KMUTNB

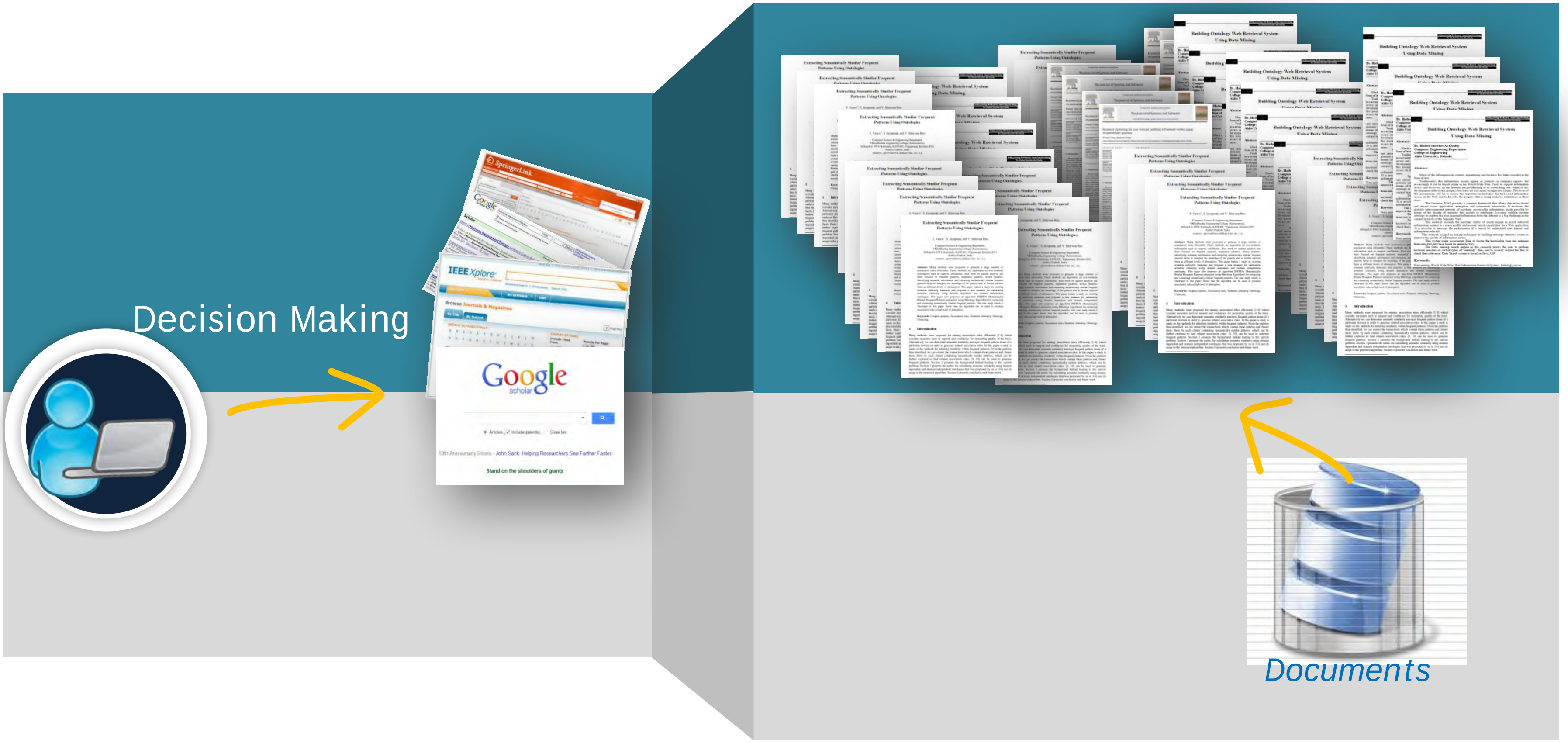
Algorithm for Information Visualization of Semantic Search Using Computer Science Research Ontology

Faculty of Industrial Technology and Management

Assoc. Prof. Dr.Anirach Mingkhwan

King Mongkut's University of Technology North Bangkok (KMUTNB),Thailand.

Keyword Based Search



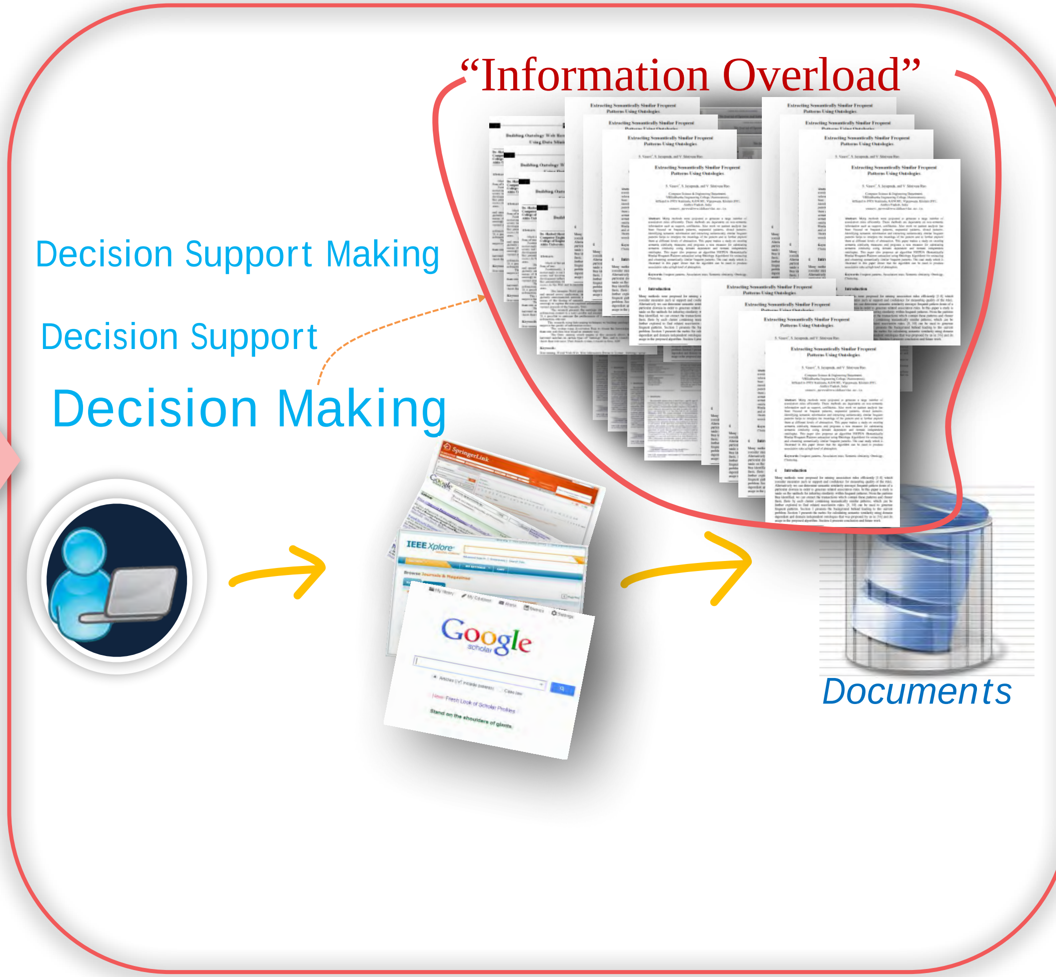
Traditional Keyword-based Matching

Keyword Based Search Limit

01
Keyword based search

Just Keyword

- ✗Synonym
- ✗Hyponym

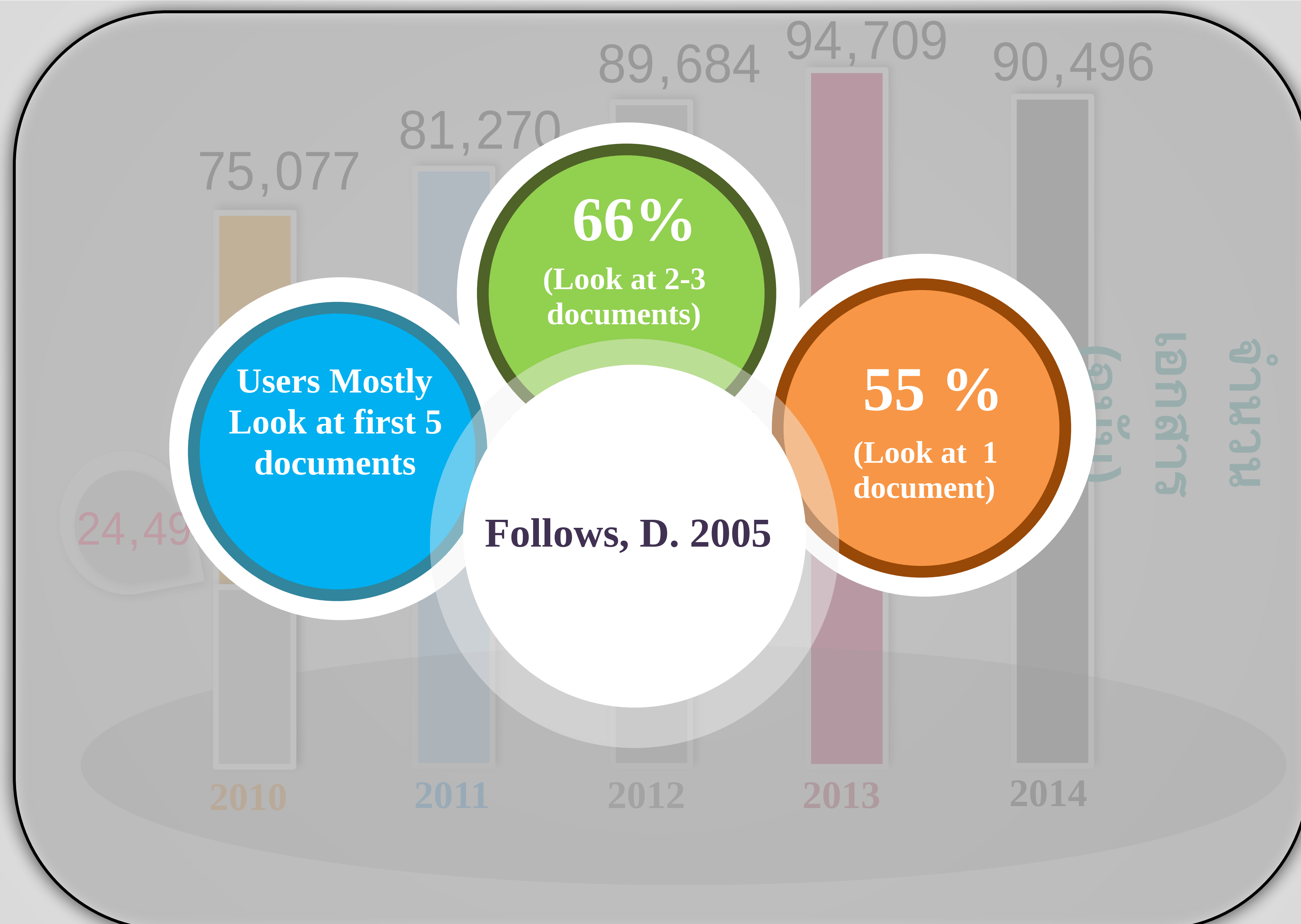


Keyword Based Search Limit

01

Keyword based search

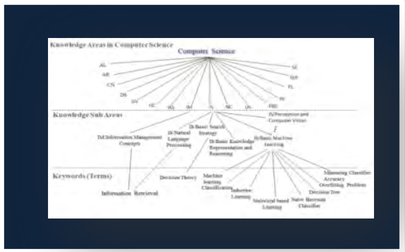
Results Example


75,077
81,270
89,684
94,709
90,496
24,49
66%
(Look at 2-3 documents)
Users Mostly Look at first 5 documents
55 %
(Look at 1 document)
Follows, D. 2005
2010
2011
2012
2013
2014

