Semantic Searching IT Careers Concepts Based on Ontology

Patsakorn Singto
Department of Information Technology, Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
E-mail: psingto@hotmail.com

Anirach Mingkhwan
Department of Information Technology, Faculty of Industrial Technology and Management
King Mongkut's University of Technology North Bangkok, Prachinburi, Thailand
E-mail: anirach@ieee.org

***Abstract*—Nowadays, there are a lot of IT Careers (ITC) which are not stored in a hierarchical structure of ITC specification and matching search based on keyword have widely use search which cannot acquire satisfactory search results. The essential reason is that traditional ITC search lacks semantic. In this paper, we propose searching ITC concepts based on ontology which is associated specification with semantic annotation. IT Careers Ontology (ITCO) defined structure consists of three main parts: IT Career Category, IT Skill and IT Education which are described with announcing the recruitment on web jobs ads that are instances of a models. The experimental results show that semantic search ITC can overcome limitations of search by comparison with traditional keyword search mode, and achieve higher recall ration and precision ratio.**

***Index Term*—semantic search, ontology, OWL-QL, IT careers.**

I. Introduction

As information IT Career (ITC) have many job posting that are listed on jobs advertisements web site, such as jobdbs.com, jobStreet.com, careerbuilder.com, etc., which have widely search format for users to retrieve acquire career on the internet. At present, two main disadvantages web search jobs [1]. Firstly, most of irrelevant careers in specifications perspective returned from web search jobs. Secondly, the display order of search result is rather in confusion. Thus, web job search cannot deal with search results effectively for those returned careers. The essential reason of these is traditional web job search lacks semantic, which is difficult for users satisfy their search requirements.

Ontology is a hierarchy of concepts with attributes and relations that defines an agreed terminology to describe semantic networks of interrelated information units. Ontology provides a vocabulary of classes and properties to describe a domain, emphasizing the sharing of knowledge.

The Semantic Web brings semantics to the current Web with formalized knowledge and data that computers can understand and deal with. Therefore, web searching could get advantage from the use of inference rules that are supported by ontology [2].

In this paper, we propose the semantic searching ITC specifications are properly in many perspectives such as search by career name, skill or education based on ontology. We extend the traditional TF/IDF(Term Frequency-Inverse Document Frequency) is often used as weighting factor in information retrieval, Apriori algorithm to reflect the relevance between ontology, Careers specification were created by Web Ontology Language (OWL) as the ontology definition language and SPARQL as the language for deductive query answering on the ontologies.

This paper is structured as follows. In Section I introduces the main concepts of semantic searching. In Section II, we review related works. In Section III, domain ontology is defined and constructed. In Section IV, we firstly describe the overall semantic search ITC model. In Section V, experimental results. Finally, Section VI concludes the paper and our future works.

II. Related Work

Domain ontology emerged as a mainstream in many applications. SC Wang and Yuzuru Tanaka [3] introduced a topic-oriented query expansion model based on the information bottleneck theory that classify terms into distinct topical clusters in order to find out candidate terms for the query expansion. Gubing Zou *et al.* [4] proposed the semantic annotation framworks and query expansion algorithms, which is inductive to implement effective documents annotation and semantic query expansion. Amar Nayak *et al.* [5] developed a semantic web mining for an educational domain as enterprise framework. The system helps to find suitable semantic data related to student, faculties and course for the clients. GuangJun Huang *et al.* [6] proposed an approach of

Manuscript received November 14, 2012; revised December 29, 2012.

doi: 10.12720/joams.1.1.102-106

expanding queries based on synonyms and hyponyms in the domain ontology, and measured with Information Gain. Brian Harrington [7] proposed a new approach to determine semantic relatedness, in which a semantic network is automatically created from a relatively small corpus using existing NLP tools. Miriam Fernandez *et al.* [8] presented a comprehensive semantic search model which, extends the classic IR model which, integrates the benefits of both keyword and semantic-based search. Juan Wang *et al.* [9] combined ontology and network curriculum resource management, and links curriculum knowledge point through the establishment of ontology, which the system can not only achieve the curriculum learning, but also the enquiry of semantic reasoning and knowledge.

These semantic annotation frameworks have limitations, which annotate documents with lexical database was not accurate to extract and compute concepts or instances. So we propose a novel semantic searching ITC method based ontology, which is inductive to implement effective documents annotation and semantic searching with adapt Apriori algorithm to define relevance.

