Semantic Search: Document Ranking and Clustering Using Computer Science Ontology and N-Grams

Thanyaporn Boonyoung[1], Anirach Mingkhwan[2]
[1]Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
[2]Faculty of Industrial and Technology Management
King Mongkut's University of Technology North Bangkok
Prachinburi, Thailand
thanyaporn.b@rmutt.ac.th, anirach@ieee.org

Journal of Digital Information Management

ABSTRACT: *Semantic similarity has become an important tool and widely been used to solve traditional Information Retrieval problems. This study adopts ontology of computer science and proposes an ontology indexing weight based on Wu and Palmer's edge counting measure and uses the N-grams method for computing a family of word similarity. The study also compares the subsumption weight between Hliaoutakis and Nicola's weight and query keywords (Decision Making, Genetic Algorithm, Machine Learning, Heuristic). A probability value (p-values) from the t-test (p = 0.105) is higher 0.05, which indicates the evidence of no of no significant differences between the two weights methods. The experimental results show the new keyword-keyword similarity matrix scores that compute from hierarchical relationship weight based on Computer Science ontology and string matching (tri-grams) for computing of string of keyword. We computed the document-document similarity matrix scores using our keyword similarity matrix scores and compared them with the keyword matching weights using Dice coefficient method. In addition, this paper, we presented a new document semantic ranking process for the semantic ranking that proposes a new weight of query term in the document based on Computer Science Ontology weight. The experimental results show that the new document similarity score between a user's query and the paper suggests that the new measures were effectively ranked.*

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing Indexing methods]; H.3.3 [Information Search and Retrieval]

General Terms: Semantics, Information Search, Ontology

Keywords: Document Ranking, Document Similarity, Vector Space Model, Computer Science Ontology, Ontology Indexing, N-Grams

Received: 18 July 2014, Revised 27 August 2014, Accepted 3 September 2014

1. Introduction

In the last few years, the amount digital documents has been increasing that is electronically available has increased rapidly and, as a result, the development of information storage and retrieval systems has become a significant challenge. A good Information Retrieval system should retrieve only those relevant documents from large databases and appropriate for user's query, not a lot of unnecessary data.

A document ranking is an ordering of the documents retrieved that reflects the relevance of the documents to the user query. Document ranking algorithm is one of importance process in retrieval model to efficiently present the search results and the top ranked documents would have highest similarity score. One of the simplest ranking functions is computed by the *tf-idf* algorithm [1], based on the value of the keyword that is stored by the frequency of the keyword that is appear in the database system and number of times the word will be appears in the document can be the value of the term vector in the document. The traditional ranking technique such as Dice coefficient, Jaccard coefficient and Cosine coefficient are common similarity measures based on vector space model [1].

The traditional document ranking (keyword-based search)

are those containing user specified keywords but they are not consider various points including context of search (Variation of words). Many researches improved document ranking based on ontology approach that and returns documents of ontology values that satisfy the query [2-8]. The term frequency *(tf)* in the given document is simply the number of times a given term appears in that document. Although the number of times that the term occurrence is more relevant, but not meant rank documents according to their proximity to user's query. So our key idea focuses on sematic similarity measure.

Ontology is an explicit specification of shared conceptualization [6]. A number of ontology libraries and search engines are in existence to facilitate retrieval of potentially relevant ontologies and provides a domain-related ontology to depict the real world applications. There is a set of standard web ontology language (OWL) which is based on RDF model to describe the concepts explicitly with their relationship. Many semantic similarity researches use domain ontologies to consider the hierarchical structure and compute the relationship between terms.

Most semantic similarity researches compute the similarity between concept using Wordnet[1], which is an online lexical and can also be seen as an ontology. It contains terms, organized into taxonomic hierarchies. Although Wordnet is widely used, it is still limited, and does not offer specific domain. The results are not in an hierarchical structure and researchers cannot compute the relationship between their terms. Consequently, specific domains such as the United States National Library of Medicine (NLM)[2] offers a hierarchical categorization of medical and biological terms called Medical Subject Headings (MeSH)[3] [2,4] facilitates searching. The domain of Computer Science field is specified as the medical and biological domain that some terms and relations not offered in Wordnet taxonomic hierarchies. For example, when users are interested in topic about "*decision making*", their hierarchical structures do not propose in Wordnet, in which decision making is related to support decision making, decision trees and many others in the knowledge of Computer Science.

In this paper, we introduce a new weighting method based on semantic similarity using Computer Science ontology [12] for support semantic search in Computer Science document repositories. The research problem of improving relevance in search and ranking of documents required techniques that consider the semantic of user's query. The search system took advantage of ontology based semantic annotation and it included the weights of document structure in ranking.

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

[2]http://www.nlm.nih.gov

[3]http://www.nlm.nih.gov/mesh

1.1 Overview of Paper

The rest of this paper is organized as follows: A review related work on document similarity measure in Section 2.; Computer Science Ontology, taxonomic hierarchy used in this work, as well as a comparison of semantic similarity methods with standard weight on the ontology in Section 3.; a semantic ranking approach in Section 4.; and the experiment results of our approach following by Conclusion in Section 5.

2. Related Work

2.1 Document Ranking

One of importance key question in document retrieval is how to rank documents based on their degrees of relevance to a query. Much effort has been placed on the development of ranking functions. Traditionally, document retrieval methods only use a small number of features (e.g., term frequency, inversed document frequency, and document length). Thus, it is possible to empirically tune the parameters of ranking functions. We refer to the method as Ranking Vector-Space Model in this paper.