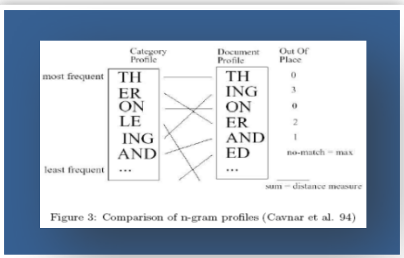
Keyword Based Search Limit

Propose Solutions

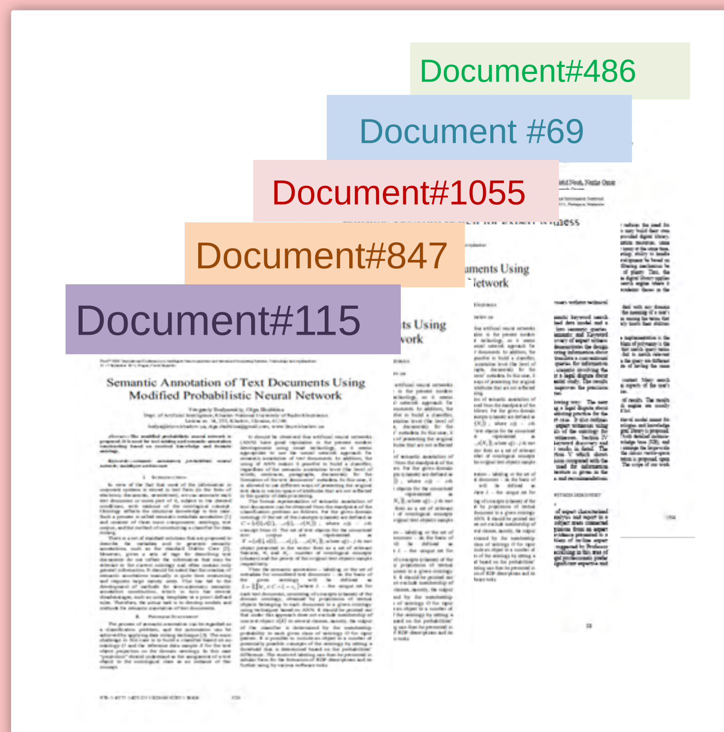
01
Keyword based search



Computer Science
Ontology



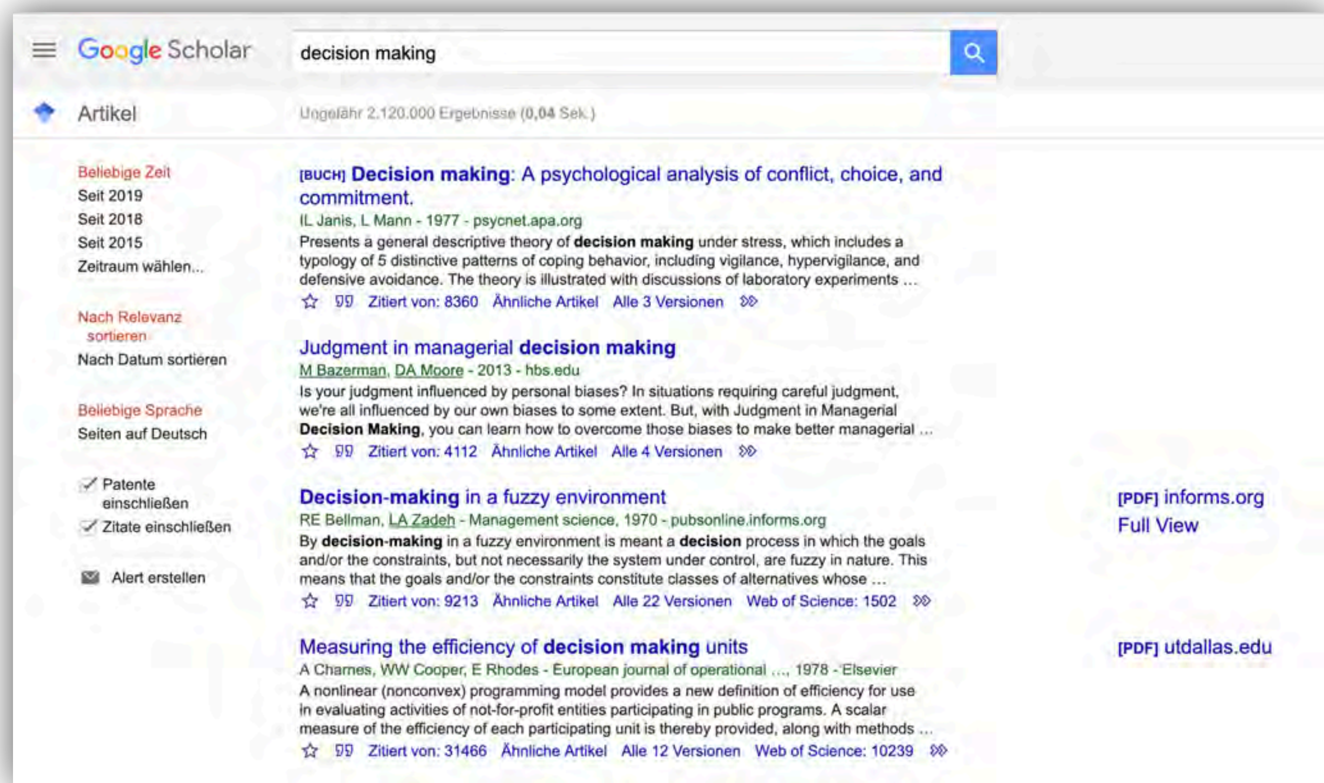
N-grams



Semantic Ranking

Keyword Based Search Limit

02
Query and Answering



Keyword Based Search Limit

02
Query and Answering

Questions:

1) What area of Computer Sciences most active in USA year 2008
2) This document related to what area in Computer Sciences.
3) Which Country are the most active in "Data Mining".

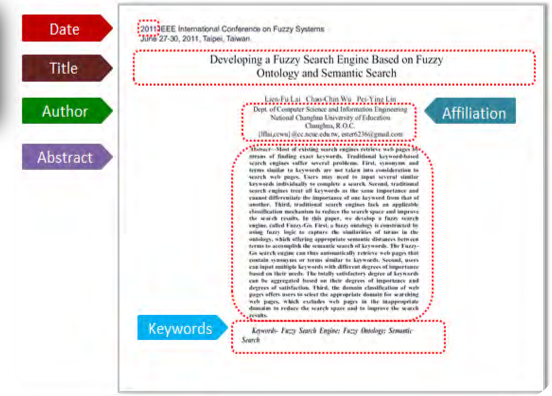
Keyword Based Search Limit

02
Query and Answering

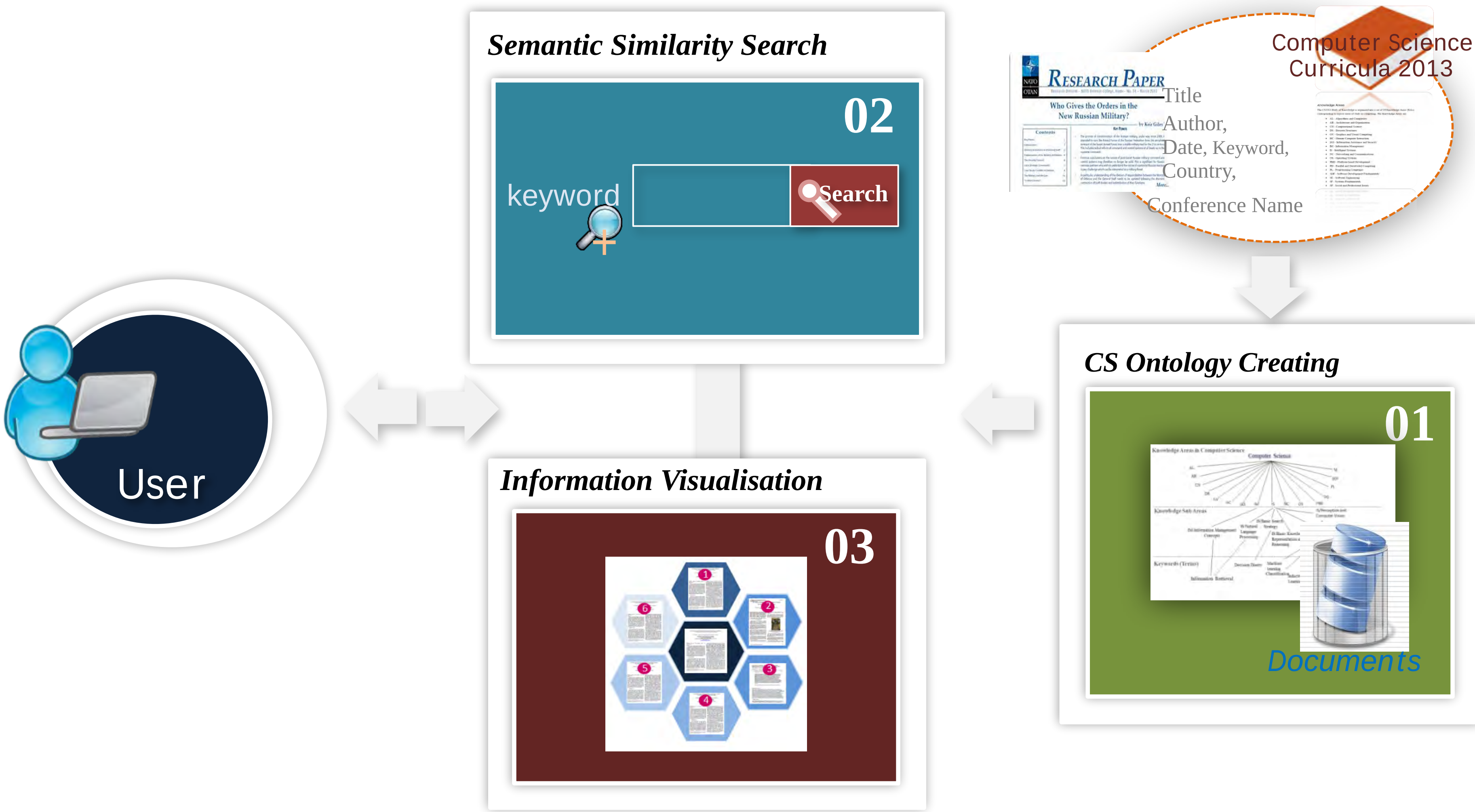
Propose **Solutions**

Research Paper

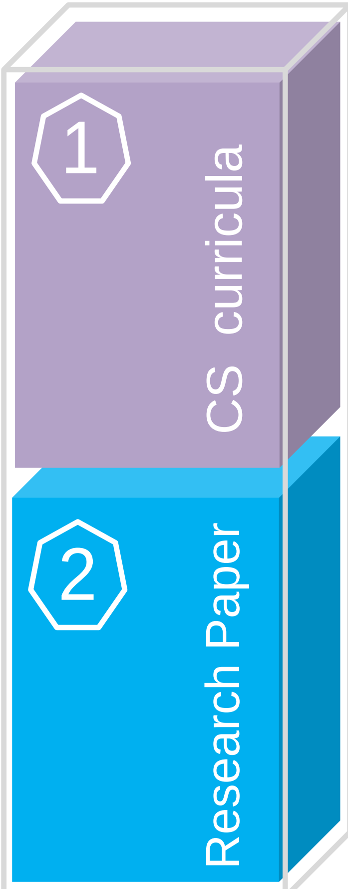
Title | Author | Country | Abstract | Keyword | Date



The Framework of Semantic Search

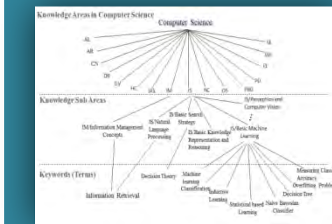


Research Database



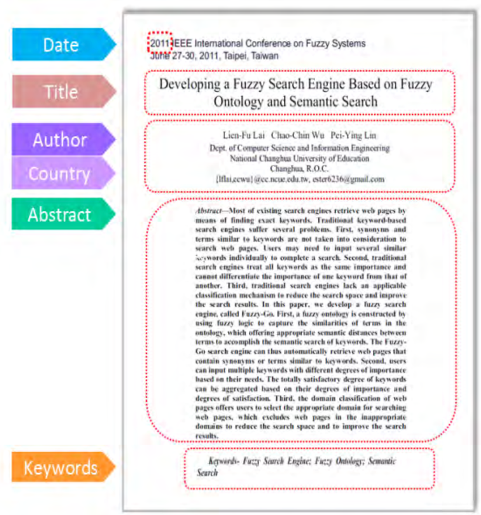
Computer science curricula 2013 report

(IEEE&ACM)



Computer Science Ontology

1800 documents
(2008-2012)