III. Construction of Ontology

The purpose of the IT Career Ontology (ITCO) is to provide a central repository of classified career in varies organization, which define relating careers to require skill and knowledge. The design of our ontology is guided by requirement of job seeker, student, academy and organization. We use Web Ontology Language (OWL) for model. We define terms of IT skill and IT Education with computer field knowledge of ACM/IEEE-CS and IT career category was defined terms by ISCO-08. Therefore, this vocabulary describes the type of objects and concepts. Standard relationships as is-a, part-of, and instance-of predefined semantics.

Fig. 1, to explain some of structure ITC can be used semantic search. The IT Careers is a graph consisting of the top class is *IT_ Careers* as the domain of ontology and *IT Career Category*, *IT Skill* and *IT Education* are defined relationships as component of to *IT Careers*. The subclass of the *IT Skills* consists of *Technology and Method of_ Software*, *Technology Application*, *Element of_ Information*, *Computer Hardware Architecture and System_ Infrastructure*. The subclass of *Technology_ and Method of Software* consists of *Operating_ System* and *Programming* and subclass of *Programming* consists of *Object Oriented_ Programming* and *Web Programming*. In the represents ontology, between class and property. Three kinds of relationships (*part-of*, *is-of*, *instance-of*) as joint edges, domain ontology can be represented as a tree structure graph (TSG).

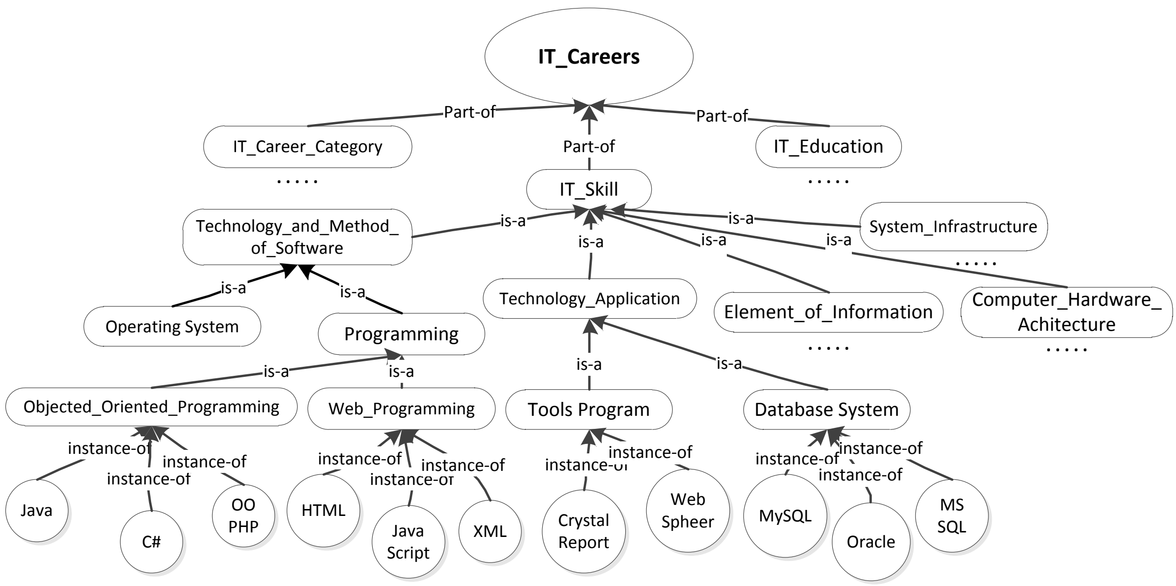

Figure 1. A portion of tree structure graph of ITC

RDF/RDFS is a language uses to represent information in the form of graph. RDF creates a data model for objects and relationships among them. RDFS provides basic vocabulary for describing properties of RDF resources. Class and instances would be defines taxonomies which object properties relates between instances of two classes, data type properties relations between instances of classes and literals with RDF/RDFS data types are show in Table I.