2.2 Vector Space Model

In the statistically based vector-space model, a document is conceptually represented by a vector of keywords extracted from the document, with associated weights representing the importance of the keywords in the document and within the whole document collection; likewise, a query is modeled as a list of keywords with associated weights representing the importance of the keywords in the query.

2.3 Cosine similarity

Cosine similarity [1] is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

$$sim(d_j, d_k) = \frac{\vec{d_j} . \vec{d_k}}{|\vec{d_j}||\vec{d_k}|} = \frac{\sum_{i=1}^{n} w_{i,j}, w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2 \sum_{i=1}^{n} w_{i,j}^2}} \quad (1)$$

2.4 Term-frequency *(tf)* Weight

The Term-frequency (*tf*) Weight [1] 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_i within the particular document can be calculated by formula (2) but 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 (3) instead.

$$tf_{i,j} = f_{i,j} / (\max_k f_{k,j}) \quad (2)$$

$$w_{t,d} = \begin{cases} 1 + log_{10} tf_{t,d} & if\ tf_{t,d} > 0 \\ 0 & , Otherwise \end{cases} \quad (3)$$

2.5 Ontology

Ontology is a shared conceptualization of a domain. Ontology is a set of definitions in a formal language for terms describing the world. Ontology is a specification of a conceptualization that is designed for reuse across multiple applications and implementations. A specification of a conceptualization is a written, formal description of a set of concepts and relationships in a domain of interest.

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

Boanergers Aleman-Meza [10] proposed SwetoDblp ontology of Computer Science publications in RDF from an XML document. The following guidelines for 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.

2.6 Edge counting methods

Wu and Palmer [14] presented a similarity measure (4) to finding the most specific common concept that subsumed both of the concepts being measured. It could be calculated by calculating the length of the path linking the ontology keyword in and ontology. Considering that the similarity *(sim)* between a pair of concepts in an upper level of the taxonomy was smaller than a pair in a lower level, they proposed a path-based measure that also took into account the depth of the concepts in the hierarchy.

$$sim_{w\&p}(a,b) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (4)$$

2.7 N-Grams

N-Grams Grzegorz Kondrak [17] formulates a family of word similarity measures based on N-grams, computes word similarity based on N-grams that can be calculated by formula (5), Sembok and Bakar [18] and proposes N-grams string similarity and stemming on Malay documents.

$$\frac{2 \times |n - \text{Keyword}(QW) \cap n\text{-Keyword}(OW)|}{|n - \text{Keyword}(QW)| + |n - \text{Keyword}(OW)|} \quad (5)$$

3. Computer Science Ontology

This section describes taxonomic hierarchy (Computer Science ontology) and ontology indexing weight.

3.1 Computer Science Taxonomy

Thanyaporn and Anirach (2012) studied and adopted ontology for knowledge of computer science, reference from computer science curricula 2013 draft report[4] that has 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). The last complete Computer Science curricular volume was 9 released in 2001 (CC2001), and an interim review effort concluded in 2008 (CS2008)[5].

The CS2013 Body of Knowledge was organized into a set of 18 Knowledge Areas (KAs) in Figure 1, 70 corresponding to topical areas of study in computing. The taxonomic hierarchy (ontology) model of computer science keywords(terms) was organized in Is-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*" was term of Information Management and Intelligent Systems is shown in Figure 1. There are four levels, eighteen taxonomies and more than 200 terms.

