Similarity Measurement in the Hybrid of Semantic Web Search Engine

Noryusliza Abdullah, Rosziati Ibrahim

Faculty of Computer Science and Information Technology , Universiti Tun Hussein Onn Malaysia, Malaysia

yusliza@uthm.edu.my

Faculty of Computer Science and Information Technology , Universiti Tun Hussein Onn Malaysia, Malaysia

rosziati@uthm.edu.my

ABSTRACT

Semantic Web approach with the assistance of ontology is widely used to give more reliable application in retrieving information and knowledge. It is capable to discover the World Wide Web (WWW) that is presented in natural-language text. Based on previous research, incorporating categorization with ontology concept has proven to give better results. However, performing hybrid of the search engine using another technique that is user profiling has a promising potency in enhancing the searching process. Utilizing searching time and giving relevant results are the contributions of this research. The proposed hybrid techniques integrate ontologies, categorization and user profiling concept. In user profiling, similarity measure is adopted in making comparison between two different ontologies. WordNet and UTHM Onto are the independent ontologies used in this process. The preliminary experimental results have given interesting results in terms of data arrangement and time usage.

Indexing terms/Keywords

Semantic Web Search; Ontology; Categorization; User Profiling; Similarity Measure

Academic Discipline And Sub-Disciplines

Information Retrieval; Semantic Web; Ontology

SUBJECT CLASSIFICATION

Information Technology Classification

TYPE (METHOD/APPROACH)

Experimental

INTRODUCTION

Many sources have given good numbers of information and knowledge. However, retrieving the real meaning of data is often fails to give the desired result. The maturity of Information Technology should provide better capturing mechanism to obtains more information and knowledge and make it meaningful to our purpose.

In retrieving knowledge or information, semantic web and ontology are discussed in detail to get the exact meaning of any data or information. Information on the Web is presented in natural-language text that understandable by human but difficult for computers to interpret. Thus, Semantic Web approach is widely used to give more reliable application.

Mikroyannidis [1] explains that Semantic Web is able to give information a well-defined meaning and better cooperation between computers and people. RDF and Microformat features allow web sites to expose semi-structured information for machine use [2]. They will deliver knowledge and assist in decision making. Ontology, on the other hand is an explicit specification of a conceptualization [3].

In applying the Semantic Web or Semantic Web Search Engine, ontology is commonly discussed. dâ€™Aquin and Noy [4] states that data interoperability property from ontologies which permits sharing and reusing features, is a key promises of the Semantic Web.

While ontologies are capable in giving good outcome, researchers are trying to enhance searching method using categorization/clustering technique [5] and user profiling/personalization [6; 7; 8]. Although previous researches are capable to give good results, Abdullah and Ibrahim [9] propose the hybrid of those three techniques (ontologies, categorization and user profiling) to ensure reliability criteria is fulfilled in retrieving knowledge. The techniques are consolidated to give more relevant searching particularly in the user's perspective. The proposed technique will provide search results based on users' own data. Our contributions are to utilize searching time and provide relevant results.

The remainder of this paper discuss the Related works followed by The framework of Semantic Web Search Engine, Case study and the Conclusion.

RELATED WORKS

In this section the details of Semantic Web, Search Engine, Ontology resources, Categorization (or clustering) and User profiling (or personalization) is discussed. User profiling part is indicates in detail which include two ontologies, WordNet and UTHM Onto mentioned earlier. Similarity among the lexical ontology resource and the domain ontology is calculated.

Semantic Web

The discussion on semantic web often leads to confusion due to the different meaning by researchers. In this subsection, concise descriptions on that matter are presented. Referring to Jiang [10] the benefit of using it is the ability for machine-understood descriptions of meaning. The web helps us to reach information that we search and other data related to it. Thus, Semantic Web is not just sharing text of a page but data and facts as well [11].

Other motivation to use Semantic Web is it helps in collecting data together from the web [11]. Referring to Mikroyannidis [1], Semantic Web is better than conventional web because of the ability to handle unstructured content. Semantic Web can overcome this problem by using software agent that can enhancing search precision and enabling logical reasoning. Joo [12] views semantic web has a potential to implement semantic integration and reduce information overload.