TABLE I. RDF/RDFS Type Properties

Classes	RDF/RDFS Type	
	Data Property	**Object Property**
IT_Careers	careerName: String	has_Education hasSkill
IT_Education	educationName: String	inverse_has_Education
IT_Skill	itSkillName:String	inverse_has_Skill

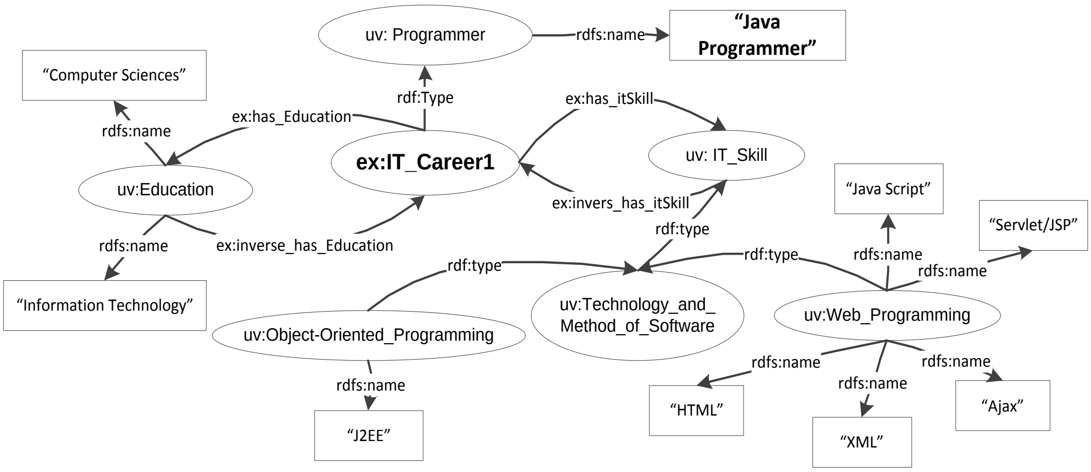

Figure 2. ITC to ontology mapping with RDF graph

ITC database consists of tables and tuples. Each tuple consists of a set of attribute values. The attributes of tuples is an RDF node. We define a semantics mapping as a process from database to an RDF graph in a final ontology. For example career name as "*Java Programmer*" show on Fig. 2. As "*IT_Career1*" name is "*Java Programmer*" and as instance of "*Programmer*". "*Education*" and "*IT_Skill*" are component of "*Java Programmer"* and has inverse property, that is, "*Computer Sciences*" and "*Information Technology*" are instance of "*Education*" is part of "*Java Programmer*" and "*Java Programmer*" has "*Education*". And "*Java Programmer*" has "*IT Skill*" and "*IT Skill*" is part of "*Java Programmer*". "*Technology_ and_ Method_of_ Software*" is subclass of "*IT_Skill*" while "*Object_Oriented_Programming*" and "*Web_Programming*" are subclass of "*Technology_and_Method_of_ Software*". "*J2EE*" is instance of "*Object_ Oriented_ Programming*" and "*HTML*", "*XML*", "*Ajax*" are instance of "*Web_Programming*".

IV. SEMANTIC SEARCHING IT CAREER

A. *Semantic Searching ITC Model*

The overall searching framework diagram is shown in Fig. 3. The processing consists of the following steps:

1) *The system transforms keyword list input as formal SPARQL query.*

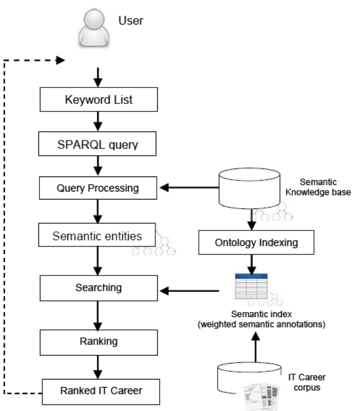

Figure 3. Semantic searching framework

2) *The SPARQL query is executed to a semantic knowledge based on query processing algorithm, which returns a list of semantic entities that as instances match with conditions in the formal query.*

3) *The documents are indexed with instances that are retrieved with semantic indexing algorithm.*

B. *Semantic Indexing*

To retrieve IT careers that are related to a user's query and to rank them according to their importance**.**

The relevance between IT careers and concepts in ontology must be measured. We quantify such relevance in three steps. First, we represent the relevance between career and components with the TF-IDF algorithm [10]; then we define several levels of relevancy between components and concepts with respect to their positions in an ontology; and finally we get the relevance between careers and concepts by combined the results of these two steps and store it in ontology indexing. These steps are detailed as follows.

Step 1. Calculate the basic TF-IDF weight algorithm of term i to career j is calculated as equation (1).

$$w_{ij} = log\left(\frac{N}{1+n_i}\right) \times \frac{freq_{ij}}{max_i\left(freq_{ij}\right)} \quad (1)$$

Let N be the total of careers and n_i number of careers which appears term t_i. The term of frequency of term i which let $freq_{ij}$ be frequency of term t_i in the careers c_j. The maximum is computed over all terms i mentioned in the career j. Then the multiply by inverse total career for c_i.