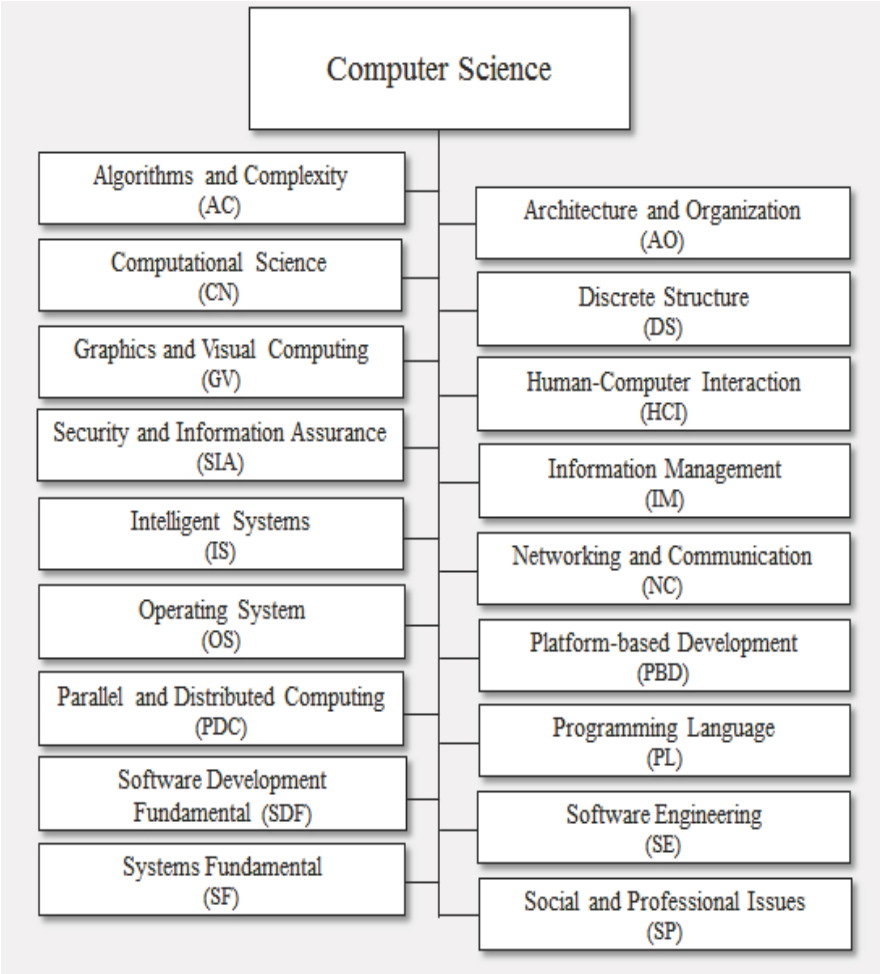
3.2 Ontology Indexing

Thanyaporn and Anirach (2012,2014) studied and proposed the hierarchy weight of subsumption (hypernym/hyponym or meronym/holonym hierarchy) in Computer Science ontology using Wu and Palmer measure [14]. These weights are shown in Table 1. The study compares the our subsumption weight with Hliaoutakis et al.[15] and Stoke's weight [16] with four query keywords (Decision Making, Genetic Algorithm, Machine Learning, Heuristic. A probability value (p-values) from the t-test (p = 0.105) is higher 0.05 and indicates no significant evidence that our weight based on Wu and Palmer Measure and their methods were not different.

For example, if query keyword(QK) and ontology keyword (OK) are the same word or synonymous, the weight is assigned as 1. For example, '*Database*' and '*DB*', '*Information Retrieval*' and '*IR*', 'Entity-Relationship' and '*E-R*' in different documents express the same meaning. The weight for '*Decision Making*' with '*Team organization*' is 0.75 Both keywords are members in Software Engineering/Software Project Management.

[1] http://ai.stanford.edu/users/sahami/CS2013

[2] http://www.acm.org/education/curricula/ComputerScience2008.pdf



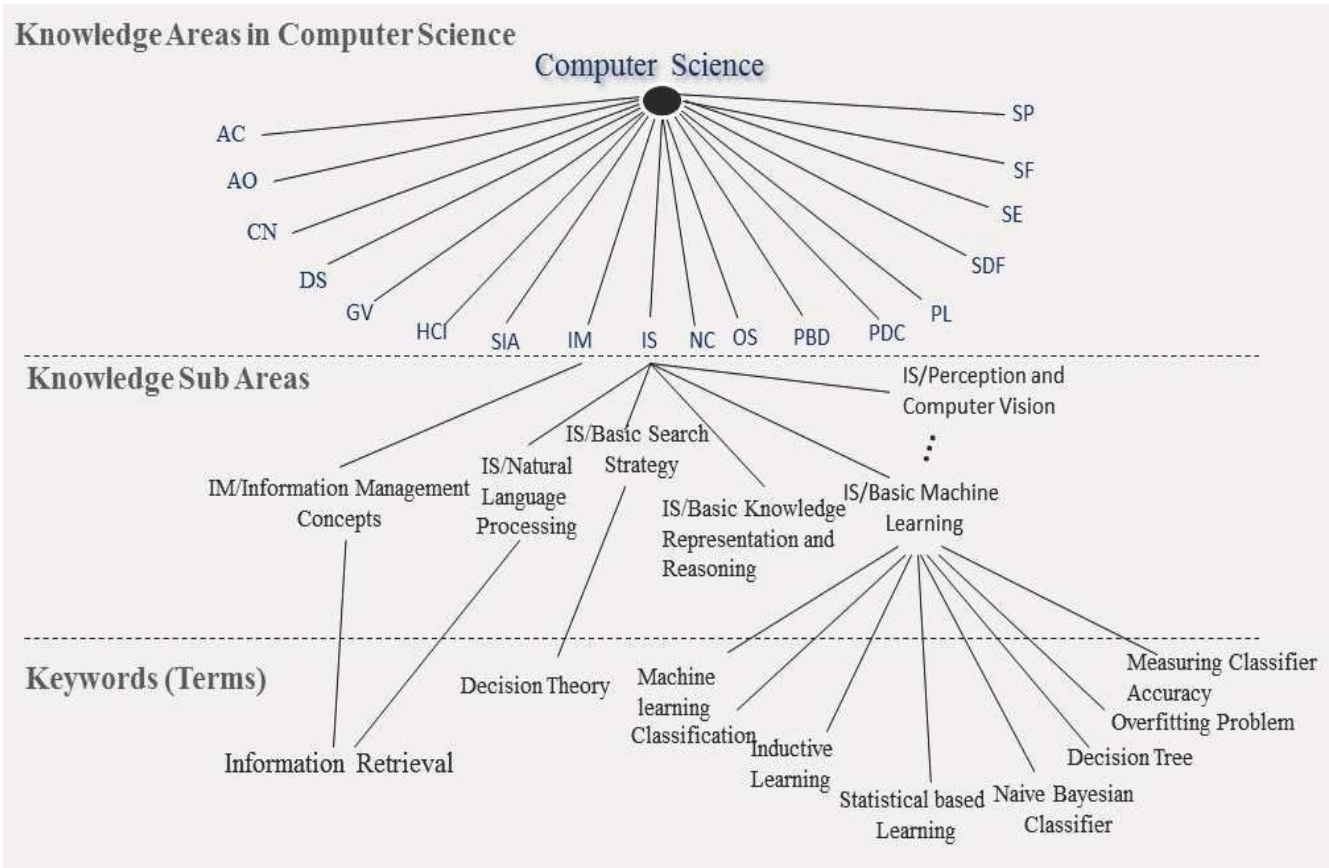


Figure 1. Knowledge Area of Computer Science

Relationship Type	Weights
Repetition /Synonymy	1
Same sub area	0.75
Same area	0.5
Term or Keyword on CS Ontology	0.25
not found In CS ontology	0

