Semantic Web based Information Query System for the Integration of Semantic Data

Okkyung Choi, Sangyong Han and Ajith Abraham
Department of Computer Science & Engineering, Chung Ang Univ. Seoul,Korea
okchoi@ec.cse.cau.ac.kr, hansy@cau.ac.kr, ajith.abraham@ieee.org

Abstract

In this paper, we propose Semantic Web based Information Query System (SW-IQS) using Ontology Server. SW-IQS is supposed to enhance efficiency and accuracy of information retrieval for unstructured and semi-structured documents. For the interoperability and easy integration, we suggest RDF based repository system. A new ranking algorithm is also proposed to measure the similarity between documents with semantic information for rapid and correct information retrieval. The proposed algorithm is implemented and compared with some of the existing models.

1. Introduction

The Internet was able to make great advancements rapidly because of its convenience and the fact that it is easily accessible to anyone around the world. However there emerged the problem of having too many search results on a single search for specific information the user requested.[2] The search engine research is a hot topic and new ideas are being tried out in order to solve such problems. But as long as the search method is based on plain data processing it would be difficult for a user to find the specific material he/she needs on the web because the search is executed fragmentarily based only on words and sentence construction having the semantic contents of the web document left out [2].

The semantic web is a new approach experimenting semantic search, automation, integration and reuse. This paper suggests the SW-IQS (Semantic Web Based Information Query System), which is a combination of agents and the automatic classification techniques upon a sub-structure of semantic web based techniques. With the suggested system, the lack of precision and recalls in the current search system can be improved. The proposed system is based on the RDF (Resource Description Framework) and the Ontology. The efficiency and accuracy of the proposed system is verified through a new method of similarity measurement using semantic metadata.

The paper is organized as follows. In Section 2, there will be an overview on the semantic web and an observation of the cosine similarity technique of the vector space model, which is one of the most widely used similarity measurement models. In Section 3, the architecture of the SW-IQS will be explained along with designing method, functions and characteristics of its modules. Next in Section 4 a new ranking algorithm technique will be introduced and applied in evaluating, comparing and analyzing the current search model and the newly proposed search model. Conclusions and future studies will be suggested in the final section.

2 Related Studies

2.1 The Semantic Web

Semantic refers to the assignment of meaning to a document. The semantic web is a highly advanced automation and intelligent technology, which allows the machine to understand the information like man. To have a machine understand and process information, the information must be disintegrated into the original data form it had been before being processed and then processed again into another appropriate form. Concerning this, the semantic web expresses the data in a way that can be interpreted by the computer and defines the relation between the different data so that the data that are automated and integrated in the application areas, such as electronic commerce, can be shared.

2.2 RQL (RDF Query Language)

RQL[14] was first introduced and proposed as a query language for RDF and RDF schema. RQL was developed by the European IST project C-Web and was included in the MESMUSES project by FORTH Institute of Computer Science based in Greece. RQL is

a function language adopting the syntax of OQL. The results for the RDF schema query are presented in the appropriate RDF code.

RQL is defined as a set of core queries or a set of basic filters. RQL is capable of organizing new queries through functional organization and repetition.

RDF uses a "select-from-where" expression, similar to that of SQL, to reorganize or filter information. The RQL filter applies the path expression in order to search for graphs from an arbitrary depth.

The query shown in Figure 1. is an RDF query for finding the property value of the name that matches with "*name" in the class. In RQL, the class variable is defined as $ and the property variable as @ for discrimination.

```
select Y, $Y
from {X} @P {Y : $ Y}
where @P like "*name"
```

Figure 1. Example of an RQL Query

2.3 The Vector Space Model and the Cosine Similarity

The Vector Space Model (VSM) is a way of representing documents through the words that they contain. It is used for information retrieval because it provides a tool that is capable of partial matching. This term weight is used in calculating the similarity between the user's query and each of the documents stored in the system. The vector space model searches for documents that partially match to the query terms and lists the searched documents starting from the one with the highest similarity to the one with the lowest. As with the vector space model, a document or a query is regarded as a certain point on the vector space. The vector space is determined by the index that appears in the document collection. Since a query is also regarded as a certain point on the vector space as same as a document, the vector space model does not use the Boolean algebra. Instead, the similarity between two documents is perceived by the cosine similarity, which is the cosine value of the angle formed by the two vectors. The smaller the angle, the greater the similarity. This is the main idea of the vector space model. [15]

In the vector space model, $W_{i,j,}$, the weight of the term and document pair (k_i, d_j), has a positive non-binary value. The query index also has a weight. If the weight of [k_i, q] is w_i,q>=0, the query vector is defined as ($w_{1,q}$, $w_{2,q}$, ..., $w_{i,q}$). Here, t indicates the total number of indexes in the system. The document vector is expressed as ($w_{1,j}$, $w_{2,j}$, ..., $w_{i,j}$).