According to Janev and Vrane [13] this is the popular area in the Information and Communication Technology field. Many research efforts are conducted to improve traditional web and making the content available on the semantic web. In line with this thought, Edwards [11] explains moving from HTML to XML is the original plan for the semantic web. Loopholes in HTML addressed by Linked Data that connect data, information and knowledge on the semantic web using Uniform Resource Identifier (URI) and Resource Description Framework (RDF).

Search Engine

In our research, we concentrate on the search engine. Internet and the World Wide Web have a tremendous growth. The evolution begins from Web 1.0, Web 2.0 and moving towards Web 3.0. In Web 1.0, WWW is merely focusing on static web pages whereby information updates are managed by webmaster. The emergence of Web 2.0 change the internet perspective to collective information among users with shared control. Information are changeable by every individual. Web 2.0 has major features including social networking sites, user created web sites, self- publishing platforms, tagging, and social bookmarking. These lead to new technologies such as Facebook, Twitter and Youtube.

The increase of users' need and requirement encourage attempts to give improvement in delivering a better usage of the World Wide Web especially when Web 2.0 has no longer fulfill new requirements. Web 3.0 (also known as Semantic Web and Web of Data) is expected to convey the users' need [14].

Primary need to explore the internet is search engine [15]. Since Web 1.0, search engine is widely exposed to retrieve information on the web. The specific purpose is to search document using keywords. In order to execute the searching, crawler, spider or bot (robot) is used to fetch documents in WWW. Indexer or catalogue will index those documents based on the words.

In conjuction with the emergence of Web 3.0 which supports Semantic Web, research on Semantic Web search engine has actively conducted. Even though the functionality of the Semantic Web search engine is resemble typical search

engine such as Google and Yahoo, in contrast, it helps us to reach a collection of information and data related to it in the Semantic Web environment built on RDF. Hakia is an instance of the Semantic Web search engine.

It is not sufficient to rely on keyword alone in finding the most related web site because substantial number of results will cause time incremental to analyze the results. According to Aguilar-Lopez [16], extracting semantic of the content by using ontologies is the method to overcome the problem. Details on the ontology resources are explained in the next section.

Ontology Resources

Ontology is the heart of the Semantic Web. It is a domain and knowledge representation [11; 17]. In consonance with Hepp [18], ontologies are the vocabulary that can be used to express a knowledge base while Diez-Rodriguez et al. [19] discusses that the intention to represent concepts in ontologies is to improve knowledge searching and discovery mechanisms.

In-depth researches are conducted on ontologies because of the function as the backbone for the semantic web [13]. Joo [12] state that research on ontology is necessary to ensure the diffusion of the semantic web. In addition, ontology- based knowledge organization can contribute to express the contents of information elements and semantic relations between them. It can also support semantic reasoning and retrieval [20]. Furthermore Maier et al. [21] explains that documented knowledge which spread across multiple sources requires identification and visualization with the help of knowledge maps and integration supported by ontologies as a manager to semantic content.

However, in the interest of ensuring ontologies and metadata to represent information correctly, they need constant updates and maintenance [1]. In order to accomplish the aim, Web Ontology Language (OWL) is used. It is a semantic markup language for publishing and sharing ontologies on the World Wide Web and used to describe the classes and relations between them [22]. Still, according to Cardoso [17], building ontology is more complex in terms of logic and structure compared to building software. The main goal of ontology engineering is to produce useful, consensual, rich, current, complete, and interoperable ontologies.

In building ontologies, linking them to the knowledge organization systems is the main priority to increase interoperability and data accessibility [18]. The highest methodologies adoption in develop ontology is Methontology. Ontologies development needed an editor. There are several editors including Protégé, SWOOP, OntoEdit, OntoStudio, Topbraid and many more used in commercial and research area. Among all, Protégé is the most used editor due to the support of wide variety of plugin and import formats and it's free open source.

In accordance with D'Amato [23], combining semantic web search with ontological background is a promising research approach. New semantic web applications discover ontologies on the web. Exploring large-scale semantics need to perform certain tasks: Find relevant resources, Select appropriate knowledge, Exploit heterogeneous knowledge sources and Combine ontologies and resources [24].