Table 1. Computer Science ontology Weight based on Wu and Palmer Measure

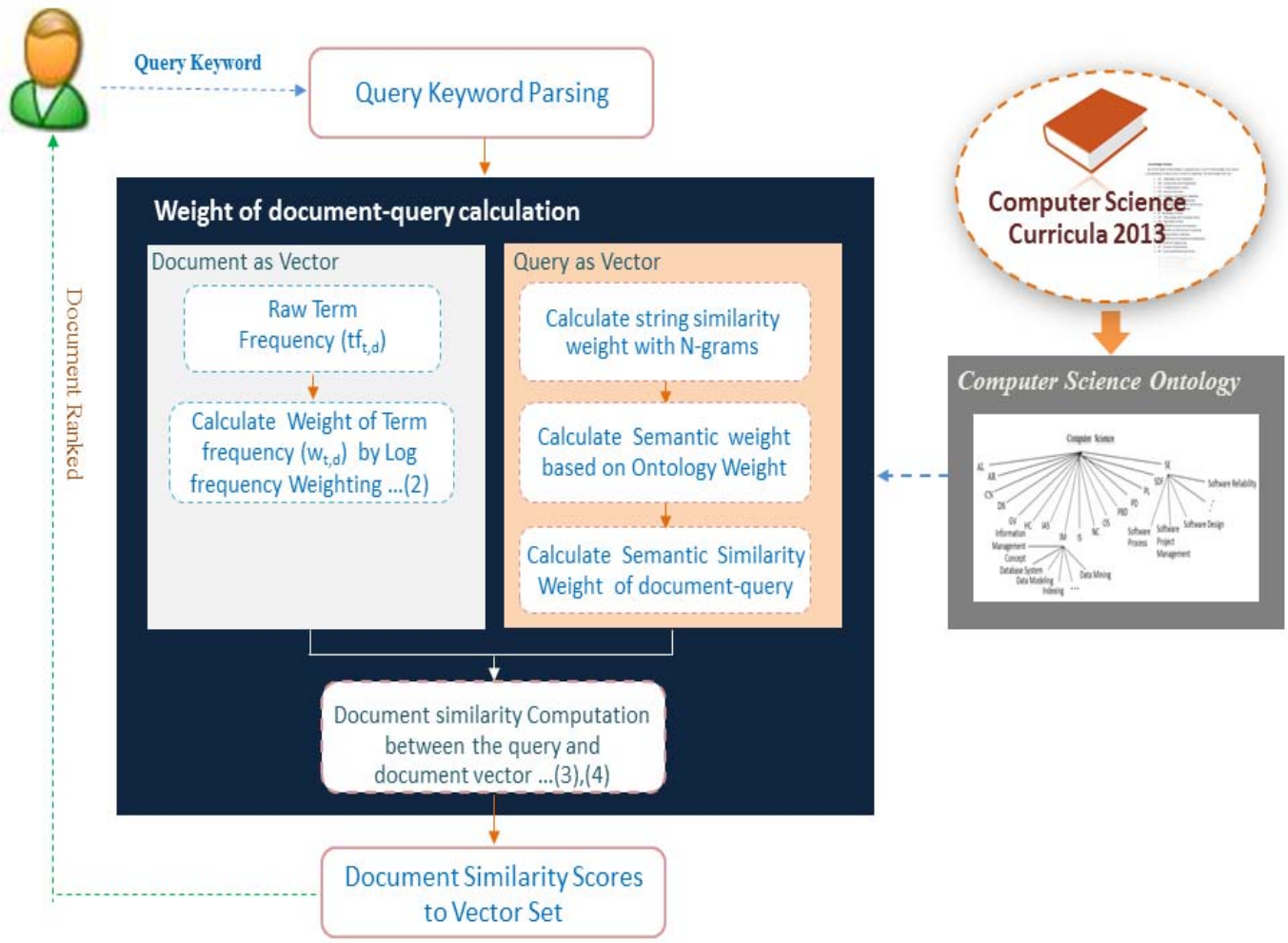


Figure 2. Illustration Document Semantic Ranking Process

4. Semantic Search Process

4.1 The Semantic Ranking Approach

This section describes the document semantic score to rank documents according to query-document matching scores that use *tf-weight* and Computer Science ontology weight. In the document keyword set, a document is represented by a keyword vector, i.e., $document = (keyword_1 \ldots keyword_j, \ldots, keyword_n)\ (1 \le i \le n)$.

We will present in the following these process, specifically we will describe the process of document semantic ranking. Figure 2 shows the current process of document semantic ranking.

- The document as vector process is a weight of term frequency that computed by formula (3).
- The query as vector process is a weight of term between query-document that using Computer Science ontology weight (table 2).
- The document similarity computation between the query and document vector by Cosine similarity measure (1) that is a measure of similarity between two vectors of an inner product space.
- Final similarity score between query and document by formula (6).

$$Simscore\ (d) = w_{t,d} * w_{t,q} \quad (6)$$

4.2 The Document similarity Approach

This section described about the document similarity score and showed that the Term(Keyword)-Term(Keyword) similarity matrix and the Document-Document similarity matrix. In the document keyword set, a document was represented by a keyword vector, i.e., $document = (keyword_1, \ldots, keyword_j, \ldots, keyword_n)$ $(1 \le i \le n)$. The first compute keyword was weighted by using Computer Science ontology. For each keyword KW_i and document D_j, a weight W_{ij}, indicating how strongly the term represents the document.