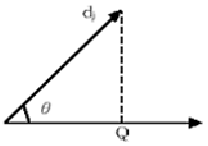

Figure 2. Cosine value applied as sim(dj, q)

As so, document d_j and the user's query q is expressed on a t dimensional vector as shown in (Figure 4). In the vector space model the similarity between document d_j and query q can be measured by the correlation between the two vectors. As an example, the similarity can be fixed to a certain value by using the cosine value of the angle formed by the two vectors as shown below

$$\mathrm{CosSim}(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{\left|\vec{d}_j\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \cdot \sum_{i=1}^{t} w_{iq}^2}$$

Figure 3. Cosine Similarity

3. The Design of SW-IQS

3.1 Techniques of System Design

In this section, we propose an overview of the SW-IQS process as illustrated in Figure 4.

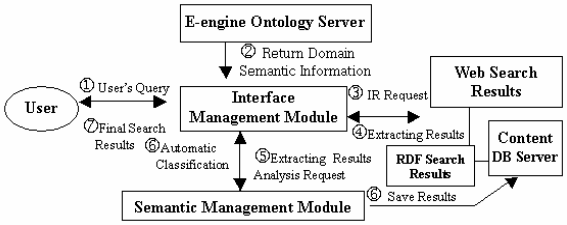

Figure 4. An overview of the SW-IQS search process

As depicted, when the user inputs the desired search information, the Interface Management Module searches for related pages on the web using the Domain Semantics information returned by the E-engine Ontology Server. Then, the Semantic Management Module automatically classifies and ranks the searched pages and the Content DB Server's RDF documents to provide the final search results to the user.

2.2 The System Architecture and the Function of Each Module

As shown in Figure 5., the architecture of the SW-IQS is composed of Interface Management Module, E-

engine Ontology Server, Content DB Server and Semantic Management Module. This section describes the functions and characteristics of each of the modules in detail.

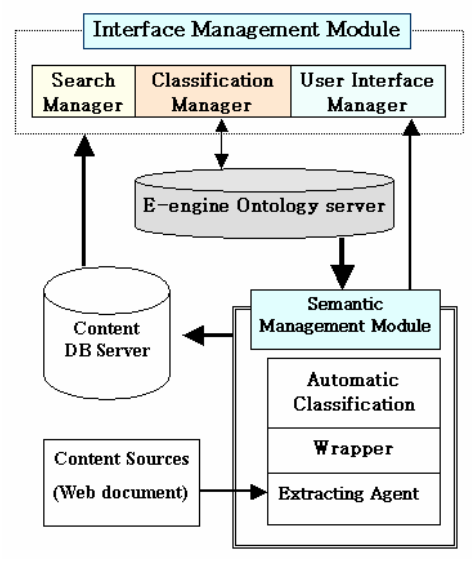

Figure 5. System Architecture

2.2.1 E-engine Ontology Server[2]

World Map (E-engine Ontology Server) is placed above the syntactic layer (XML) and semantic layer (RDF). It is a systematic method of expression that can improve the present condition where information is processed simply as data and the semantic context must be provided by man, and allow information to have value as knowledge.

E-engine Ontology Server is composed of three layers, the Content Manager, Schema Manager and the thesaurus manager. The Content Manager applies the definition of the semantic metadata and the definition of the classification model for semantic data search along with succession and equivalence to define the relation between different metadata. The Schema Manager defines the standard data type and format of the Content Manager's standard classification model and the thesaurus manager's semantic integration model. The Thesaurus Manager is like an encyclopedia. It defines the identification and property standards in accordance with the international standards for electronic commerce. The Thesaurus Manager integrates schema and unifies and reorganizes similar terms. In other words, it is in charge of integrating terms that are semantically the same.

2.2.2 Interface Management Module

(1) Search Manager

The Search Manager induces accurate search results by bringing the related domain semantic information from the ontology server based on the query input by the user and requesting the user for a second query. In other words, when a multiple number of domain semantic information has been found the manager suggests the subject words and descriptions of the semantic information for the user to select.

(2) Classification Manager

The Classification Manager's method of searching by subject is quite distinctive compared to the current layered structure method. Searching by subject determines the relations between terms based on the RDF documents and enables more precise and efficient search for documents by applying a flexible network structure.

(3) User Interface Manger

The User Interface Manager provides various users' search input screens, ontology information selecting screens and final information search results screens.

2.2.3 Semantic Management Module

As with the Semantic Management Module, the information-extracting agent is used for extracting related web pages, and then the wrapper is used to return XML documents based on the material. After this, the Automatic Classification Module is used to automatically classify the pages and then the results are stored in the Content DB Server. Here, the similarity of the web pages must be measured in order to automatically classify and rank them. The term *relationship variable* is used to measure the synonymous relations between terms (*i*) and the *Semantic distance variable* to measure the relationship between terms, that is, measuring the distance between the two terms to see how close they are.

The *relationship variable* uses the similarity level (the range of similarity) of each term to measure the similarity. Within a range of 1 to 9, those scoring close to 1 have high similarity and those scoring close to 9 have low similarity. The factors that discriminate the levels are the search words, terms contained in the extracted documents, synonyms of the search words extracted through ontology and the synonyms of terms contained in the documents extracted through ontology.