In semantic web, ontology resources are used to interpret the natural language properly. Many ontology resources are available and they grow rapidly. Instance of some ontology resources are WordNet, Cyc, Basic Formal Ontology (BFO) and General Formal Ontology (GFO). Table 1 shows the characteristic of specified ontology.

Table 1. Ontology characteristic

Ontology	Characteristic
Wordnet	Lexical reference system.
Cyc	Provides knowledge base of everyday common sense knowledge.
Basic Formal Ontology (BFO)	Support domain ontologies developed for scientific research
General Formal Ontology (GFO)	Specialties in persistence and time model.

Although there are several ontology resources are available, specifically WordNet is selected. It is due to the suitability in finding similar English words through lexical resource. Compared to the similar product that is CYC, Wordnet has become an ideal tool for disambiguation of meaning, semantic tagging and information retrieval based on its designed for manipulating by computers and cost-free use [25]. English language is emphasized due to large number of internet

resources are written in that language. Information retrieval using other languages such as Malay, Japanese, German, France and several others are possible to implement, nevertheless, it is out of scope for this paper.

Categorization/Clustering

Extension to the current approach, Trillo et al. [5] proposes categorization or clustering method which turns up with a semantic technique to group the output of searching keywords into different categories. They use online ontologies to define the possible categories.

User Profiling/Personalization

Research on personalization or user profiling in the semantic web is actively conducted. Jie [6] uses information on the homepage for profile extraction. Data for instance, interest and publications are extracted to get more information on users. Other researchers are based on the history of visited site for personalization.

In order to improve browsing result, personalization mechanism is used. This mechanism is based on user preferences and monitoring process of user navigation. Antoniou et al. [7] suggests the method of suggesting highly accessed pages from the past users' navigational patterns to the new users. This method has overcome very frequent accessibility for short periods of time using advance data structures technique. Yoo [8] supports effective retrieval of personalized information on the semantic web by using hybrid query processing method. The hybrid of two methods, query rewriting method and reasoning method are able to process query when individual requirements change.

Many researchers are using user profiling and personalization term interchangeably and refer them as the same entity. However, some researchers adopt them as two different things. Personalization refers to the navigational behaviour while user profiling is user's personal data. User profiling term is used from now onwards to avoid confusion. While most researchers are concentrating on browsing history and using web data for personalization or user profiling, we choose to hybrid our Semantic Web Search engine using data in our Relational Databases (RDB) to get more info on users.

SEMANTIC WEB SEARCH ENGINE

In this section, the proposed hybrid Semantic Web Search is presented.

Semantic Web Search Construction

In retrieving knowledge, there are several techniques can be implemented. Semantic Web is chosen based on certain advantages stated in the previous section. Ensuring results obtained are more reliable, method in Trillo et al. [5] is used with modification in user profiling concept.

Search Result based on User Profiling

This research focuses on University Tun Hussein Onn Malaysia (UTHM) dataset by connecting to the relational databases. Emphasizing on the user profiling, members' own data are extracted and used to ensure results are more reliable in user's perspective. Components that need to be examined are Staff/Student ID, Staff Name, Faculty ID and Faculty Name.

Querying the data needs five joined tables from Staff and Student databases of UTHM. They are Student, Faculty for Student, Staff, Staff Detail and Faculty for Staff.

Data Description

Important attributes are listed in Table 2 and Table 3 with the structure and description. The data structure is based on real data from UTHM's Relational Database (RDB). In the implementation phases, actual UTHM data will be used as datasets. Although Staff and Student are involve in this research, only Staff data are used in this paper.

Table 2. Datasets structure – Faculty table

Field	Structure	Description
facID	Varchar2 (3)	Faculty in UTHM.
facName	Varchar2 (50)	Name of faculty.

Table 3. Datasets structure – Staff table

Field	Structure	Description
staff ID	Varchar2 (10)	ID for every staff. Unique. Used as identification.
staffName	Varchar2 (50)	Name of staff.
facID	Varchar2 (3)	Faculty for staff.