2011 IEEE International Conference on Fuzzy Systems
June 27-30, 2011, Taipei, Taiwan

Developing a Fuzzy Search Engine Based on Fuzzy Ontology and Semantic Search

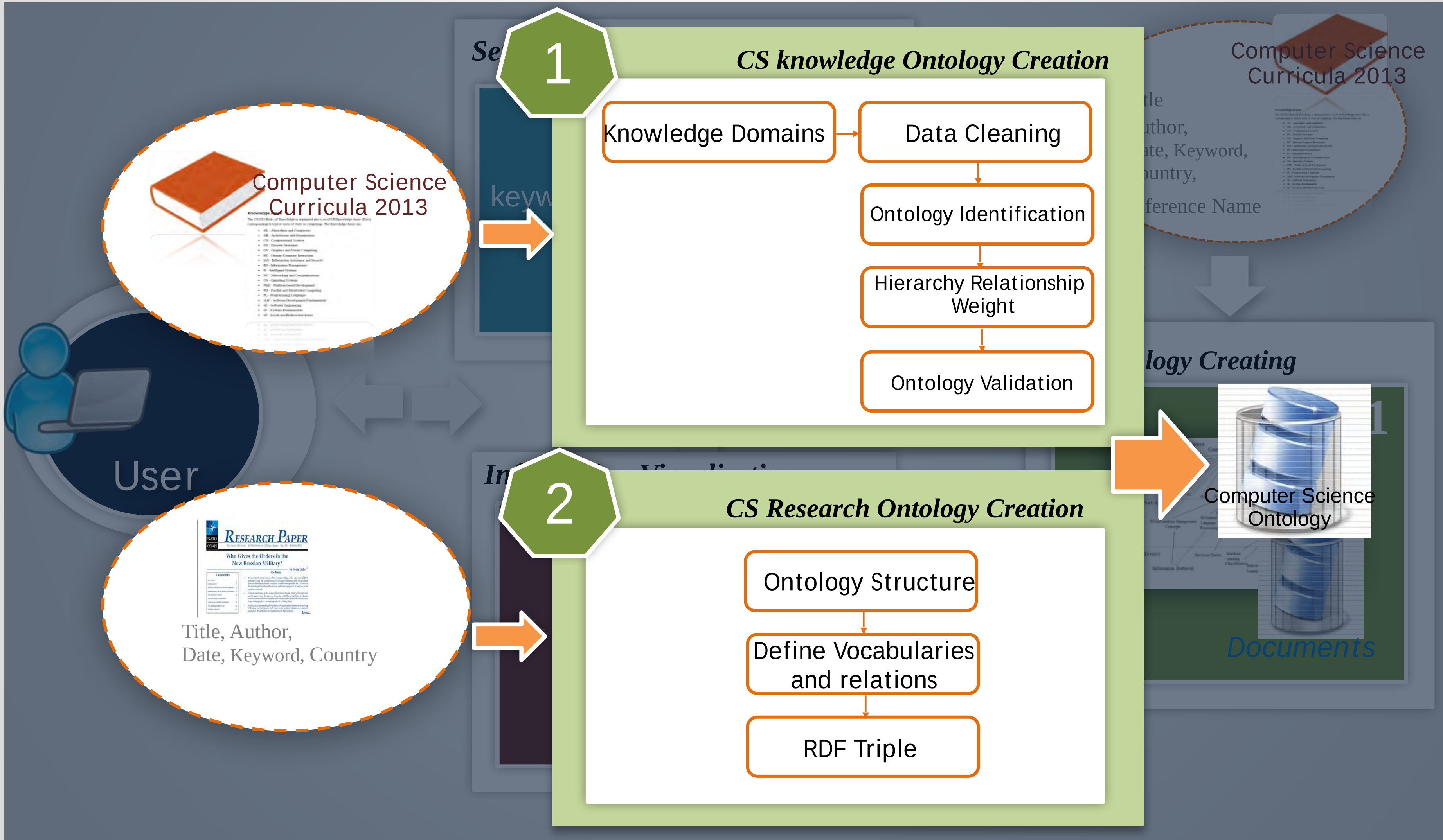
Lien-Fu Lai Chao-Chin Wu Pei-Ying Lin
Dept. of Computer Science and Information Engineering
National Changhua University of Education
Changhua, R.O.C.
{lflai,ccwu}@cc.ncue.edu.tw, ester6236@gmail.com

Abstract—Most of existing search engines retrieve web pages by means of finding exact keywords. Traditional keyword-based search engines suffer several problems. First, synonyms and terms similar to keywords are not taken into consideration to search web pages. Users may need to input several similar keywords individually to complete a search. Second, traditional search engines treat all keywords as the same importance and cannot differentiate the importance of one keyword from that of another. Third, traditional search engines lack an applicable classification mechanism to reduce the search space and improve the search results. In this paper, we develop a fuzzy search engine, called Fuzzy-Go. First, a fuzzy ontology is constructed by using fuzzy logic to capture the similarities of terms in the ontology, which offering appropriate semantic distances between terms to accomplish the semantic search of keywords. The Fuzzy-Go search engine can thus automatically retrieve web pages that contain synonyms or terms similar to keywords. Second, users can input multiple keywords with different degrees of importance based on their needs. The totally satisfactory degree of keywords can be aggregated based on their degrees of importance and degrees of satisfaction. Third, the domain classification of web pages offers users to select the appropriate domain for searching web pages, which excludes web pages in the inappropriate domains to reduce the search space and to improve the search results.

Keywords- Fuzzy Search Engine; Fuzzy Ontology; Semantic Search

Documents

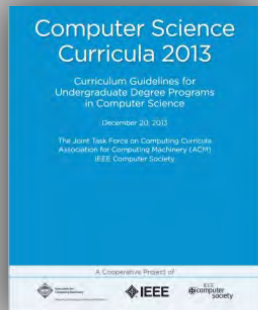
01 Cs Ontology Creating



01 Cs Ontology Creating *(CS knowledge Ontology Creation)*

1

Computer Science Knowledge Domain



(Computer science curricula 2013 report)

(Endorsed by Association for Computing Machinery (ACM) and IEEE Computer Society)

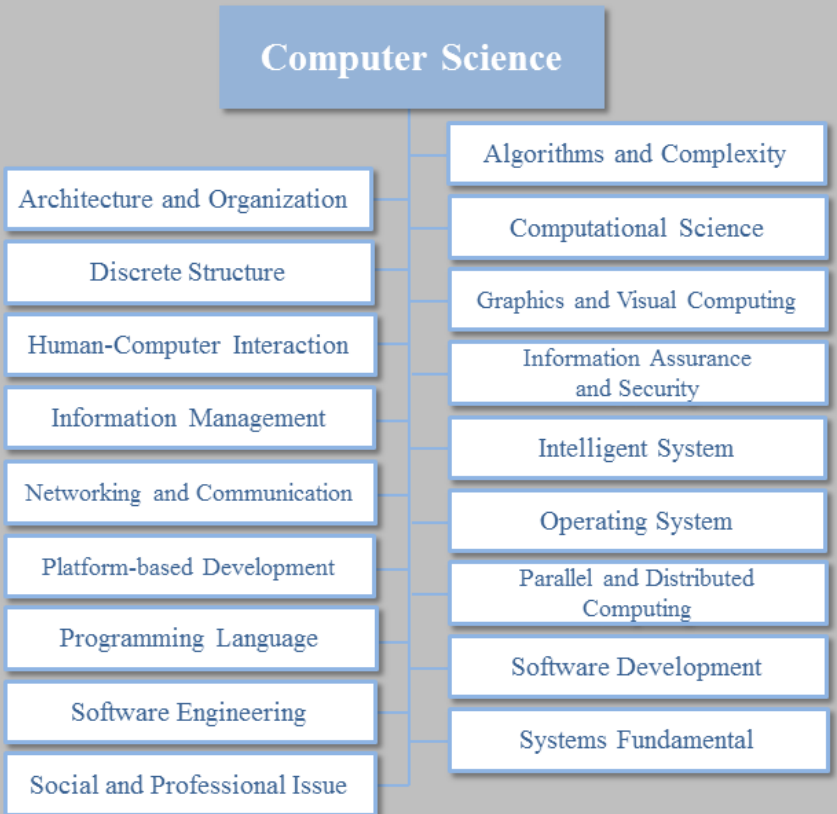
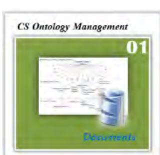