4.2.1 Calculating the Keyword-Keyword similarity matrix Scores

Thanyaporn and Anirach (2014) proposed two features for the Keyword-Keyword similarity computing [19]:

• The Keyword-Keyword Similaity Matrix Scores of Subsumption (Hypernym/Hyponym Hierarchy) from Computer Science Ontology, shown in Table 2.

• The string similarity with N-grams: The measure based on tri-grams (5) [17], are defined as the ratio of the number of N-grams (Tri-grams) that are shared by two strings and the total number of N-grams in both strings.

$$SWK(KW_i, KW_j) = KWW_{n\text{-}grams} + KWW_{Cs\text{-}Onto} \quad (7)$$

where $SWK(KW_i, KW_j)$ was the total Keyword-Keyword Similaity Matrix Scores ; $KWW_{n\text{-}grams}$ was similarity weight between Keyword-Keyword in documents based on N-grams; $KWW_{Cs\text{-}Onto}$ was similarity weight between Keyword-Keyword in documents based on Computer Science ontology weight (Table 2); i, j indicated the keyword number;

4.2.2 Calculating the Document-Document similarity matrix Scores

The use of Dice coefficient [20] in semantic similarity the Document-Document similarity matrix scores between two documents was calculated as follows:

$$Dss(D_i, D_j) \cong \frac{\sum_{k=1}^{n} KWW_{ki} KWW_{kj}}{\alpha \sum_{k=1}^{n} KWW_{ki}^2 + (1-\alpha) \sum_{k=1}^{n} KWW_{kj}^2} (\alpha = \frac{1}{2}) \quad (8)$$

5. Experiment

This section we summarizes the main experiments and the results obtained in the study. To test the proposed system, this present study used Computer Science Ontology [12] and Computer Science documents, which consist of 1769 documents.

5.1 Results of Semantic Ranking

As an example, '*Decision Making*' is the query for the experiment. First, the research found that the knowledge area of 'Decision Making' in Computer Science ontology and its Computer Science area are Software Engineering and Social and Professional Issues. This paper also computes the document semantic similarity scores by cosine similarity measure (3). The result in Table 2 shows that the term frequency document similarity scores of documents and the query ("*Decision Making*"). This paper also computes the document semantic similarity scores by cosine similarity measure (3). The result in Table 2 shows that the term frequency document similarity score of documents and the query ("*Decision Making*").

Document				Query	Product
Terms	***tf-raw***	***tf-wght***	**n'lized**	**Semantic weight**	
Architecture	2	1.3010	0.2433	0.25	0.0608
client	2	1.3010	0.2433	0.5	0.1217
Decision	6	1.7782	0.3325	1	0.3325
Design	3	1.4771	0.2762	0.5	0.1381
making	6	1.7782	0.3325	1	0.3325
Public	5	1.6990	0.3177	0.25	0.0794
quality	3	1.4771	0.2762	0.75	0.2072
Tender	5	1.6990	0.3177	0	0.0000
Case	3	1.4771	0.2762	0.5	0.1381
European	3	1.4771	0.2762	0	0.0000
procedure	2	1.3010	0.2433	0.25	0.0608
restricted	1	1.0000	0.1870	0	0.0000
studies	2	1.3010	0.2433	0	0.0000
Similarity Score (*Semantic Weight*) :					**0.7864**
Similarity Score (*Non-Semantic Weight*) :					**0.5242**

Table 2. Term Weight (Query,Documents PID#115)

Document				Query	Product
Terms	***tf-raw***	***tf-wght***	**n'lized**	**Semantic weight**	
Brief	3	1.4771	0.2967	0	0.0000
client	6	1.7782	0.3571	0.5	0.1786
Culture	7	1.8451	0.3706	0	0.0000
Decision	9	1.9542	0.3925	1	0.3925
process	2	1.3010	0.2613	0.75	0.1960
Public	5	1.6990	0.3412	0.25	0.0853
User	2	1.3010	0.2613	0.5	0.1307
impact	2	1.3010	0.2613	0	0.0000
strategic	4	1.6021	0.3218	0.25	0.0804
values	2	1.3010	0.2613	0.5	0.1307
Similarity Score (*Semantic Weight*) :					**0.4899**
Similarity Score (*Non-Semantic Weight*) :					**0.2840**

Table 3. Term Weight (Query,Documents PID#1706)

5.2 Results of Document similarity

For example, 'Decision Making' was the query for the experiment. First, the research found that the knowledge area of 'Decision Making' in Computer Science ontology and its Computer Science area were Software Engineering and Social and Professional Issues. As a result, the study computed both the distance area of the query and the document keywords (see Figure 3). This paper also computed the string similarity with N-grams and used Wu

and Palmer measure to find the distance in Computer Science Ontology describing in section 3. The result in Table 5, 6 and 7 showed the document similarity matrix scores of documents ("*Decision Making*").