CASE STUDY OF THE SEMANTIC WEB SEARCH ENGINE

This section describes a case study of Semantic Web Search Engine for UTHM members using three techniques: ontology, categorization and user profiling. Total of 968 academic staffs from UTHM are expected to utilize this finding. However, for testing requirement, only staff from Information Technology (IT) faculty which selected randomly will undergo the testing phase.

Step 1 - User Identification

The goal of this process is to capture user's profile. Figure 1 shows the Graphic User Interface (GUI) for identification. This search engine classify user's faculty. To facilitate uses, computer's data stored in web log might be used to avoid users from enter ID every time they use this application. Otherwise, ID and password are verified using Lightweight Directory Access Protocol (LDAP) data. Users are using the same ID and password for almost all systems in UTHM. This will avoid problems in remembering those data for every system.

Fig 1: GUI of user identification

Step 2 - Categorization or Clustering

In this step, user enters keywords. They are then mapped with WordNet ontology. The categories are listed into specific group based on sense and synset. Tables used in this processes are shown in Figure 2.
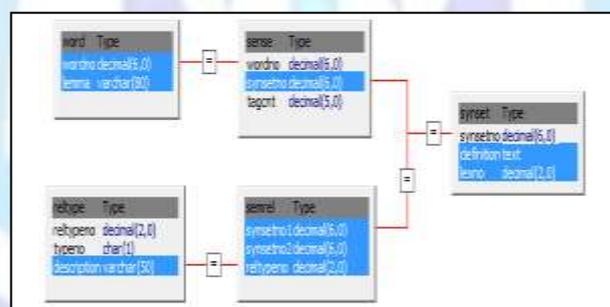
Fig 2: Categorization or clustering based on WordNet

Step 3 - User Profile and UTHM Ontology Mapping

UTHM ontology as shown in Figure 3 is developed to ensure changes are not done to the database. Modification to the databases will affect current systems since we use actual UTHM datasets. After clustering, the user's faculty captured earlier is compared with UTHM ontology.
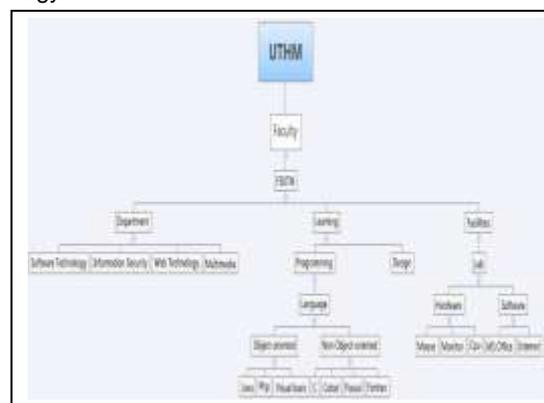
Fig 3: UTHM Ontology

Step 4 - Ranking

In this final stage, categories are ranked depending on user's data. This Semantic Web Search use entered ID as identification. Name is captured from the RDB.

Using previous framework, categories are mixed up and listed randomly. Excessive numbers of random categories will cause confusion. Hence, we are expected to produce results that are relevant towards user preferences by rank the categories using user profiling technique. In this study, ranking activity is utilizing Similarity Measurement that discussed below.