Figure : Knowledge Area of Computer Science

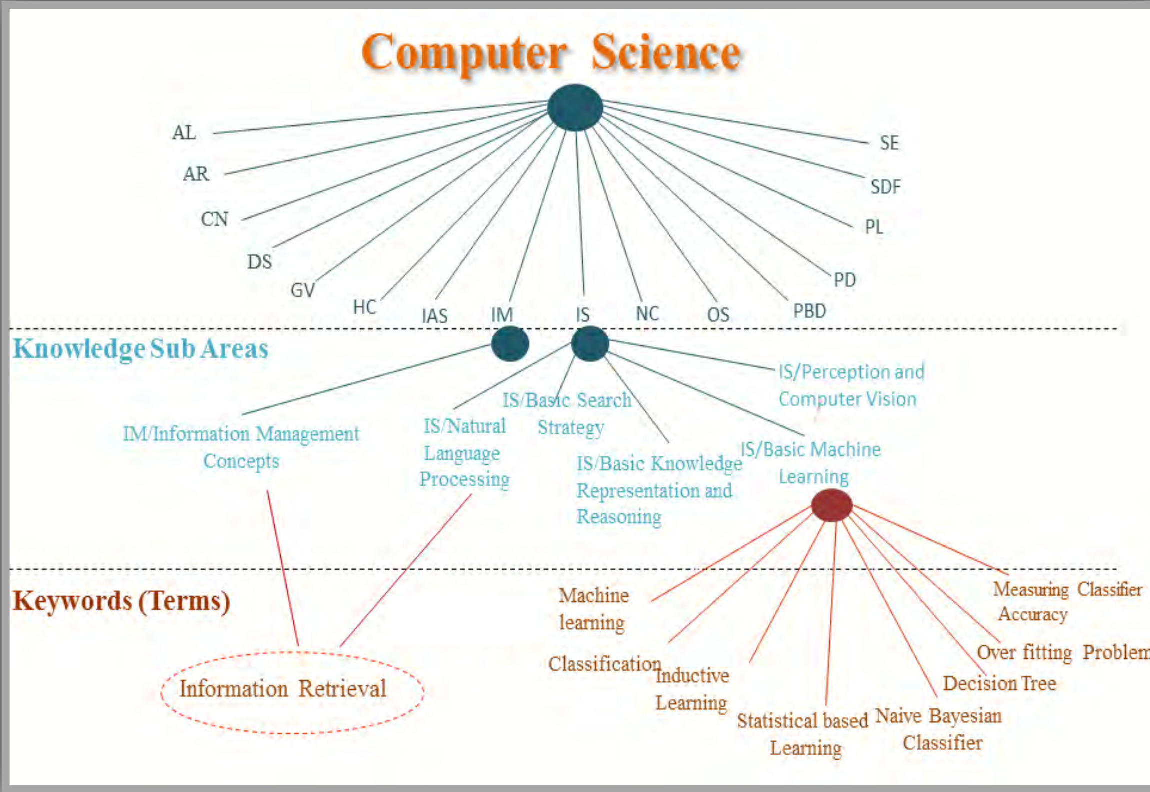


Ref : http://ai.stanford.edu/users/sahami/CS2013/

01 Computer Science Ontology Creating

2

Taxonomic hierarchy & Weight



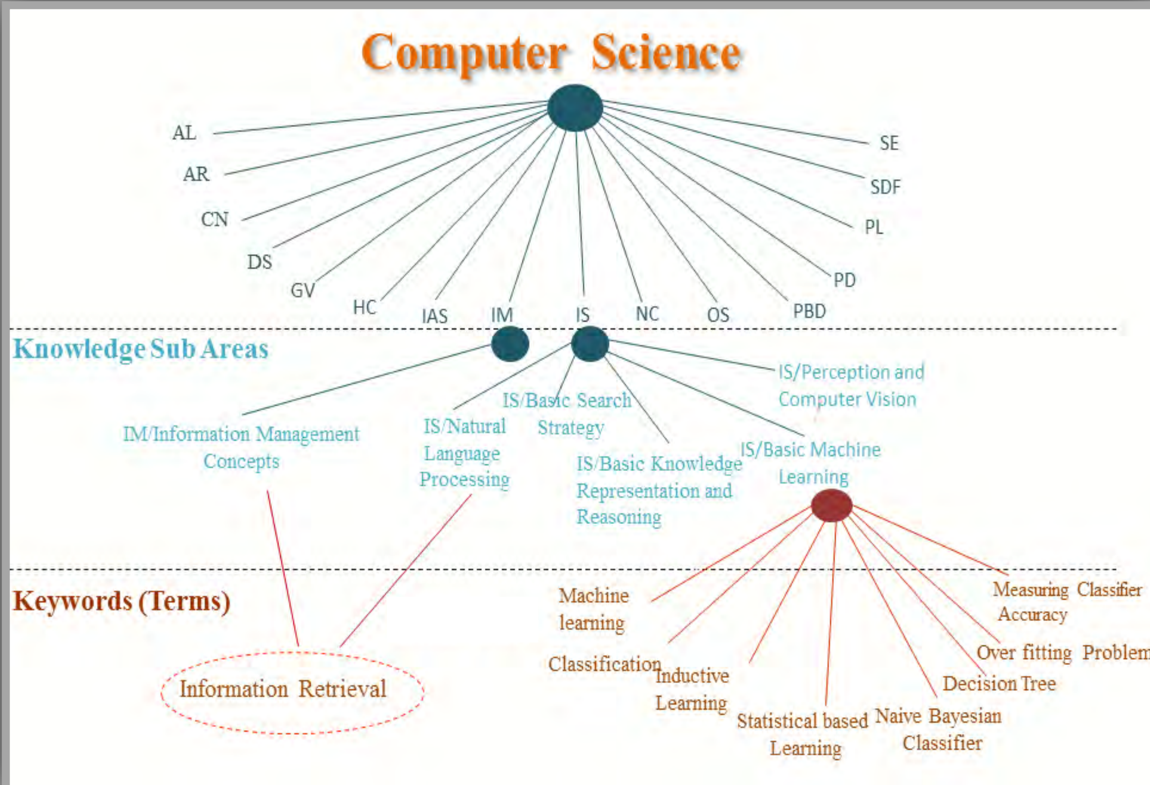
18 Taxonomies

The taxonomic hierarchy of Computer Science

01 Computer Science Ontology Creating

2

Taxonomic hierarchy & Weight



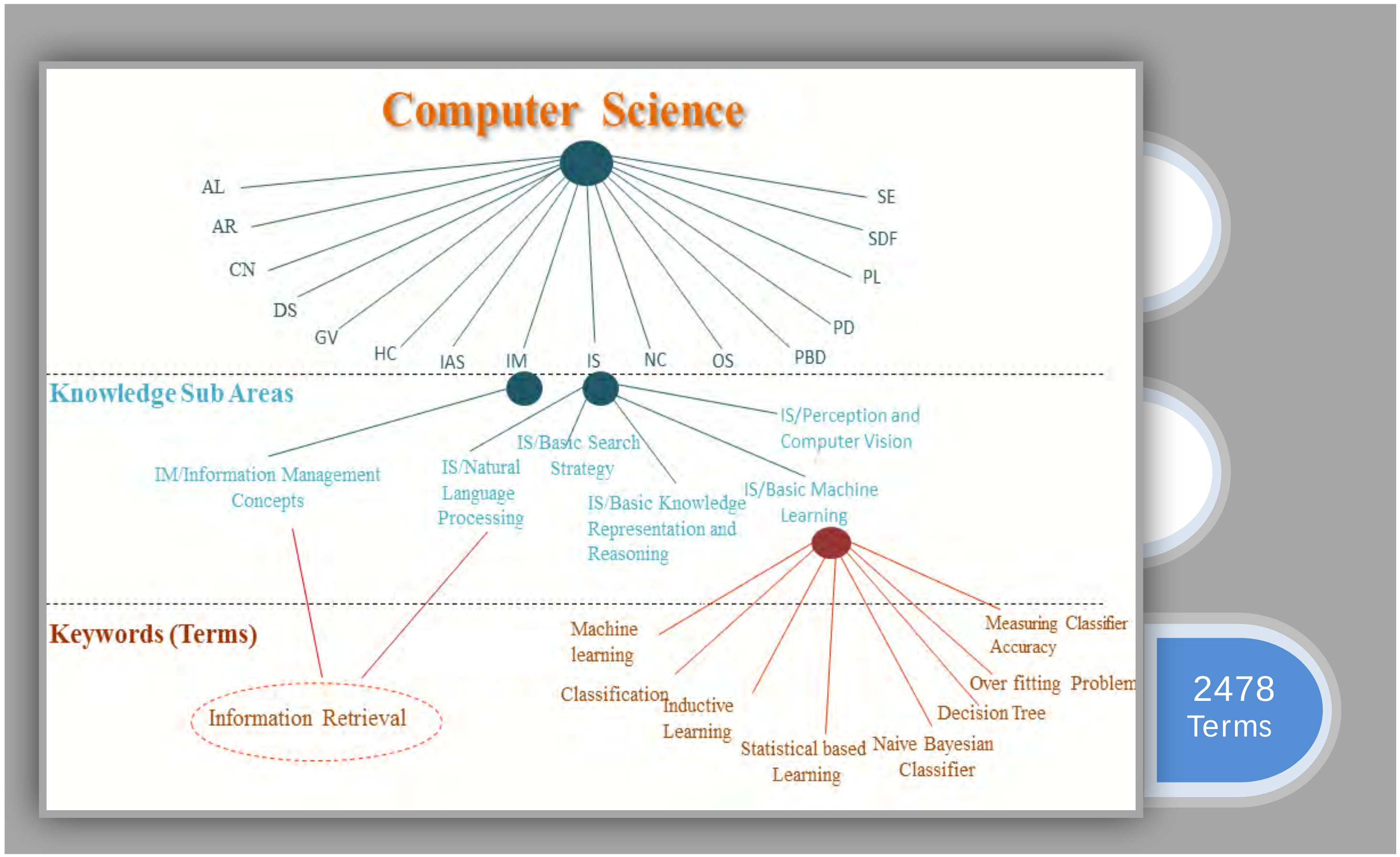
155 Terms

The taxonomic hierarchy of Computer Science

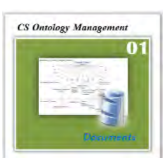
01 Computer Science Ontology Creating

2

Taxonomic hierarchy & Weight



The taxonomic hierarchy of Computer Science



01 Computer Science Ontology Creating

2

Taxonomic hierarchy & Weight

Similarity Measure (Edge Counting)

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

(Wu, Z. and Palmer, Edge Counting Measure, 1994.)

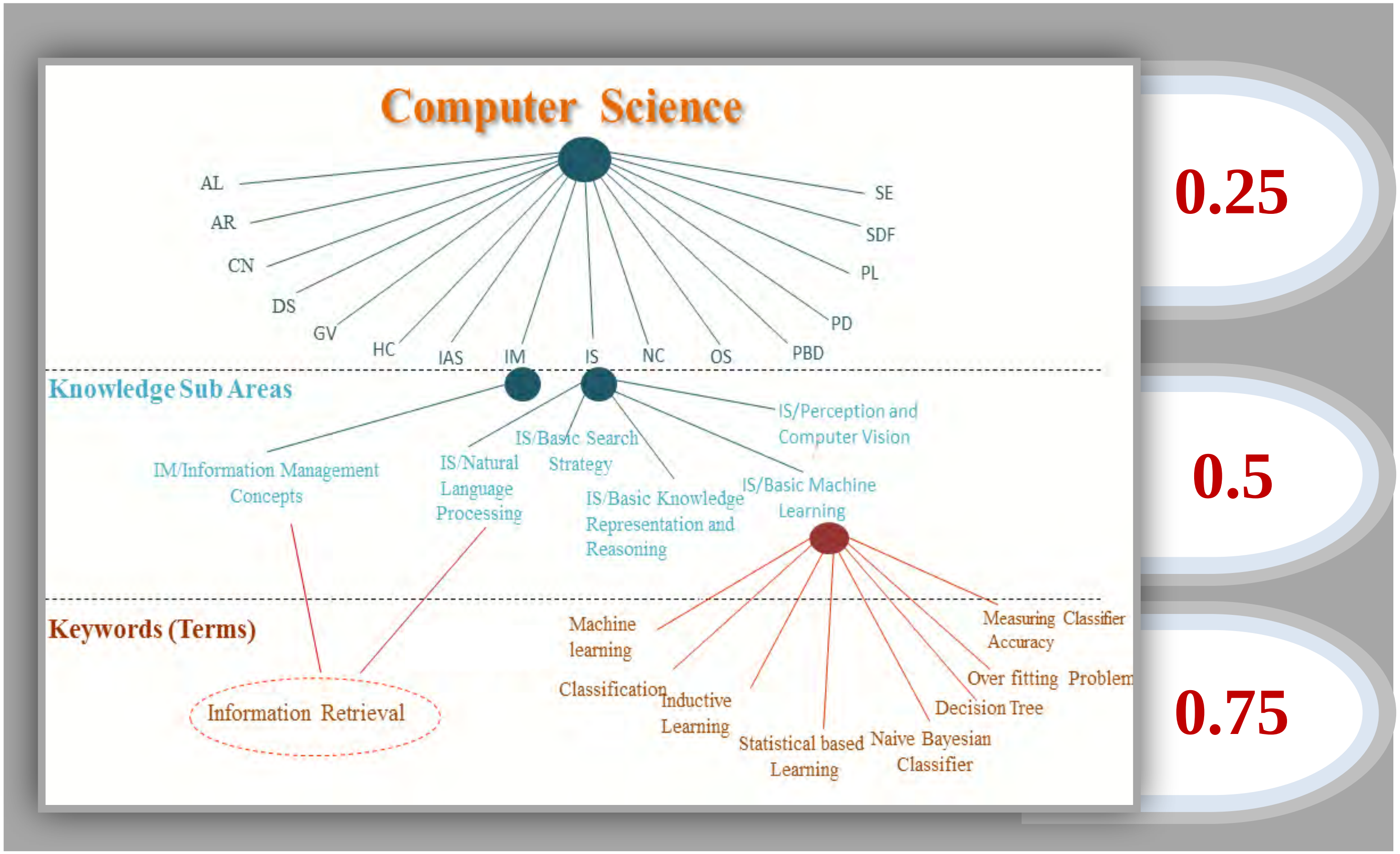
Weight of the ontology relation table

Relationship Type	Weight
Repetition / Synonymy	1
Same Sub Area	0.75
Same Area	0.5
different area	0.25
not found In CS ontology	0

01 Computer Science Ontology Creating

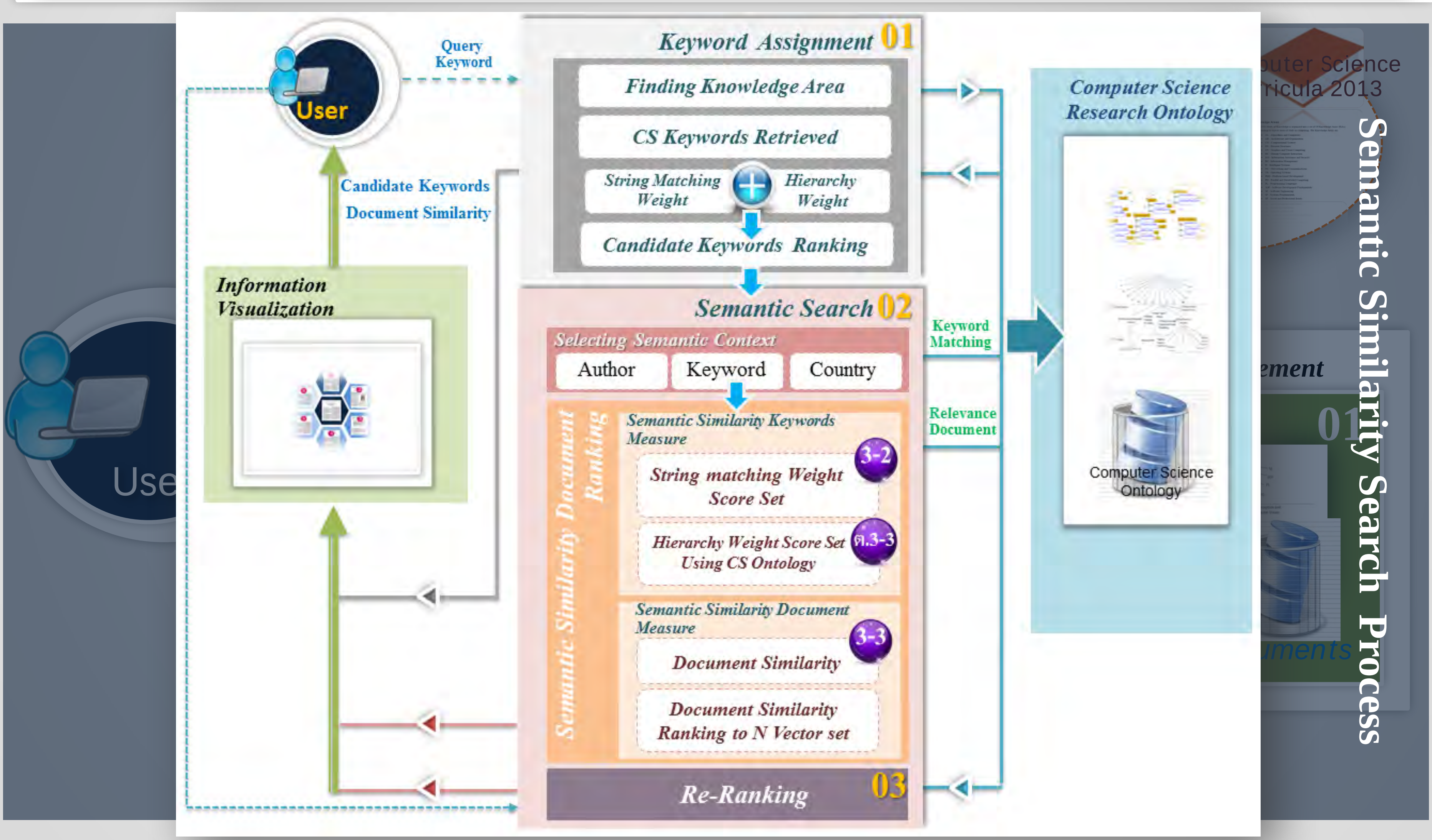
2

Taxonomic hierarchy & Weight



The taxonomic hierarchy of Computer Science

The Framework of Semantic Search (02 Semantic Similarity Search)



02 Semantic Similarity Search (Keyword Semantic Similarity)

1) String Matching Weight Measure

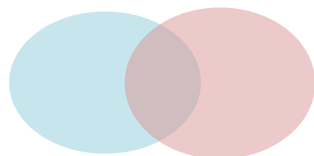
Medical Decision Making

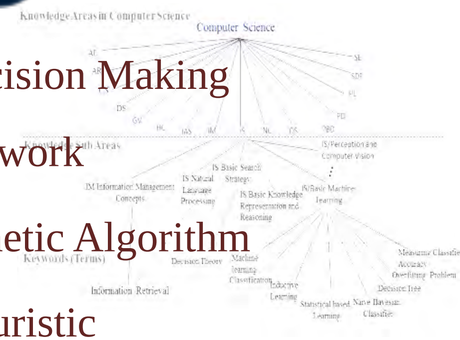
Electrical Network

Improved Genetic Algorithm

Heuristic Evaluation

Document Keywords





Decision Making

Network

Genetic Algorithm

Heuristic

Ontology Keywords
(Computer Science Ontology)

N-grams
(Trigrams)

M *Me Med edi dic ica cal al* l

D *De Dec cis isi sio ion on* n

M *Ma Mak aki kin ing ng* g

Grams Grzegorz Kondrak formulates a family of word similarity measures based on N-grams.

$$\frac{2x|n\text{-}Keyword(QW)\cap n\text{-}Keyword(OW)|}{|n\text{-}Keyword(QW)|+|n\text{-}Keyword(OW)|} \quad \text{.....} \quad A$$

Ref : (Grzegorz Kondrak,2005)

Semant
Pro

2) Hierarchical Relationship Weights

Relationship Type	Weight
Repetition /Synonymy	1
same sub area	0.75
same area	0.5
different area	0.25
not found In CS ontology	0

For Example

Repetition /Synonymy
Database, DB → 1
Information Retrieval, IR
Entity-Relationship, E-R

Steps

Calculating Document Keywords Score

Calculate Document Similarity Score

Calculating Document Similarity Score

(a) Calculate N-grams Weight Set of Similarity between Query Keyword with Ontology Keyword.