		Similarity Score (Using CS ontology Weight)		Similarity Score (Non-Semantic Weight)		
No.	**Paper**	**Cosine**	**Sum-Score**	**No.**	**Cosine**	**Sum-Score**
1	852	0.8438	1.4616	1	0.5447	0.7703
2	115	0.7864	1.4712	3	0.5242	0.7413
3	584	0.7863	1.6090	7	0.4703	0.6651
4	1055	0.6560	1.0110	4	0.4048	0.5725
5	1700	0.6361	0.96732	2	0.4009	0.5670
6	553	0.6312	0.9854	5	0.3925	0.3925
7	1455	0.6056	1.0155	9	0.3831	0.5418
8	847	0.5912	0.9230	8	0.3640	0.5147
9	1706	0.4899	1.1941	6	0.2840	0.4017
10	1506	0.3328	1.3729	10	0.1848	0.3696

Table 4. A comparative of Document Semantic Ranking between Semantic Weight and Common Weight

Document Keyword	Decision making	Architecture	Design quality	Public clients	Tendering	Mobile robots	Sensor virtualization	Situation awareness	Testbed	Urban traffic	Cordon and search	Petri Nets	Robustness	Critical decision Making	Gaming	Learning	Training	Video game	Medical decision making	Interaction design	Computer-supported cooperative work	Multi-disciplinary team meetings
Decision making	2	0.50	0.67	0.50	0.21	0.31	0.39	0.40	0.00	0.31	0.60	0.50	0.25	0.89	0.48	0.46	0.71	0.25	0.89	0.51	0.25	0.61
Architecture	0.50	2	0.75	0.75	0.00	0.57	0.50	0.56	0.00	0.25	0.42	0.50	0.25	0.50	0.50	0.50	0.50	0.57	0.50	0.25	0.29	0.50
Design quality	0.67	0.75	2	0.75	0.00	0.25	0.40	0.30	0.00	0.31	0.30	0.25	0.25	0.63	0.25	0.25	0.50	0.25	0.64	0.67	0.25	0.57
Public clients	0.50	0.75	0.75	2	0.00	0.62	0.25	0.30	0.00	0.37	0.55	0.45	0.32	0.54	0.25	0.25	0.50	0.50	0.50	0.50	0.29	0.54
Tendering	0.21	0.00	0.00	0.00	2	0.00	0.00	0.00	0.20	0.07	0.00	0.00	0.00	0.15	0.32	0.29	0.38	0.00	0.16	0.00	0.00	0.13
Mobile robots	0.31	0.57	0.25	0.62	0.00	2	0.25	0.30	0.00	0.50	0.75	0.39	0.54	0.30	0.75	0.50	0.25	0.82	0.30	0.75	0.29	0.58
Sensor virtualization	0.39	0.50	0.40	0.25	0.00	0.25	2	0.80	0.00	0.30	0.68	0.25	0.25	0.37	0.50	0.50	0.25	0.61	0.37	0.43	0.34	0.50
Situation awareness	0.40	0.56	0.30	0.30	0.00	0.30	0.80	2	0.00	0.30	0.69	0.31	0.49	0.37	0.50	0.50	0.25	0.25	0.37	0.44	0.31	0.28
Testbed	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	2	0.08	0.07	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.08	0.09
Urban traffic	0.31	0.25	0.31	0.37	0.07	0.50	0.30	0.30	0.08	2	0.55	0.25	0.25	0.30	0.25	0.25	0.48	0.25	0.30	0.30	0.25	0.29
Cordon and search	0.60	0.42	0.30	0.55	0.00	0.75	0.68	0.69	0.07	0.55	2	0.50	0.50	0.37	0.50	0.56	0.25	0.25	0.33	0.35	0.38	0.50
Petri Nets	0.50	0.50	0.25	0.45	0.00	0.39	0.25	0.31	0.00	0.25	0.50	2	0.33	0.25	0.75	0.25	0.25	0.75	0.25	0.25	0.25	0.33
Robustness	0.25	0.25	0.25	0.32	0.00	0.54	0.25	0.49	0.00	0.25	0.50	0.33	2	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.29
Critical decision Making	0.89	0.50	0.63	0.54	0.15	0.30	0.37	0.37	0.00	0.30	0.37	0.25	0.25	2	0.42	0.41	0.66	0.25	1.80	0.45	1.03	0.84
Gaming	0.48	0.50	0.25	0.25	0.32	0.75	0.50	0.50	0.00	0.25	0.50	0.75	0.25	0.42	2	0.83	0.58	1.04	0.42	0.50	0.25	0.54
Learning	0.46	0.50	0.25	0.25	0.29	0.50	0.50	0.50	0.00	0.25	0.56	0.25	0.25	0.41	0.83	2	0.65	0.50	0.41	0.50	0.25	0.54
Training	0.71	0.50	0.50	0.50	0.38	0.25	0.25	0.25	0.11	0.48	0.25	0.25	0.25	0.66	0.58	0.65	2	0.25	0.66	0.25	0.25	0.59
Video game	0.25	0.57	0.25	0.50	0	0.82	0.61	0.25	0.00	0.25	0.25	0.75	0.25	0.25	1.04	0.50	0.25	2	0.25	0.50	0.29	0.50
Medical decision making	0.89	0.50	0.64	0.50	0.16	0.30	0.37	0.37	0.00	0.30	0.33	0.25	0.25	1.80	0.42	0.41	0.66	0.25	2	0.46	0.25	0.91
Interaction design	0.51	0.25	0.67	0.50	0	0.75	0.43	0.44	0.00	0.30	0.35	0.25	0.25	0.45	0.50	0.50	0.25	0.50	0.46	2	0.32	0.28
Computer supported cooperative work	0.25	0.29	0.25	0.29	0	0.29	0.34	0.31	0.08	0.25	0.38	0.25	0.25	1.03	0.25	0.25	0.25	0.29	0.25	0.32	2	0.25
Multi-disciplinary team meetings	0.61	0.50	0.57	0.54	0.13	0.58	0.50	0.28	0.09	0.29	0.50	0.33	0.29	0.84	0.54	0.54	0.59	0.50	0.91	0.28	0.25	2

Table 5. Document-Document Similarity Matrix Scores (Keyword Matching)