1) Similarity Measure

In this subsection, we are presenting a similarity measure. Two independent ontologies are compared. They are WordNet and UTHM Ontology. The similarity assessment is based on the Feature Matching [26]. In this assessment it indicates entity classes similarity. To perform this assessment, depth of entity_classes (α) is required. Connecting two independent ontologies and get the distance of these entity classes is conducted by creating imaginary and more general entity class called 'anything'. Figure 4 shows connected ontologies between WordNet and UTHM Onto.
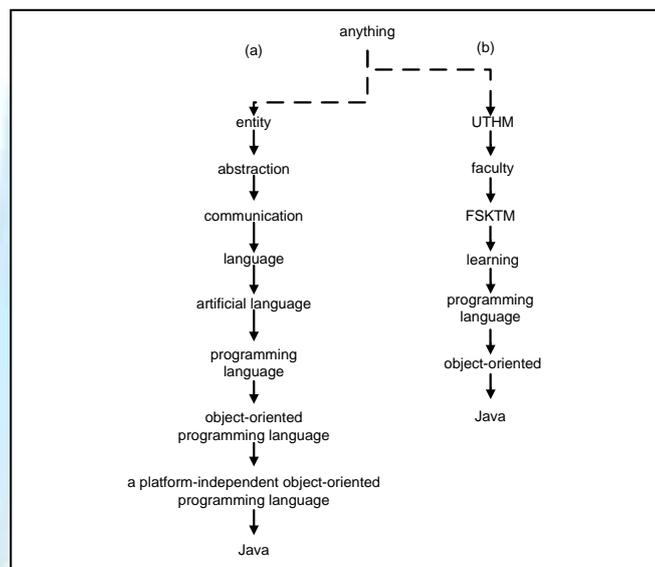
Fig 4: Connecting independent ontologies. (a)WordNet (b)UTHM Ontology (*anything corresponds to an imaginary root)

The connection of two ontologies has permitted in depth finding. Subsection 2 below shows the WordNet depth.

2) WordNet Depth

For the assimilation purpose, 'Java' keyword is used as a sample. WordNet as a lexical ontology is using hypernym-hyponym concept. It acts as a parent-child relationship. Traversing through the hypernyms of word will contribute in depth searching. Every word, sense (and synset) is represented with number. Figure 5 shows the route from word to root for 'Java'.
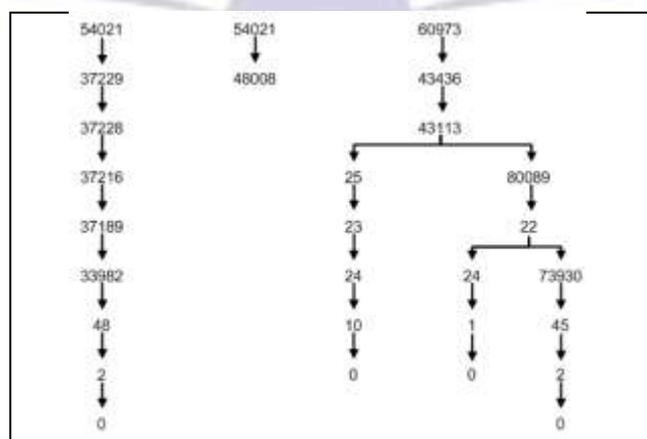
Fig 5: Synset in WordNet with hypernyms

For simplicity, only Java programming language depth is calculated. SQL is used over WordNet database and recursive technique is utilized in acquiring its depth. Figure 6 is the SQL command:

```
SELECT t1.synsetno1 AS parent,

GROUP_CONCAT(t2.synsetno1  ORDER BY t2.synsetno1 )
AS node,

COUNT(*)  FROM semrel AS t1

INNER JOIN semrel t2 ON

t1.synsetno2 = t2.synsetno1

where t1.synsetno1 = 43113 and t1.reltypeno in
```

Fig 6: SQL command to retrieve depth from node to root

3) UTHM Ontology Depth

UTHM Ontology is built based on the necessity in getting user's profile. SPARQL (Simple Protocol and RDF Query Language) is used to query the ontology. Depth calculation from word to its root is done by running the SPARQL statement until no superclasses are obtained. Both depth from Wordnet and UTHM Onto are used in the next subsection.

4) Feature Matching

The obtained depths from both ontologies are used in getting depth of entity_classes (α) in (1) or (2).

$$\alpha(a^w, b^u) = \frac{depth\,(a^W)}{depth\,(a^W) + depth\,(b^u)} \quad \{\ if\ depth\,(a^W) \leq (b^u) \quad (1)$$

$$= 1 - \frac{depth\,(a^W)}{depth\,(a^W) + depth\,(b^u)} \{\ if\ depth\,(a^w) > (b^u) \quad (2)$$

Where α denotes the depth of the entity classes, w is a WordNet ontology, u is a UTHM ontology, a and b are entity classes, depth (a^w) is the shortest path from the entity class to the imaginary root for WordNet and depth (b^u) is the shortest path from the entity class to the imaginary root for UTHM onto.