(b) Calculate Weight Set of Semantic Similarity between Query Keyword with Ontology Keyword

$$SWK_{d,i}(QK,DK) = SWK_{n\text{-}gram} + SWK_{cs\text{-}onto}$$

--- $SWK_{d,i}(QK,DK)$: the total Semantic Weight Keyword (SWK)

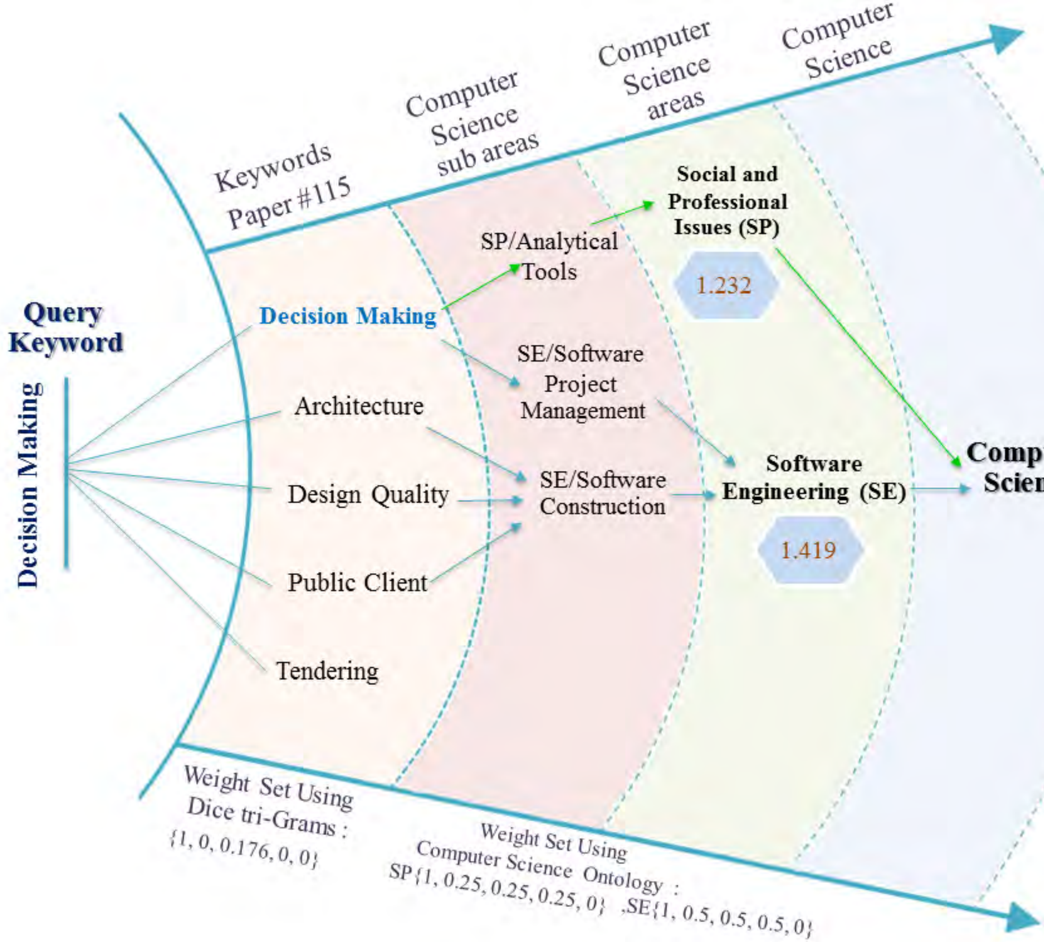
--- SWKn-gram : semantic weight keyword each document based on N-grams

--- SWKcs-onto : semantic weight keyword each document based on Computer Science ontology

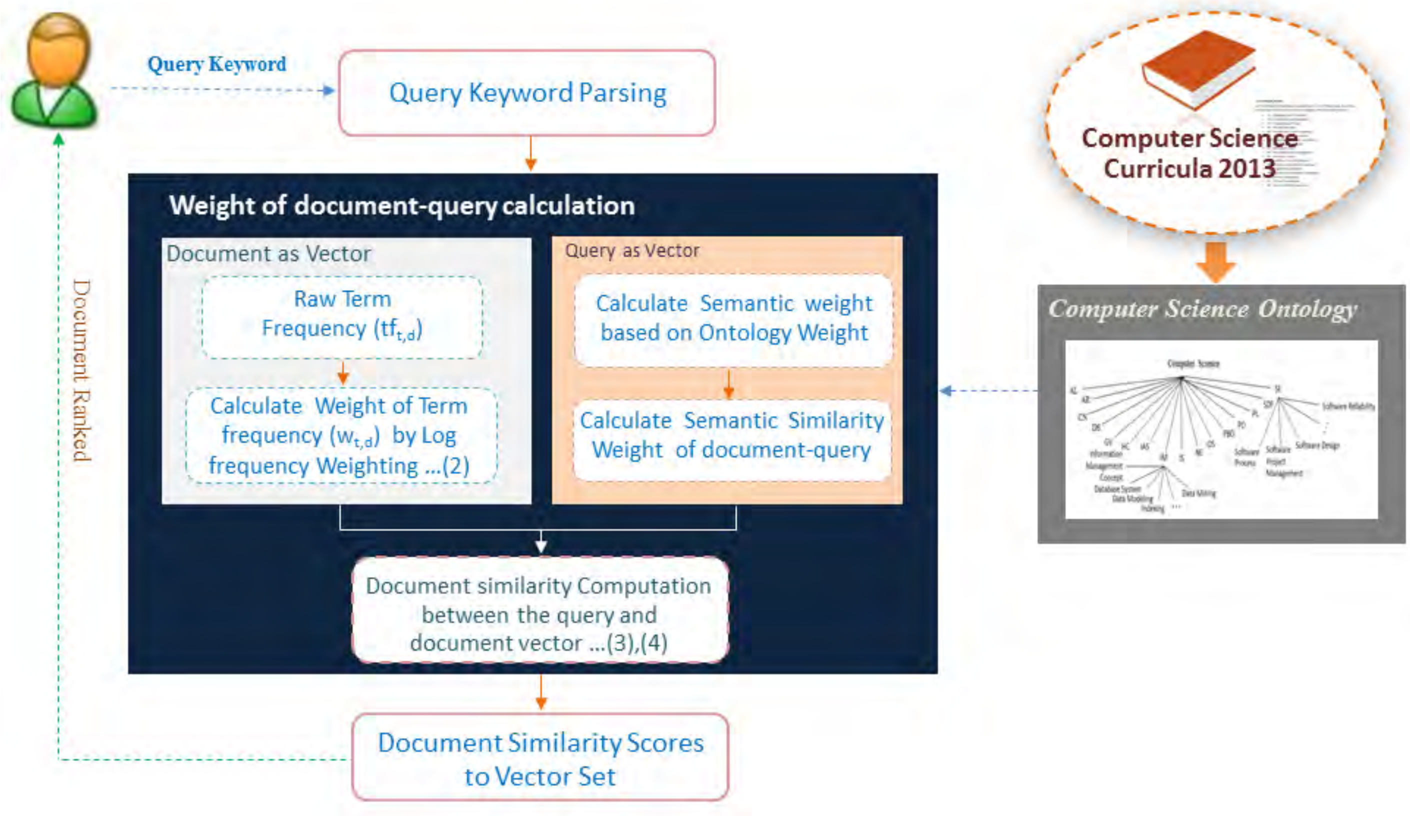
$$\mathbf{DSS}\left(D_d, \mathbf{q}\right) = \begin{cases} 1 + \dfrac{\sum_{i=1}^{N} SWK_{d,i}}{N-1} & ;\ \text{One of the query and Keywords document } d \text{ are exactly alike.} \\ \dfrac{\sum_{i=1}^{N} SWK_{d,i}}{N} & ;\ \text{otherwise} \end{cases}$$

02 Semantic Similarity Search (Calculating Document Similarity Score)

Calculation Example



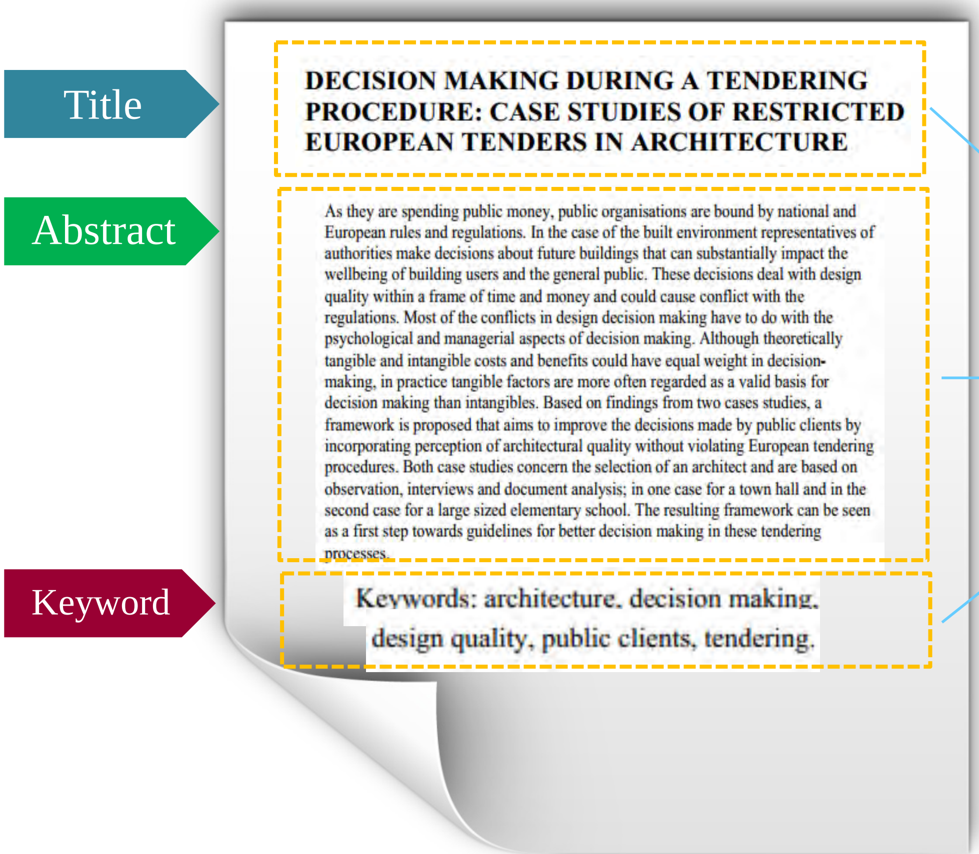
02 Semantic Similarity Search (Vector Space Model)



Vector Space Model Process

Semantic Ranking Process

Document as Vector | *Term-Frequency*



DECISION MAKING DURING A TENDERING PROCEDURE: CASE STUDIES OF RESTRICTED EUROPEAN TENDERS IN ARCHITECTURE

As they are spending public money, public organisations are bound by national and European rules and regulations. In the case of the built environment representatives of authorities make decisions about future buildings that can substantially impact the wellbeing of building users and the general public. These decisions deal with design quality within a frame of time and money and could cause conflict with the regulations. Most of the conflicts in design decision making have to do with the psychological and managerial aspects of decision making. Although theoretically tangible and intangible costs and benefits could have equal weight in decision-making, in practice tangible factors are more often regarded as a valid basis for decision making than intangibles. Based on findings from two cases studies, a framework is proposed that aims to improve the decisions made by public clients by incorporating perception of architectural quality without violating European tendering procedures. Both case studies concern the selection of an architect and are based on observation, interviews and document analysis; in one case for a town hall and in the second case for a large sized elementary school. The resulting framework can be seen as a first step towards guidelines for better decision making in these tendering processes.

Keywords: architecture, decision making, design quality, public clients, tendering.