Paper ID	115	847	1055	553	584	Query
115	1	0.1818	0.2222	0	0	0.3333
847	0.1818	1	0.2	0	0	0.2857
1055	0.2222	0.2	1	0	0	0.4000
553	0	0	0	1	0	0
584	0	0	0	0	1	0
Query	**0.3333**	**0.2857**	**0.4000**	**0**	**0**	1

Table 6. Document-Document Similarity Matrix Scores (Keyword Matching)

Paper ID	115	847	1055	553	584	Query
115	1	0.4346	0.6853	0.3990	0.4334	0.8231
847	0.4346	1	0.4551	0.3574	0.3425	0.6424
1055	0.6853	0.4551	1	0.3944	0.3578	0.8290
553	0.3990	0.3574	0.3944	1	0.4999	0.7167
584	0.4334	0.3425	0.3578	0.4999	1	0.6959
Query	**0.8231**	**0.6424**	**0.8290**	**0.7167**	**0.6959**	**1**

Table 7. Document-Document Similarity Matrix Scores (Using Our Semantic Weight)

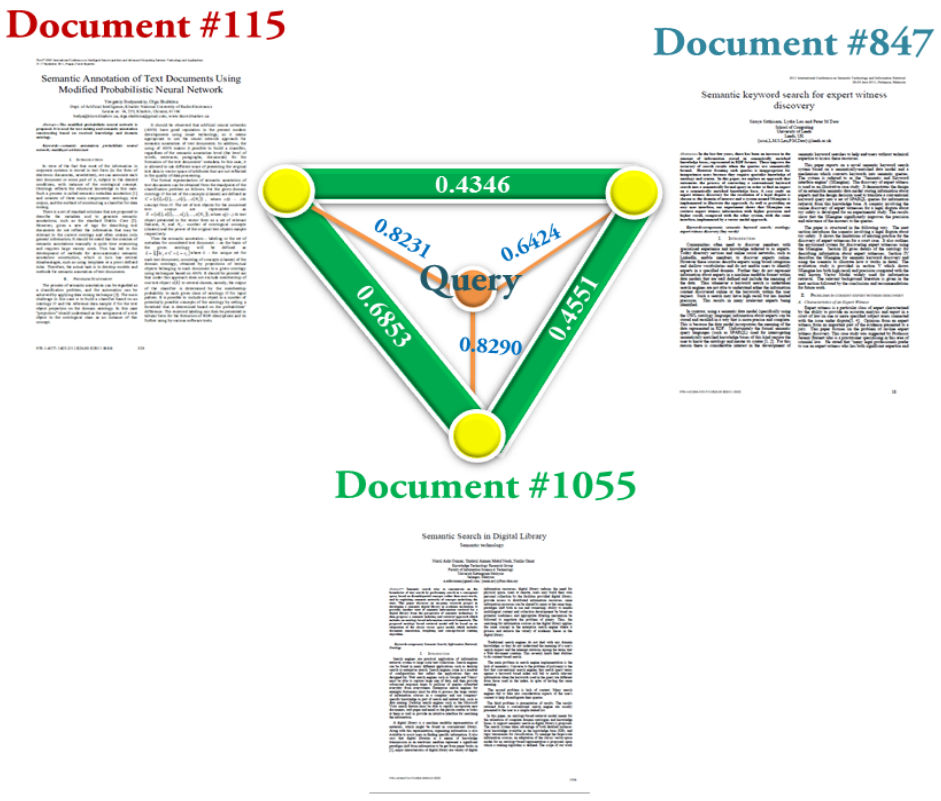


Figure 3. Distance Similarity between Document and Query

6. Conclusion

In this paper, we have presented a new document Semantic ranking process for the semantic ranking that proposes a new weight of query term in the document based on Computer Science Ontology weight to prevent a bias towards higher term frequency. We analyzed and compared two methods that used semantic weight and term frequency weight (non-semantic). Table 4 shows that the ontology weight can be evaluated and ranked results focus on the meaning of user's query. The experimental results show that the document similarity score between a user's query and the paper suggests that the new measure were effectively ranked.

As an example, "*Decision*" and "*Making*" are the query for the experiment. The query term ("*Making*") is not found in the paperID#1706 ,but similarity score by non-semantic weight is the sixth in the result because of the term "*Decision*" is a high term frequency. On the contrary, the similarity score based on Computer Science ontology weight is the ninth in the result.

We propose a new keyword weight calculating method for Keyword-Keyword matrix scores and Document-Document matrix scores. The paper adopted a Computer Science Ontology to hierarchical computation based on Wu and Palmer measure and tri-grams method to find the similarity of keywords string matching that kept a comprehensive keyword and discard ponderous keywords. The experimental results present document similarity scores using our Keyword-Keyword matrix scores. We therefore suggest the new document similarity measures are effectively document similarity scores.

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

References

[1] Manning, C. D., Raghavan, P., Schütze, H. (2009). An Introduction to Information Retrieval, Online Edition, MA: Cambridge University Press. [E-book] Available : http://www.informationretrieval.org/

[2] Sridevi, U. K., Nagaveni, N. (2010). Ontology based Similarity Measure in Document Ranking, *Computer Applications*., 1 (26) 135-139.

[3] Bouramoul, A., Kholladi, M -K., Doa, B-L. (2012). An ontology-based approach for semantic ranking of the web search engines results, *In*: proceeding of the 2012 3rd International Conference on Multimedia Computing and Systems, ICMCS 2012, 10-12 May, Tangier-Morocco. Available: IEEE Xplore, http://www.ieee.org.