From the α, similarity measure using set theory/model of Tversky [27], Feature Matching measurement is conducted. In this matching process Parts, Functions and Attributes are classified.

Entity_class features for Java (Programming language) in WordNet and UTHM Onto are listed in Figure 7. Parts, Functions and Attributes will have their own data and calculated in (3).

```
<entity_class>::= entity_class { Name: {java}

              Description: < a platform-independent object-

              oriented programming language>

              Is_a: <object oriented programming language*>

              Part_of: <part_of>

              Whole_of: <whole_of>

              Parts: <parts>

              Functions: <functions>

              Attributes: <attributes>
```

Fig 7: Features in wordnet and uthm onto

From the above listing, similarity assessment is calculated using the following formula for Parts (S_p), Functions (S_f) and Attributes (S_a):

$$S_{p/f/a}\,(a,b) = \frac{|A \cap B|}{|A \cap B| + \alpha\,(a,b)\,|A\,/\,B| + (1 - \alpha\,(a,b))\,|B\,/\,A|} \quad (3)$$

$$For\ 0 <= \alpha <= 1$$

Where S corresponds to a Similarity function, a and b are the entity classes, A and B are the description sets of a and b, | | is the cardinality of a set, α is the depth of entity classes, $A \cap B$ is the intersections of A and B, A / B and B / A is the difference between A and B.

Assuming X is 4, Y is 5 and Z is 2, calculation is implemented in (4). The examples only have Parts in Features so the values are shown as below.

$$S(\text{java}^w, \text{java}^u) = \frac{4}{4 + \alpha\,(a,b) * 5 + (1 - \alpha\,(a,b)) * 2} \qquad (4)$$

Otherwise, total number of Parts (S_p), Functions (S_f) and Attributes (S_a) are aggregated accordance to the indicated formula:

$$S_u = w_p . S_p\,(a^p, b^q) + w_f . S_f\,(a^p, b^q) + w_a . S_a\,(a^p, b^q) \qquad (5)$$

$$\text{For } w_p, w_f \text{ and } w_a \geq 0 \text{ and } \quad w_p + w_f + w_a = 1.0$$

In WordNet, 'Java' has three different meaning called synset(/sense). They are referring to An island in Indonesia, Beverage and Programming language. The assessment mentioned above, give different values for all synset. This value is used to determine ranking. Most probable category for users based on his profile is listed in the highest rank.

This technique generates results in Table 5.

Table 5. Web category ranking

ID	Name	Faculty	Field	Web Category	Rank
718	Yusliza	IT	Computer	Beverage	2
				Island	3
				Programming language	1

CONCLUSIONS

The propose framework hybrid Semantic Web Search has been discussed. They are three criteria namely ontology, categorization and user profiling have been used in this research. Enhancement using user profiling criteria is embedded to the current practice which only uses ontology and categorization. It will give more releveant search results by considering users' own data in RDB. In this research, Similarity Measurement is utilizes to fulfilled user profiling part. Two independent ontologies are used throughout the measurement. To produce better illustration, example is enclosed in this paper with detail explanation in the case study section. This hybrid Semantic Web Search Engine implementation is capable to give the desired result in terms of user's profile.

ACKNOWLEDGMENTS

REFERENCES

[1] Mikroyannidis, A. Toward a Social Semantic Web. Computer, 40(11), 113-115, 2007.

[2] Renaud, D. SIREn: entity retrieval system for the web of data. Proceedings of the Third BCS-IRSG conference on Future Directions in Information Access. Padua, Italy, British Computer Society, 2009.

[3] Gruber, T. R. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2):199-220, 1993.

[4] d'Aquin, M., & Noy, N. F. Where to publish and find ontologies? A survey of ontology libraries. Web Semantics: Science, Services and Agents on the World Wide Web, 11(0), 96-111, 2011.

[5] Trillo, R., Po, L., Ilarri, S., Bergamaschi, S. and Mena, E. Using semantic techniques to access web data. Information Systems, 36 (2), 117-133, 2011.