Term-Frequency

Terms	*tf-raw*	*tf-wght*
Architecture	2	1.3010
client	2	1.3010
Decision	6	1.7782
Design	3	1.4771
making	6	1.7782
Public	5	1.6990
quality	3	1.4771
Tender	5	1.6990
Case	3	1.4771

Three parts (Title,Abstract,Keyword) for Representation of Term within a document

02 Semantic Similarity Search (Vector Space Model)

Keyword : Decision Making

Document PID#115				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**

Retrieval Results

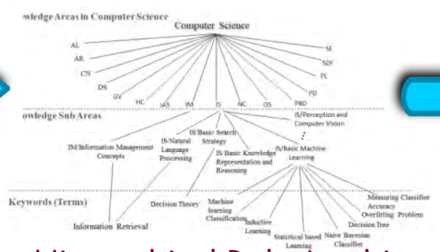
DECISION MAKING DURING A TENDERING PROCEDURE: CASE STUDIES OF RESTRICTED EUROPEAN TENDERS IN ARCHITECTURE

As they are spending public money, public organisations are bound by national and European rules and regulations. In the case of the built environment representatives of authorities make decisions about future buildings that can substantially impact the wellbeing of building users and the general public. These decisions deal with design quality within a frame of time and money and could cause conflict with the regulations. Most of the conflicts in design decision making have to do with the psychological and managerial aspects of decision making. Although theoretically tangible and intangible costs and benefits could have equal weight in decision-making, in practice tangible factors are more often regarded as a valid basis for decision making than intangibles. Based on findings from two cases studies, a framework is proposed that aims to improve the decisions made by public clients by incorporating perception of architectural quality without violating European tendering procedures. Both case studies concern the selection of an architect and are based on observation, interviews and document analysis; in one case for a town hall and in the second case for a large sized elementary school. The resulting framework can be seen as a first step towards guidelines for better decision making in these tendering processes.

Keywords: architecture, decision making, design quality, public clients, tendering.

TF-IDF Weights

+



Hierarchical Relationship Weight
(Using CS Ontology)

Semantic Ranking

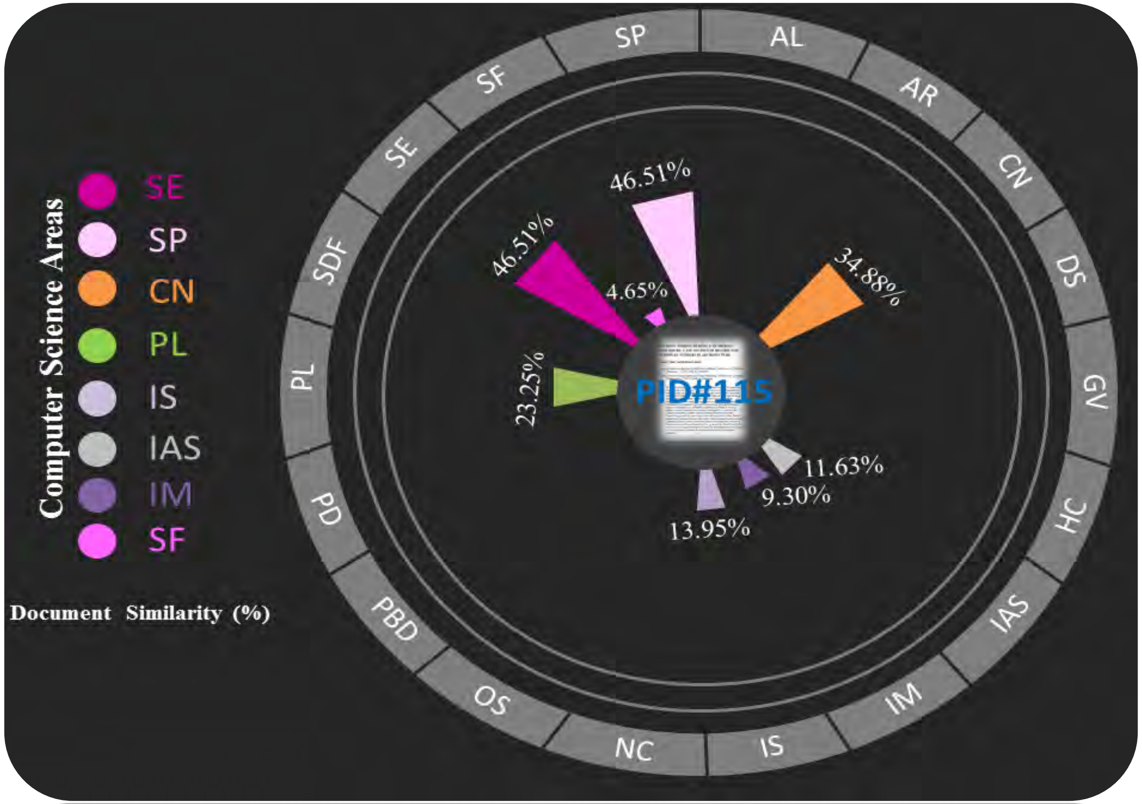
Cosine Similarity Measure

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

2

Query
And
Answering

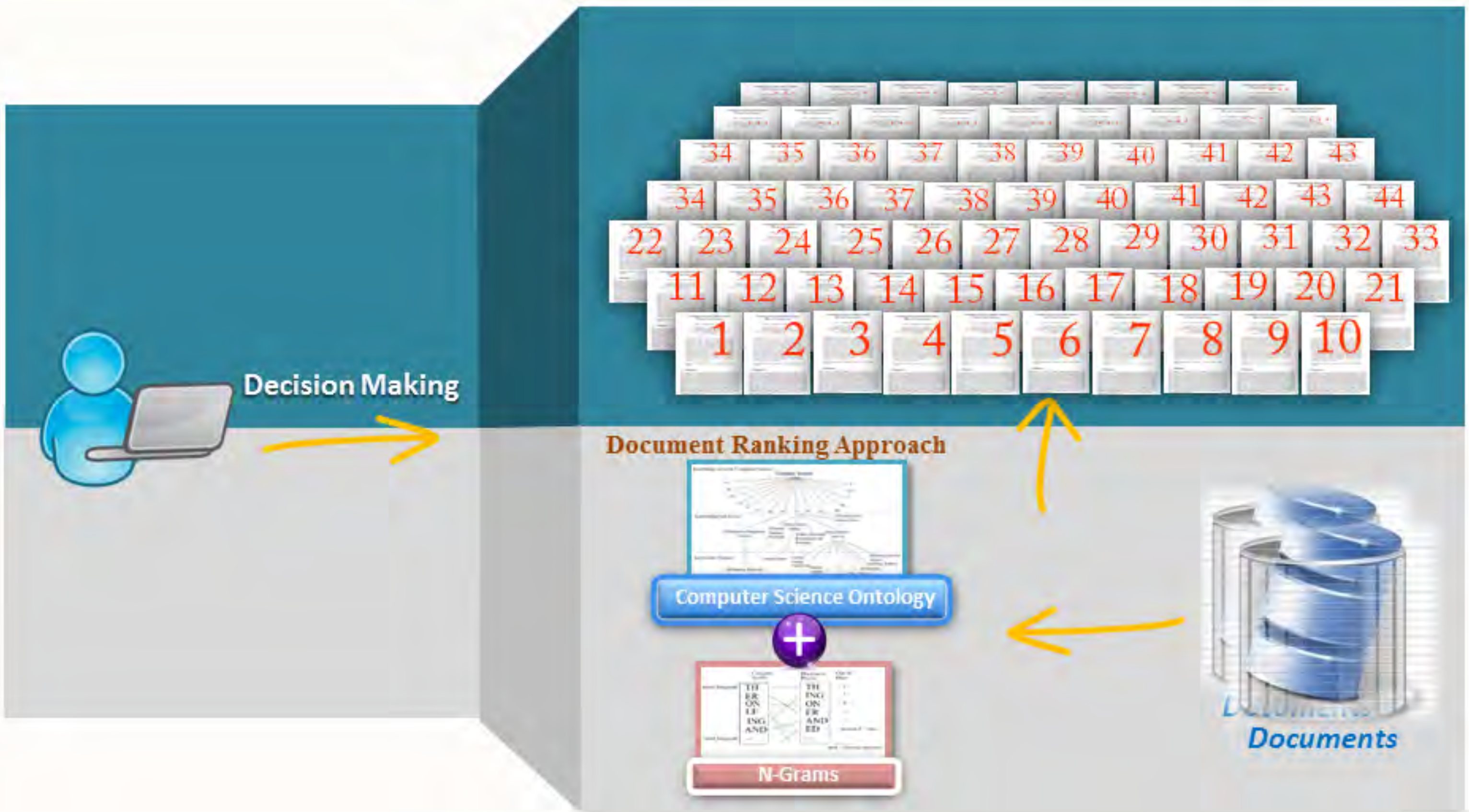
Paper Similarity % in Computer Science Area