[4] CASTELLS, P., FERNÁNDEZ, M., VALLET, D. (2007). An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. Knowledge and Data Engineering., 19 (2) 261 – 272, Feb.

[5] Jones, M., Alani, H. (2006). Content-based Ontology Ranking. *In*: proceeding of the 2006 9th International Protégé Conference, 23-26 July, Stanford California. Available: http://eprints.soton.ac.uk/id/eprint/ 262605

[6] Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J., Lorés, J. (2004). Neptuno: Semantic Web Technologies for a Digital Newspaper Archive, *In*: proceeding of the 2004 1st European Semantic Web Symposium, Lecture Notes in Computer Science, 3053., 10-12 May , Crete Greece. Available : Springer Berlin Heidelberg Xplore,http://link.springer.com/

[7] Maedche, A., Staab, S., Stojanovic, N., Studer, R., Sure, Y. (2003). SEmantic portAL: The SEAL Approach," In : Fensel, D., Hendler, J. A., Hendler, J. A., Lieberman, H., Wahlster, W. (eds.): *Spinning the Semantic Web*. MIT Press, p. 317-359, Cambridge London.

[8] Khan, L., McLeod, D., Hovy, E. (2004). Retrieval Effectiveness of an Ontology-Based Model for Information Selection, The International Journal on Very Large Data Bases, 13 (1) 71-85, January. [Online]. Available: http://dl.acm.org/

[9] Weng, S -S., Tsai, H -J., Hsu, C- H. (2006). Ontology construction for information classification, *Journal of Expert Systems with Applications*, 31 (1) 1–12. Available : www.elsevier.com/locate/eswa

[10] Aleman-Meza, B., Hakimpour, F., Budak Arpinar, I. (2007). SwetoDblp ontology of Computer Science publications, The Journal on Web Semantics: Science, Services and Agents on the World Wide Web, 5 (3), p.151-155, September. Available : Available: http://dl.acm.org/

[11] Association for Computing Machinery (ACM),IEEE Computer Society. Computer Science Curricula2013 draft report, February 2012. Retrieved October 13, 2012, from http://ai.stanford.edu/users/sahami/CS2013/

[12] Boonyound, T., Mingkhwan, A. (2014). Semantic Search using Computer Science Ontology based on Edge Counting and N-Grams, *In*: proceeding of the 10th International Conference on Computing and Information Technology, IC2it 2014, 8-9 May 2014, Thailand.

[13] Boonyound, T., Mingkhwan, A. (2014). Semantic Ranking based on Computer Science Ontology, *In*: proceeding of the Ninth International Conference on Digital Information Management (ICDIM 2014), 29 Sep – 01 Oct, Thailand.

[14] Wu, Z., Palmer. M. (1994). Verb semantics and lexical selection. *In*: 32nd. Annual Meeting of the Association for Computational Linguistics. (p. 133-138), New Mexico State University, Las Cruces, New Mexico.

[15] Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G. M., Milios, E. (2006). Information Retrieval by Semantic Similarity, *International Journal on Semantic Web and Information Systems* (IJSWIS), 3 (3). p. 55-73.

[16] Stoke, N. (2004). Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain, A thesis submitted for the degree of Doctor of Philosophy in Computer Science Department of Computer Science Faculty of Science National University of Ireland, Dublin.

[17] Kondrak, G. (2005). N-Gram Similarity and Distance. Lecture Notes in Computer Science. *In*: Proceeding of the 12th String Processing and Information Retrieval. (p. 115-126), Aires, Argentina.

[18] Sembok,T. M., Bakar, Z. A. (2011). Effectiveness of Stemming and N-grams String Similarity Matching on Malay Documents. *International Journal of Applied Mathematics and Informatics*,5 (3) 208-215.

[19] Boonyoung, T., Mingkhwan, A. (2014). Document Similarity using Computer Science Ontology based on Edge Counting and N-Grams, In proceeding of the 15th Annual PostGraduate Symposium on the Convergence of Telecommunication, Networking and Broadcasting, PG NET 2015, 23-24 June 2014, Liverpool England.

[20] Dice, L. R. (1945). Measures of the amount of ecologic association Between species, *Ecology*, 26 (3), 297-302.

[21]Watthananon, Julaluk., Mingkhwan, Anirach. (2012). A Comparative Efficiency of Correlation Plot Data Classification, The Journal of KMUTNB., 22 (1).

[22] Lertmahakrit, Wilaiporn., Mingkhoan, Anirach. (2010). The Innovation of Multiple Relations Information Retrieval, *The Journal of KMUTNB*., 20 (3).

Thanyaporn Boonyoung was born on July 2st, 1977. She earned Master's degree in Information Technology from King Mongkut's University of Technology North Bangkok, Thailand in 2003.

She is currently a doctoral student in faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Thailand. Her research areas cover Information Retrieval and semantic web.

Dr. Anirach Mingkhwan was born on August 7th, 1969, and he is the Associate Professor. He earned doctorate in computer network, School of Computing and Mathematical Sciences from Liverpool John Moores University, United Kingdom in 2004.

He is teaching Information Technology and currently serving as a Dean of Faculty for Industrial Technology and Management, King Mongkut's Institute of Technology North Bangkok, Thailand. His main research interest include networks, information graphics, information retrieval, ubiquitous computing, library science, computer networks, information technology, knowledge discovery in databases, information visualization, network forensics, service oriented computing, wireless sensor network, digital libraries, information security, wireless security, network security, mobile computing, computer science, distributed computing, systems, wireless, Ad Hoc Networks, Computer Forensics, Digital Investigation, Vehicular Ad Hoc Networks, and Mobile Ad Hoc Networks.