[6] Jie, T., Limin, Y., Duo, Z., & Jing, Z. A Combination Approach to Web User Profiling. ACM Trans. Knowl. Discov. Data, 5(1), 1-44, 2010.

[7] Antoniou, D., Paschou, M., Sourla, E. and Tsakalidis, A. A Semantic Web Personalizing Technique: The Case of Bursts in Web Visits. Proceedings of the Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on. 530-535, 2010.

[8] Yoo, D., Hybrid query processing for personalized information retrieval on the Semantic Web, 2011.

[9] Abdullah, N. and R. Ibrahim (2012). Knowledge retrieval using hybrid semantic web search. Computer & Information Science (ICCIS), 2012 International Conference on Computer & Information Sciences.

[10] Jiang, H. Information retrieval and the semantic web. Proceedings of the Chongqing, China. IEEE Computer Society, V3461-V3463, 2010.

[11] Edwards, C. Analysis: Semantic web's hidden meanings. Engineering and Technology, 5 (16), 52-53, 2010.

[12] Joo, J. Adoption of Semantic Web from the perspective of technology innovation: A grounded theory approach. International Journal of Human Computer Studies, 69 (3), 139-154, 2011.

[13] Janev, V. and Vrane, S. Applicability assessment of Semantic Web technologies. Information Processing and Management, 47 (4), 507-517, 2010.

[14] Abdullah, N. and R. Ibrahim (2013). Knowledge retrieval in lexical ontology-based semantic web search engine. Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication. Kota Kinabalu, Malaysia, ACM.

[15] Kassim, J. M. and M. Rahmany. Introduction to Semantic Search Engine. Electrical Engineering and Informatics, 2009. ICEEI '09. International Conference on, IEEE, 2009.

[16] Aguilar-Lopez, D., I. Lopez-Arevalo, et al. Toward the semantic search by using ontologies. Electrical Engineering, Computing Science and Automatic Control, 2008. CCE 2008. 5th International Conference on, 2008.

[17] Cardoso, J. The semantic web vision: Where are we? IEEE Intelligent Systems, 22 (5), 84-88, 2007.

[18] Hepp, M. Ontologies: State of the Art, Business Potential, and Grand Challenges. In M. Hepp, P. D. Leenheer, A. d. Moor & Y. Sure (Eds.), Ontology Management. Semantic Web, Semantic Web Services, and Business Applications. New York: Springer, 2008.

[19] Diez-Rodriguez, H., Morales-Luna, G., & Olmedo-Aguirre, J. O. Ontology-based Knowledge Retrieval. Paper presented at the 2008 Seventh Mexican International Conference on Artificial Intelligence, 2008.

[20] Hao, Y. and Zhang, Y.-f. Research on Knowledge Retrieval by Leveraging Data Mining Techniques. Proceedings of the 2010 International Conference on Future Information Technology and Management Engineering. IEEE, 479- 484, 2010.

[21] Maier, R., Hadrich, T., & Peinl, R. Enterprise Knowledge Infrastructures (2nd Edition ed.). Berlin: Springer, 2009.

[22] Wecel, K. Towards an Ontological Representation of Knowledge on The Web. in W. Abramowicz (Eds.). Knowledge-based Information Retrieval and Filtering From the Web. USA. Kluwer Academic Publisher, 2003.

[23] D'Amato, C., Esposito, F., Fanizzi, N., Fazzinga, B., Gottlob, G. and Lukasiewicz, T. Inductive reasoning and semantic web search. Proceedings of the SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing. Sierre, Switzerland. Association for Computing Machinery, 1446-1447, 2010.

[24] D'Aquin, M., Motta, E., Sabou, M., Angeletou, S., Gridinoc, L., Lopez, V. and Guidi, D. Toward a new generation of semantic web applications. IEEE Intelligent Systems, 23 (3), 20-28, 2008.

[25] Morato J., Marzal M. A., Llorens J., Moreiro J. Wordnet Applications. Proc. Of the 2nd Int. Conf. Global Wordnet, 2004.

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

[27] Tversky, A. Features of similarity. Psychological review 84(4): 327-352, 1977.