02 Semantic Similarity Search (Re-Ranking)

Semantic Similarity Process

Re-Ranking



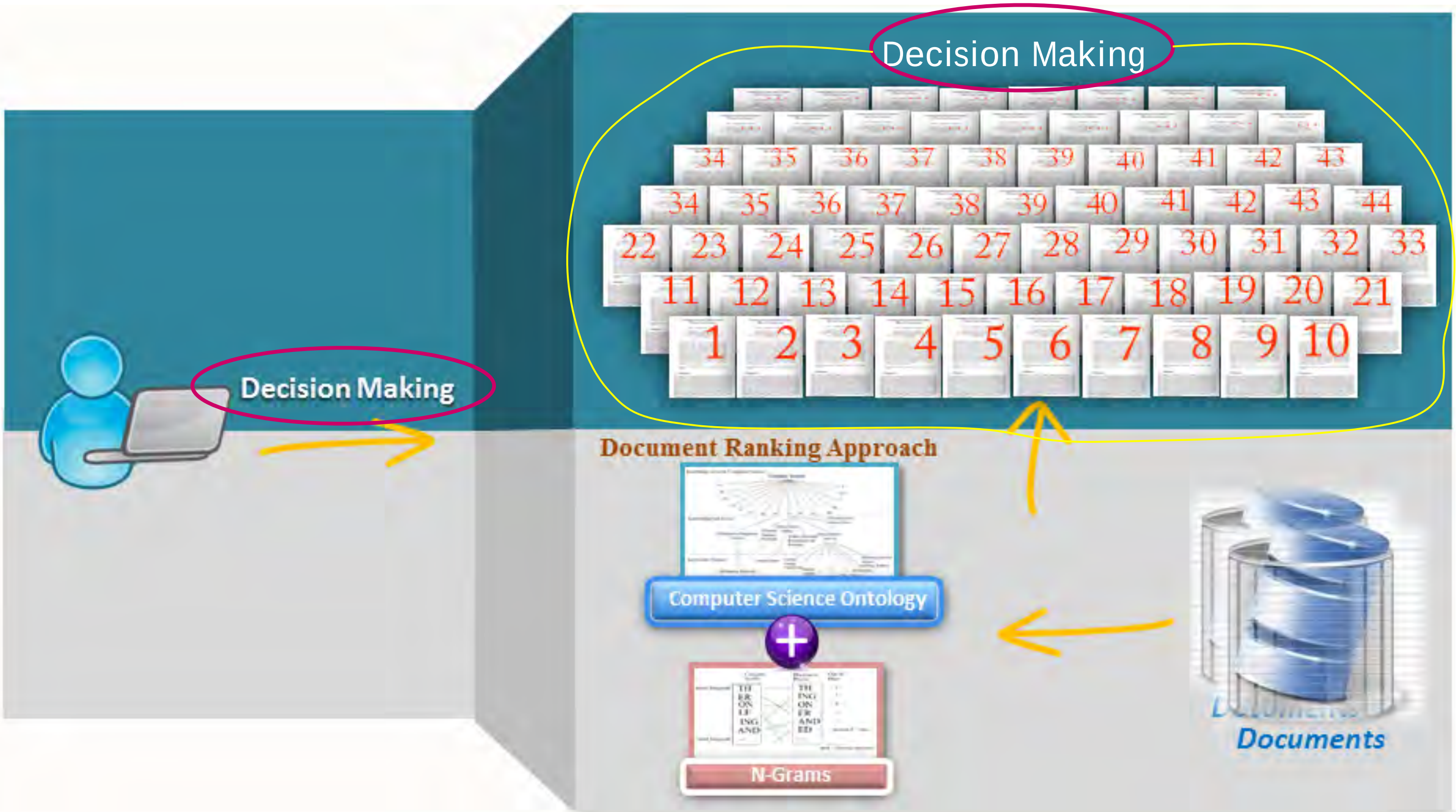
Documents ranking

02 Semantic Similarity Search (Re-Ranking)

2

Semantic Similarity Process

Re-Ranking

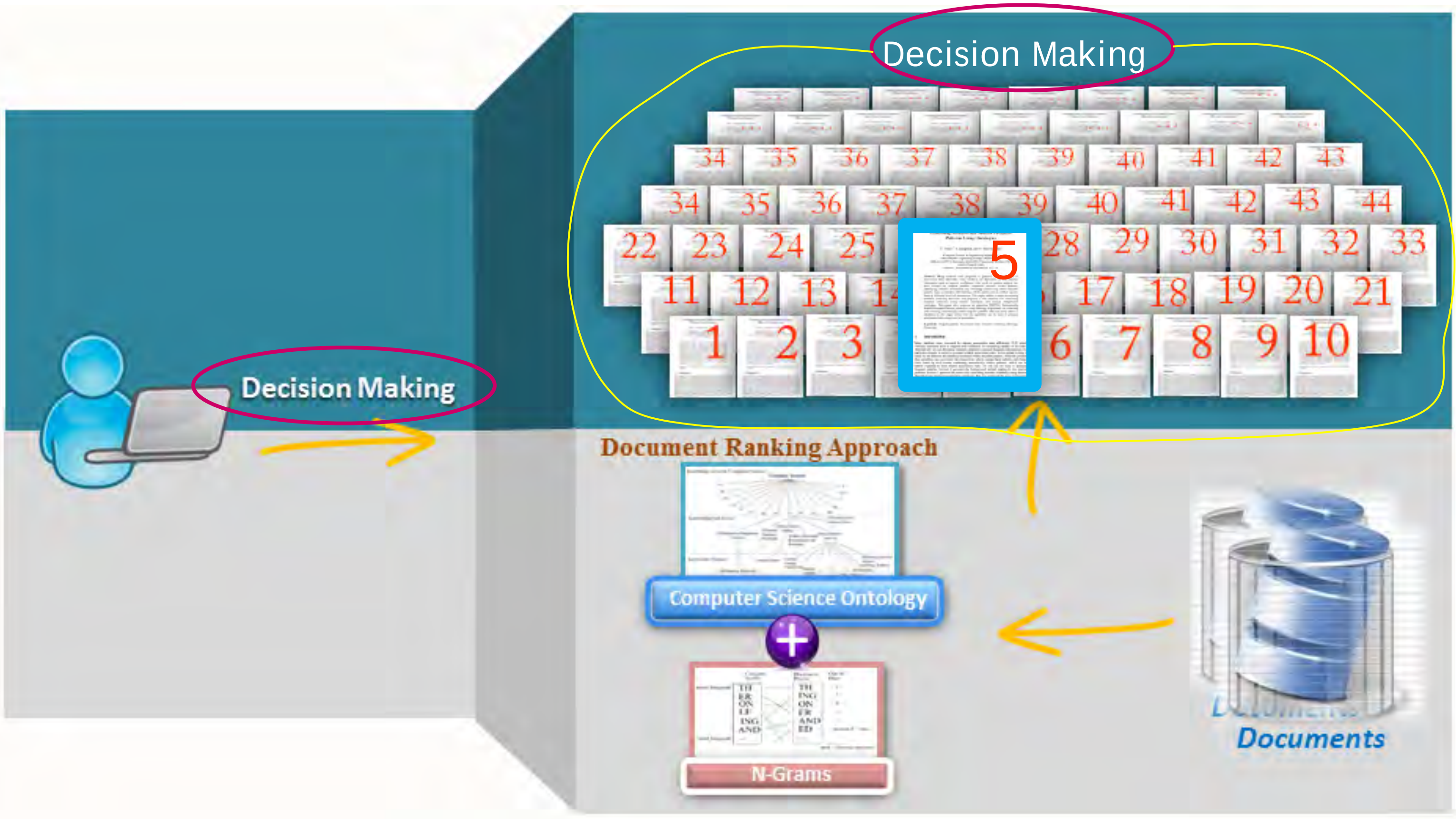


Document ranking

02 Semantic Similarity Search (Re-Ranking)

Semantic Similarity Process

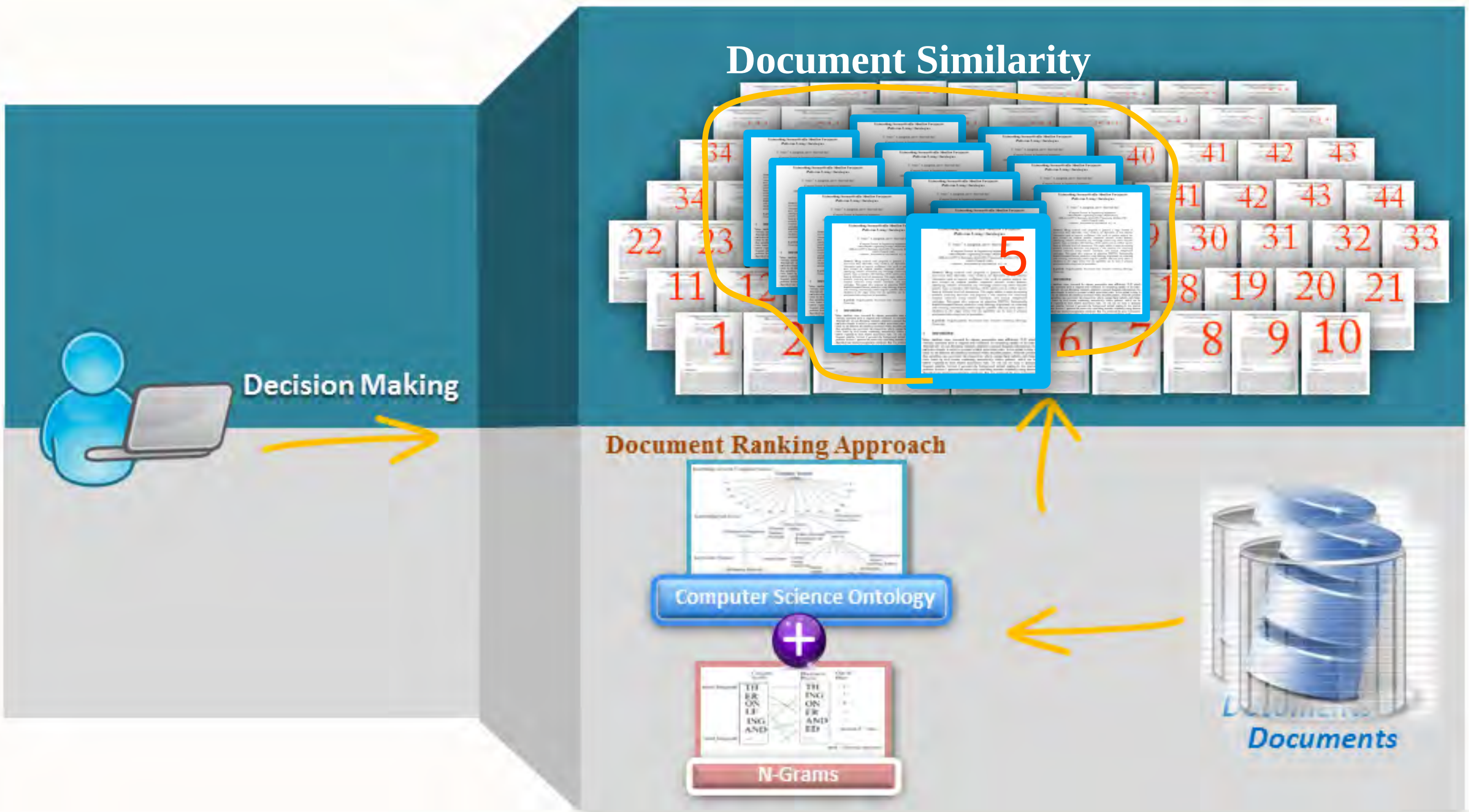
Re-Ranking



02 Semantic Similarity Search (Re-Ranking)

Semantic Similarity Process

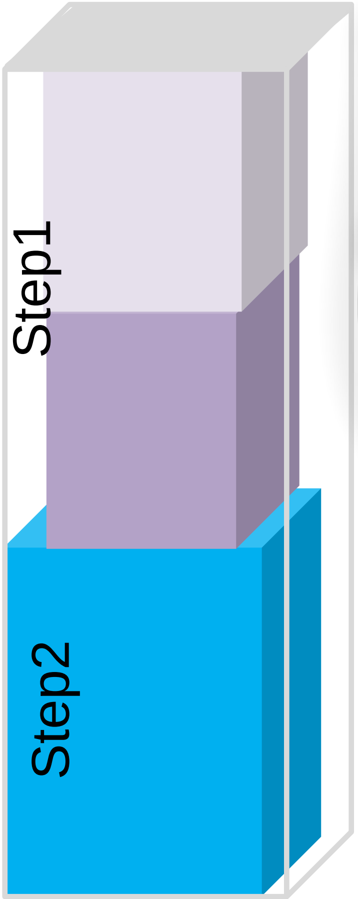
Re-Ranking



02 Semantic Similarity Search (Re-Ranking)

Semantic Similarity Process

Re-Ranking



Step1 : Calculating the Keyword-Keyword similarity matrix Scores

- Calculating the Subsumption Weight (Hypernym/Hyponym Hierarchy).
- Calculating the string Similarity weight.

Step2 : Calculating The Document-Document similarity matrix Scores

Document Similarity Process

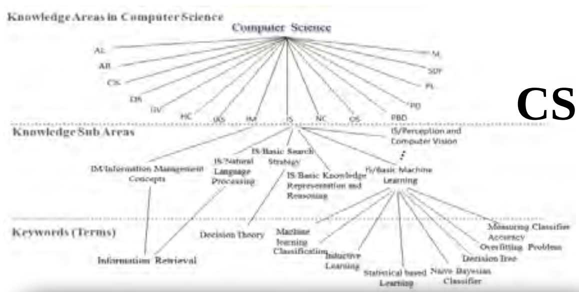
02 Semantic Similarity Search (Re-Ranking)

2

Semantic Similarity Process

Re-Ranking

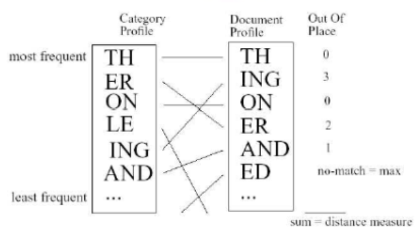
- **Calculating the Keyword-Keyword similarity matrix Scores**



CS Ontology

Hierarchical Relationship Weight

Tri-Grams +



Word Similarity Weight

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

Summation of Semantic Keyword Weight.

2

Semantic Similarity Process

Re-Ranking

- **Calculating The Document-Document similarity matrix Scores**

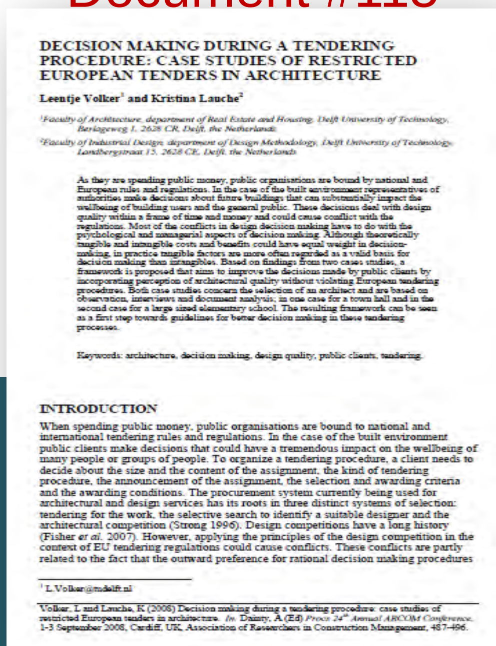
$$DSS(D_i, D_j) \cong \frac{\sum_{k=1}^{n} KWW_{ki} KWW_{kj}}{\alpha \sum_{k=1}^{n} KWW_{ki}^2 + (1-\alpha)\sum_{k=1}^{n} KWW_{kj}^2} (\alpha = \frac{1}{2})$$

$$= \frac{2\sum_{k=1}^{n} KWW_{ki} KWW_{kj}}{\sum_{k=1}^{n} KWW_{ki}^2 + \sum_{k=1}^{n} KWW_{kj}^2}$$

Ref : Dice, L. R. (1945).

Document - Document Similarity

Document #115


DECISION MAKING DURING A TENDERING PROCEDURE: CASE STUDIES OF RESTRICTED EUROPEAN TENDERS IN ARCHITECTURE

Leentje Volker[1] and Kristina Lauche[2]

INTRODUCTION

Document Keywords

Decision making
Architecture
Design quality
Public clients
Tendering

Hierarchical Relationship Weight
(Using CS Ontology)

N-grams Weight
(Trigrams)

Document #847


Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA, October 3-7, 2009

WeAT3.2

An Indoor Intelligent Transportation Testbed for Urban Traffic Scenarios

978-1-4244-5521-8/09/$26.00 ©2009 IEEE

721

Document Keywords

Decision making
Mobile robots
Sensor virtualization
Situation awareness
Testbed
Urban traffic

Document #115

Architecture

Design quality

Public clients

Tendering

→ Decision making ←

Hierarchical Relationship Weight
(Using CS Ontology)

N-grams Weight
(Trigrams)

Document #847

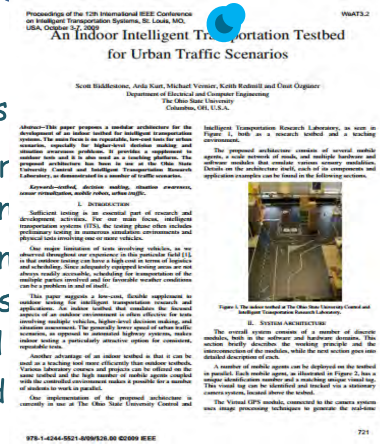
Mobile robots

Sensor virtualization

Situation awareness

Testbed

Testbed



Keyword-Keyword Similarity Matrix Scores

Document Keyword	Decision making	Architecture	Design quality	Public clients	Tendering	Mobile robots	Sensor virtualization	Situation awareness	Testbed	Urban traffic
Decision making	2	0.50	0.67	0.50	0.21	0.31	0.39	0.40	0.00	0.31
Architecture	0.50	2	0.75	0.75	0.00	0.57	0.50	0.56	0.00	0.25
Design quality	0.67	0.75	2	0.75	0.00	0.25	0.40	0.30	0.00	0.31
Public clients	0.50	0.75	0.75	2	0.00	0.62	0.25	0.30	0.00	0.37
Mobile robots	0.31	0.57	0.25	0.62	0.00	2	0.25	0.30	0.00	0.50
Sensor virtualization	0.39	0.50	0.40	0.25	0.00	0.25	2	0.80	0.00	0.30
Situation awareness	0.40	0.56	0.30	0.30	0.00	0.30	0.80	2	0.00	0.30
Urban traffic	0.31	0.25	0.31	0.37	0.07	0.50	0.30	0.30	0.08	2
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0[illegible]	[illegible]	[illegible]