Step 2. Define the levels of relevance between ontology members. We define four relevance levels, including direct, strong, normal and weak. Each of them can be given a number. These four levels are given 1.0, 0.9, 0.6, and 0.3 respectively shown in Table II. We define weak relevance with frequent itemset generation in Apriori Algorithm. Apriori is the first association rule mining algorithm that the use of support-based pruning to systematically control the exponential growth of candidate itemsets. Association rule has a support level and a confidence level. The support is the percentage of the population which satisfies the rule and the confidence is percentage in which the consequent is also satisfies rule. We selected to relate two itemsets that minimum support as 0.5 values. We show example relevance between member and term on Fig. 4. The relevance between an ontology level and a term t_j is calculated as equation (2).

$$R(c,m) = \sum_{i=1}^{N_c} r(t_i, m) \quad (2)$$

TABLE II. RELEVANCE BETWEEN ONTOLOGY LEVELS

Relevance Level	Concept/Individual	Relevance Quantification
Direct	Synonyms	1
Strong	Hyponyms, Hypernyms (is-a)	0.9
Normal	Meronyms, holonym (part-of)	0.6
Weak	Support>=0.5	0.3

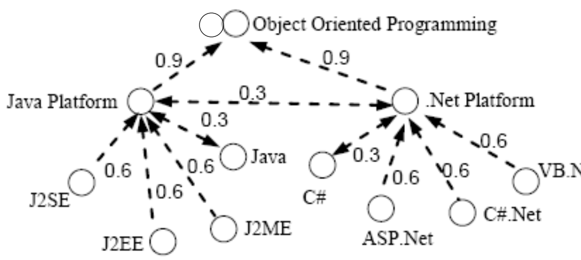

Figure 4. An example relevance between member and term

Step 3. Calculate the extended term of career matrix: where R_{ij} is the relevance between the term i and career *j* calculated with equation (3).

$$W = \{w_{ij}\} = \{w'_{ij} \times R_{ij}\} \quad (3)$$

C. Query Processing and Searching

SPARQL Protocol and RDF Query Language (SPARQL) are ontology query languages. We are used to extract specific information form RDF graphs.

The search process begins with the keyword list forms of a user's query. We transform keyword list forms to formal SPARQL query, which returns a list of instance tuples that satisfy the query. The SPARQL queries supported by our model can express conditions involving domain ontology instances, career properties such as *careerName, skillName, educationName*, etc. The query keywords are assigned a weight that classified relevant documents.

As in the example, the WHERE sentence to a query contains a set of standard RDF form (*subject, predicate, object*). The retriever will get all the documents that correspond to bound concepts in these triples. For example, if the user searches "What is the IT career have a skill ass Java or J2EE skill" show on Fig. 5.

```
prefix ns:<http://www.owl-ontologies.com/ITOnt.owl#>
SELECT  ?careerName  ?skillName
WHERE { ?career ns:careerName ?careerName.
?career ns:hasSkill    ?skill.
?skill     ns:skillName ?skillName.
FILTER (?skillName= "Java" || ?skillName="J2EE")
Order by weight}
```

Figure 5. A query example in SPARQL

Semantic search computes a similarity value between the query and each career, using the Vector Space Model (VSM) [10]. We represent each career in the search space as a career vector, where career vector value is the weight of the annotation of the career with ontology concept. We defined query vector element corresponding to the variable weight in ontology. Results set tuple have more than one the same instance appears as satisfying value. The similarity measure between a career x_i and the query q is computed as:

$$sim(q, x_i) = \frac{q \cdot x_i}{|q| \times |x_i|} \quad (4)$$

V. EXPERIMENTAL AND PERFORMANCE MEASUREMENT

A. Experimental

The experimental use testing IT Career from JobsDB.com amount 50 careers and based on the framework above, we take the through our semantic search model using five term list specifications sample and compared with keyword search model traditional keyword search. Table III shows the results of our experiment implemented framework which term IT skill list as require a career name and result as shown a list of relevant careers, which order follow by relevant weight.

