
</tr>
<tr>
<td>Author</td>
<td>foaf:person</td>
<td><a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a></td>
</tr>
<tr>
<td></td>
<td>swrc:author</td>
<td><a href="http://swrc.ontoware.org/ontology0">http://swrc.ontoware.org/ontology0</a></td>
</tr>
<tr>
<td>Affiliation</td>
<td>foaf:Organization</td>
<td><a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a></td>
</tr>
<tr>
<td></td>
<td>swrc:affiliation</td>
<td><a href="http://swrc.ontoware.org/ontology0">http://swrc.ontoware.org/ontology0</a></td>
</tr>
<tr>
<td>Keyword</td>
<td>dc:Subject</td>
<td><a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a></td>
</tr>
<tr>
<td>Year</td>
<td>dc:Date</td>
<td><a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a></td>
</tr>
<tr>
<td>Document Type</td>
<td>foaf:Document</td>
<td><a href="http://xmlns.com/foaf/0.1/TDocument">http://xmlns.com/foaf/0.1/TDocument</a></td>
</tr>
<tr>
<td>Reference</td>
<td>dc:Reference</td>
<td><a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a></td>
</tr>
</tbody>
</table>

Figure 1. Illustration of Computer Science Research Entity Concepts and Relationships
B. RDF Triple

We represented with three values for RDF is called the triple. The identifier for the row was called the domain (subject) of the triple. The identifier for the column was called the object property (predicate) of the triple and the value in the cell was called the range (object) of the triple. Table 2 shows example for the triple in article.

Table 2. Example : RDF Triple in the document

<table>
<thead>
<tr>
<th>Domain</th>
<th>Object Property</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class : dc:Title</td>
<td>dc:Subject</td>
<td>Class : dc:Subject</td>
</tr>
<tr>
<td>Class : foaf:Person</td>
<td>swrc:Author</td>
<td>Class : dc:Title</td>
</tr>
<tr>
<td>Class : foaf:Person</td>
<td>swrc:Affiliation</td>
<td>Class : foaf:Organization</td>
</tr>
</tbody>
</table>

As an example, Keywords in Research document about : “Data Model”. We extracted possible meaning of keyword based on Computer Science ontology. Finding synonym, hypernym/ hyponym meronym/holonym and related term shown that in Figure 5.  

**Hypernym Term**: Information Management  
**Hyponym Terms are**: Conceptual models, Object-oriented model, Relational data model, Semi structured data model

![Figure 2](image2.png)  
**Figure 2** Example of Concepts and Relationships Related a Research Paper

**B. Computer Science Knowledge Taxonomy**

Thanyaporn and Anirach (2012) adopted ontology for knowledge of computer science, reference from computer science curricula 2013 draft report [13] that had been endorsed by the Association for Computing Machinery (ACM) and IEEE Computer Society. Computer Science Curricula 2013 (CS2013), represented a comprehensive revision. 11 CS2013 redefined the knowledge units in Computer Science (CS).

![Figure 3](image3.png)  
**Figure 3** Knowledge Area of Computer Science

The last complete Computer Science curricular volume was 9 released in 2001 (CC2001), and an interim review effort concluded in 2008 (CS2008). The CS2013 Body of Knowledge was organized into a set of 18 Knowledge Areas (KAs) in Figure 3, 70 corresponding to topical areas of study in computing. The taxonomic hierarchy (ontology) model of computer science keywords/terms was organized in ls-A relationships (Hyponym/Hypernym) with more general terms (e.g. “Operating System and Digital Forensic”, “Information Management and Database System”) higher in Information Management taxonomy than more specific terms (e.g. “Object-oriented model”, “Indexing”). A keyword (term) may appear in more than one taxonomy, such as “Information Retrieval” term of Information Management and Intelligent Systems is shown in Figure 4. There was four levels, eighteen taxonomies and more than 200 terms.

![Figure 4](image4.png)  
**Figure 4** Illustration of Computer Science Ontology Hierarchy

![Figure 5](image5.png)  
**Figure 5** Illustration Data Model Context
IV. QUERY AND ANSWERING PROCESS

In previous experiment, we introduced a new weighting method based on semantic similarity using Computer Science ontology [4] for support semantic search in Computer Science document repositories. The next experiment, we proposed document similarity using Computer Science Ontology based on Edge Counting and N-Grams for a re-ranking process of search results. The last experiment, we represented using the vector-space model (cosine similarity) for semantic ranking [6]. In the model, we used Term-Frequency (TF) for compute document as vector and considered hierarchy relationship using Computer Science ontology for compute query as vector.

This section described the query and answering process to Computer Science Knowledge. In the document keyword set, a document was represented by a keyword vector, i.e., \( \text{document} = (\text{keyword}_1, \ldots , \text{keyword}_i, \ldots , \text{keyword}_n) \). We described the process for answering from two user’s questions.