Document-Document Similarity Matrix Scores

Document	#115	#847
#115	1	0.4346
#847	0.4346	1

Experiment Results

Table 1 : Document-Document Similarity Matrix Scores
(Keyword Matching Weight)

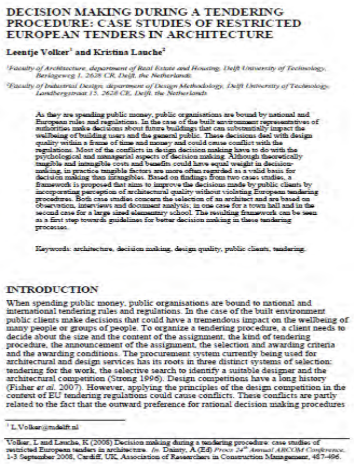
Paper ID	115	847	1055	553	584	Query
115	1	0.1818	0.2222	0	0	0.3333
847	0.1818	1	0.2	0	0	0.2857
1055	0.2222	0.2	1	0	0	0.4000
553	0	0	0	1	0	0
584	0	0	0	0	1	0
Query	0.3333	0.2857	0.4000	0	0	1

Table 2 : Document-Document Similarity Matrix Scores
(Using Our Semantic Weight)

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

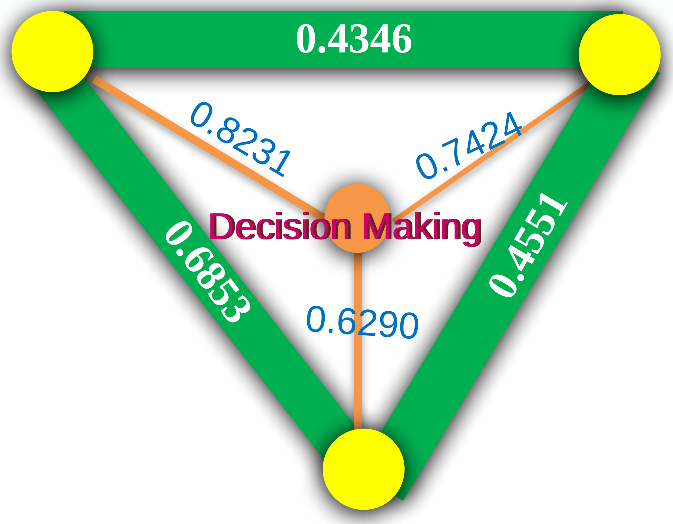
Experiment Results

Document #115


DECISION MAKING DURING A TENDERING PROCEDURE: CASE STUDIES OF RESTRICTED EUROPEAN TENDERS IN ARCHITECTURE

Leentje Volker[1] and Kristina Lauche[2]

INTRODUCTION



Document #847


Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA, October 3-7, 2009

WeAT3.2

An Indoor Intelligent Transportation Testbed for Urban Traffic Scenarios

Scott Biddlestone, Arda Kurt, Michael Vernier, Keith Redmill and Ümit Özgüner

Department of Electrical and Computer Engineering

The Ohio State University

Columbus, OH, U.S.A.

I. INTRODUCTION

II. SYSTEM ARCHITECTURE

978-1-4244-5521-8/09/$26.00 ©2009 IEEE

721

Document #1055


A Cordon and Search Model and Simulation using Timed, Stochastic, Colored Petri Nets for Robust Decision-Making

Paul Maxwell[1,2], Anthony A. Maciejewski[1], Howard Jay Siegel[1,3], Jerry Potter[1]

ISTeC People-Animals-Robots (PAR) Laboratory

Electrical and Computer Engineering Department

Computer Science Department

Colorado State University, Fort Collins, CO 80523-1373 USA

United States Army

paul.maxwell@us.army.mil

Keywords: robustness, decision-making, cordon and search, Petri Nets.

Abstract

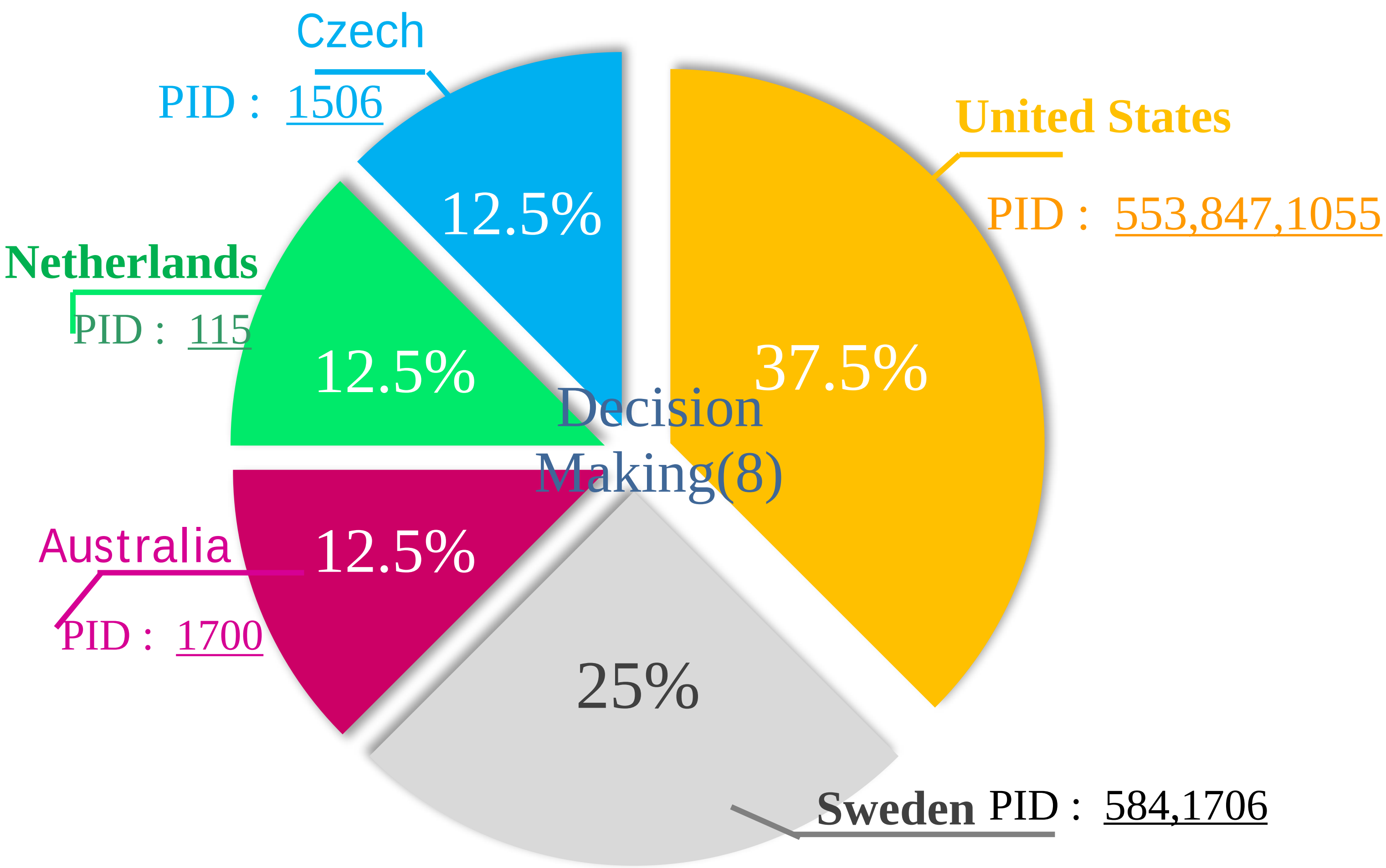
1. INTRODUCTION

02 Semantic Similarity Search (Semantic Similarity)

2

Selecting Semantic Context

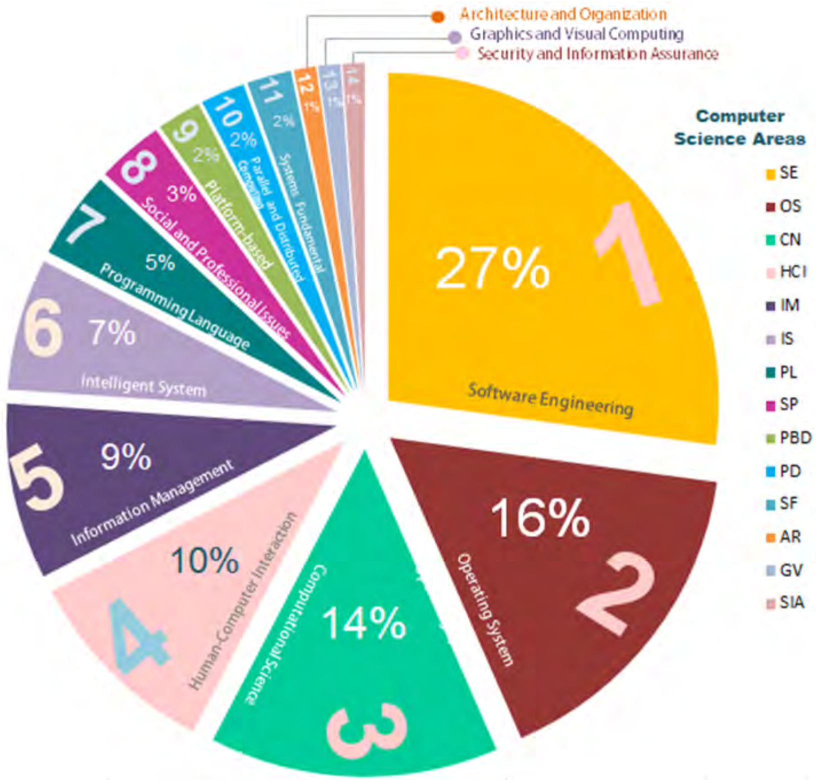
(Country)



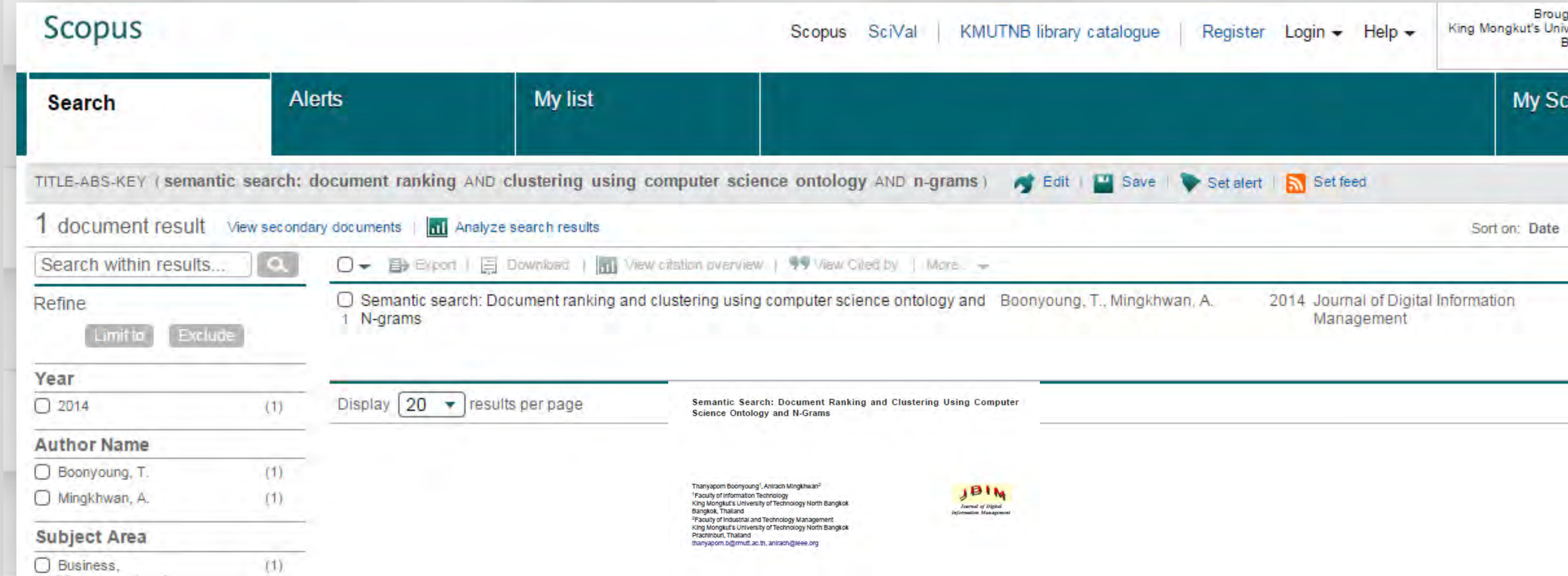
3

Query And Answering

Computer Science Research Area in UK



Journal Published



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

JDIM 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)

Thank you