TABLE III. EXPERIMENTAL RESULTS SEMANTIC SEARCH

Term List	Career List
Java, J2EE	Java Programmer Analyst, Software Developer
.Net, Ajax, C/C++	Application Programmer, .Net Programmer
PHP,HTML, CSS	Software Developer, Programmer
Ajax, CSS, JSP	Software Developer, Java Programmer
Ajax, JavaScript, MySQL	.Net Programming, PHP Programming

B. Performance Measurement

In this paper, recall ratio and precision ratio are applied to evaluate efficiency of search results. Recall refers to proportion of retrieved related careers out of all rerated career in system. Precision is defined as proportion of retrieved rerated careers.

$$Re\,call = \frac{retrieved\ Re\,latedDocuments}{all\ Re\,latedDocumentsInSystem} \quad (5)$$

$$Precision = \frac{retrieved\,RelatedDocuments}{retrievedDocuments} \tag{6}$$

In order to compare and analyze search efficiency among two kinds of search method as keyword search and semantic search, Table IV show the results of the evaluation using IT skill name 5 queries, which performance of system in terms of precision and recall. The semantic search outperforms the keyword search in all queries, which the average precision and recall values are shown at bottom.

TABLE IV. COMPARISON EFFICIENCY OF TWO SEARCH METHOD

Query	Keyword Search		Semantic Search	
	Recall	Precision	Recall	Precision
1	0.50	0.67	0.50	1.00
2	0.50	0.75	0.65	0.80
3	0.33	0.50	0.50	0.75
4	0.30	0.33	0.45	0.55
5	0.33	0.17	0.35	0.50
Average	0.39	0.48	0.49	0.72

VI. CONCLUSIONS

Our study in this paper shows that ontology-based searching is essential and feasible for supporting semantic searching results. The proposed framework can be viewed and extension of traditional VSM with semantic support. Future effort around this framework will be focused on the optimization of the numbers given to relevance levels, and the improvement of the reasoner's performance. We process focuses on lessening the execution time by querying the results with the resources from databases after extracting resources from ontology as well as providing abbreviated inference information.

ACKNOWLEDGMENT

The authors wish to thank King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

REFERENCES

[1] Z. Xiaogang and L. Mingshu, "Research on search engine technologies," *Computer Engineering and Application*, no. 24, pp. 67-70, 2001.

[2] L. Li, Z. Dong, and K. Xie, "Ontology of general concept for semantic searching," in *Proc. Computer Modeling and Simulation Conf.*, 2010, pp. 81-84.

[3] SC. Wang and Y. Tanaka, "Topic-oriented query expansion for web search," in *Proc. 15th international conf. World Wide Web*, New York, 2006, pp. 1029-1030.

[4] G. Zou, B. Zhang, Y. Gan, and J. Zhang, "An ontology-based methodology for semantic expansion search," in *Proc. Fuzzy Systems and Knowledge Discovery Conf.*, 2008, pp. 453-457.

[5] A. Nayak, J. Agarwal, VK. Yadav, and S. Pasha, "Enterprise architecture for semantic web mining in education," in *Proc. Computer and Electrical Engineering Conf.*, 2009, pp. 23-26.

[6] G. Huang, P. Musilek, and J. Sun, "Searching messages based on semantic content," in *Proc. Wireless Communications, Networking and Mobile Computing Conf.*, 2008, pp. 1-4.

[7] B. Harringto, "A semantic network approach to measuring relatedness," in *Proc. Computational Linguistics Conf.*, 2010, pp. 356-364.

[8] M. Fernandez, I. Cantador, V. Lopez, D. Vallet, P. Castells, and E. Motta, "Semantically Enhanced Information Retrieval: An ontology-based approach," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 9, no. 4, pp. 434-452, 2011.

[9] J. Wang and Y. Li, "Design and implementation of curriculum resource management model based on domain ontology," in *Proc. Computer Science and Information Technology Conf.*, 2010, pp. 217-221.

[10] DL. LEE, H. Chuang, and K. Seamons, "Document ranking and the vector-space model," *IEEE Trans. on Computer Graphics and Computer-Aided Design, Software,* vol. 14, pp. 67-75, March/April 1997.

Patsakorn Singto, Female, born on March 1st, 1971. She earned master's degree in Information Technology from King Mongkut's University of Technology North Bangkok, Thailand in 2003. She currently is doctoral student in faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Thailand. Her research areas cover knowledge base system, XML technology and semantic web.

Anirach Mingkhwan, Male, born on August 7th, 1969, Ph.D., 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 instructor of Information Technology and the current 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, embedded systems, wireless mesh networks, Ad Hoc Networks, Computer Forensics, Digital Investigation, Vehicular Ad Hoc Networks, and Mobile Ad Hoc Networks.