- For the question 1, we computed by formula (3)

\[
\text{Sim}_{\text{CSarea}} = \frac{\sum \text{tf}_{t,d} \times 100}{N}
\]

(3)

where \( \text{Sim}_{\text{CSarea}} \) was Document Similarity in Area of Computer Science; \( \text{area} \) was the body of knowledge in Computer Science Ontology; \( \text{tf} \) was number of terms where the term \( t \) appeared in area of Computer Science Ontology; \( N = \text{Total number of term frequency in document} \); \( i \) was a term(keyword) in document;

V. EXPERIMENTS

In this section we summarized the main experiments and the results obtained in the study. The 1769 Computer Science Ontology and Computer Science documents were used to test the proposed system.

For example 1, user’s question: User want to find the document similarity of Document #115 to the Computer Science areas.

(Document Title : Decision making during a tendering procedure : Case studies of restricted European tenders in Architecture)

We also computed document similarity in Computer Science areas by formula (3), the results shown in Table 3 and figure 6.

Table 3 Document Similarity with Computer Science Areas

<table>
<thead>
<tr>
<th>No.</th>
<th>Computer Science Areas</th>
<th>Document Similarity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SE</td>
<td>46.51163</td>
</tr>
<tr>
<td>2</td>
<td>SP</td>
<td>46.51163</td>
</tr>
<tr>
<td>3</td>
<td>CN</td>
<td>34.88372</td>
</tr>
<tr>
<td>4</td>
<td>PL</td>
<td>23.25581</td>
</tr>
<tr>
<td>5</td>
<td>IS</td>
<td>13.95349</td>
</tr>
<tr>
<td>6</td>
<td>IAS</td>
<td>11.62791</td>
</tr>
<tr>
<td>7</td>
<td>IM</td>
<td>9.302326</td>
</tr>
<tr>
<td>8</td>
<td>SF</td>
<td>4.651163</td>
</tr>
</tbody>
</table>

For example 2, user’s question: What is a Computer Science knowledge areas that most popular researching in England?

For Semantic Search, We considered related word that found United Kingdom Set = \{United Kingdom, UK, England\}

Some authors may identified their difference country name, such as England, United Kingdom and UK.

Table 4 Number of Document and Evaluation

<table>
<thead>
<tr>
<th>Country Name</th>
<th>Number of Doc</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>United Kingdom</td>
<td>80</td>
<td>0.684</td>
<td>1</td>
<td>0.812</td>
</tr>
<tr>
<td>UK</td>
<td>36</td>
<td>0.308</td>
<td>1</td>
<td>0.471</td>
</tr>
<tr>
<td>England</td>
<td>1</td>
<td>0.008</td>
<td>1</td>
<td>0.017</td>
</tr>
<tr>
<td>United Kingdom Set</td>
<td>117</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

After retrieving relevant documents, we also count all maximum keyword that found of all, the results show in Table 4,5 and figure 7.
Table 5 Total of Document Related in Computer Science Knowledge Areas

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Software Engineering</td>
<td>SE</td>
<td>25</td>
<td>Social and Professional Issues</td>
<td>SP</td>
<td>3</td>
</tr>
<tr>
<td>Operating Systems</td>
<td>OS</td>
<td>15</td>
<td>Platform-based Development</td>
<td>PBD</td>
<td>2</td>
</tr>
<tr>
<td>Computational Science</td>
<td>CN</td>
<td>13</td>
<td>Parallel and Distributed Computing</td>
<td>PD</td>
<td>2</td>
</tr>
<tr>
<td>Human-Computer Interaction</td>
<td>HCI</td>
<td>9</td>
<td>Systems Fundamentals</td>
<td>SF</td>
<td>2</td>
</tr>
<tr>
<td>Information Management</td>
<td>IM</td>
<td>8</td>
<td>Architecture and Organization</td>
<td>AR</td>
<td>1</td>
</tr>
<tr>
<td>Intelligent Systems</td>
<td>IS</td>
<td>6</td>
<td>Graphics and Visual Computing</td>
<td>GV</td>
<td>1</td>
</tr>
<tr>
<td>Programming Languages</td>
<td>PL</td>
<td>4</td>
<td>Security and Information Assurance</td>
<td>SIA</td>
<td>1</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

In this paper, we have presented the answering for Computer Science documents using Computer Science Ontology, such as, document similarity with Knowledge of Computer Science and the finding a popular terms in Country.

The results in table 3 show the percentile of document similarity with Computer Science knowledge areas. The similarity score of the document #115 are Software Engineering (46.512%), Social and Professional Issue (46.512%), Computational Science (34.884%), Programming Languages(23.256%), Intelligent Systems (13.953%), Security and Information Assurance (11.628%), Information Management(9.302%) and Systems Fundamentals (4.651%).

The results in Table 4 show the semantic similarity measure significantly improve the accuracy (F-measure of 1) and Table 5 show the total of document related with Computer Science Knowledge in England.

Future studies should apply the proposed method to applications of semantic search using Computer Science ontology, and display the results using information visualization technique.

ACKNOWLEDGEMENTS

I gratefully acknowledge the funding received toward my research from Rajamangala University of Technology Thanyaburi. I would like to thank my advisor, Asst. Prof. Dr. Anirach Mingkhwan who has supported me throughout this period Additionally, thank you the lecturers and researchers for sharing their knowledge on Computer Science for the Computer Science Ontology Evaluation.

REFERENCES