The definition of *relationship* is as follows.

Definition 1: term *relationship* (synonymous relations between each term)

$$R_j = \frac{f_{ij}}{t_r}$$

f_{ij} : The number of occurrence of term*(j)* in document*(i)*

$t_{r;}$: Measurement variable for similarity between terms*(j)* = *level*(t_r)

The *Semantic Distance Variable* uses the proximity between each horizontal node (***H**p*) and the proximity between each vertical node (***V**p*) of each of the documents' structures. Figures 6 and 7 illustrate a comparison of *Semantic Distance* measured for the XML documents and RDF documents respectively.

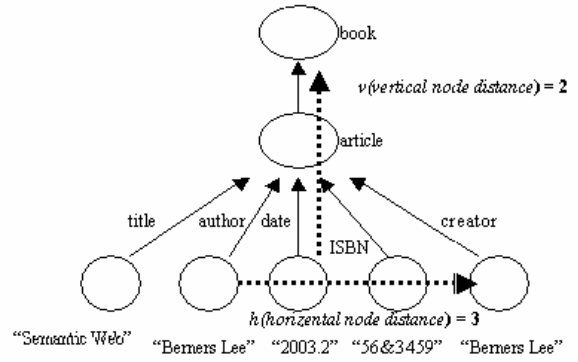

Figure 6. Semantic distance using XML document

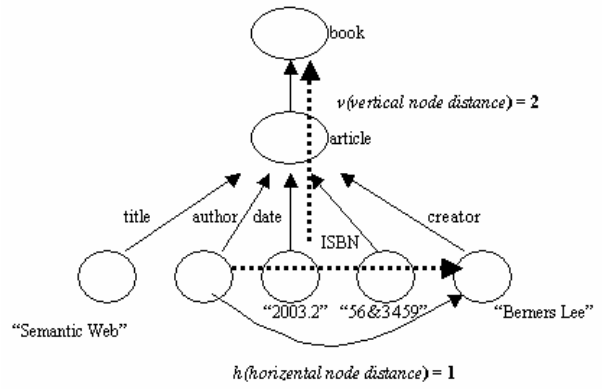

Figure 7. Semantic distance using RDF document

The reason why the *Semantic Distance Variable* differs between the XML and RDF documents is, because the XML document is based on a tree structured hierarchical method while the RDF document is based on the graph structured method. When the user is looking for "a book whose author and publisher are both Berners Lee," the horizontal node distance between "author" and "creator" in the case of RDF documents is "1," which means they are very closely related. Meanwhile, the same distance is "3" in the case of XML documents, which shows that "author" and "creator" are less related than in the former case. Different modeling methods are required for accessing and handling XML and RDF documents, which are formed in different structures. The XML document having a tree based hierarchical structure; it requires DOM (Document Object Model) to handle XML data. However, for the RDF document, a new modeling method is required that is capable of intensive analysis and expression of the attributes or relations between attributes and the relation between classes. The definition of *Semantic Distance* is as follows.

Definition 2: *Semantic Distance* (The relationship between terms)

$$D_j = H_p * V_p$$

$$H_p = \frac{1}{C^h}$$

$$h = |k - j|$$

$$C = \frac{level(i_j)}{max\ V(i)}$$

H_p : The horizontal proximity between each node

h : Horizontal proximity between each term

C : The measurement variable for the level of each tree

$level(i_j)$: The level value of the term *(j)*'s location in document *(i)*

$max\ V(i)$: The highest-level value in document *(i)*

$$V_p = \frac{1}{F^v}$$

V_p : The vertical proximity between each node

F : The vertical proximity determining factor (0<F<1)

V : $level(i_k) - level(i_j)$: The vertical node distance between each term

And finally, the proportionally applied value of weight (***k**j*) for automatic classification and ranking is as follows.

Definition 3: Proportionally applied value of weight (***k**j*)

$$k_j = \frac{R_j}{D_j}$$

***R**j* : term relationship variable measuring the synonymous relation between terms *(i)*

***D**j* : Semantic Distance variable measuring relationship between terms.

In the next section, a new ranking algorithm using the proportionally applied value of weight (***k**j*) is given.

3 Evaluation of Performance

This Section analyzes some of the current issues of the search models suggested in former studies [2] and suggest a new ranking algorithm for improvement. Furthermore, the new algorithm is illustrated for evaluating the performance of SW-IQS.

3.1 The Ranking Algorithm

Previous studies [2] exclude semi-structural and non-structural documents, which possibly may contain meaningful information, and only refers to whether semantic information (RDF documents) are included or not to determine the similarity and ranking. This method rather lowers the preciseness of the ranking. Our research attempts to rank documents by using

$$sim(d_j q) = k_i \times \frac{\vec{d}_j . \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}\right|} \qquad (1)$$

which applies the proportionally applied value of weight (k_j) and the cosine similarity of the vector model, and does not classify and extract the documents.

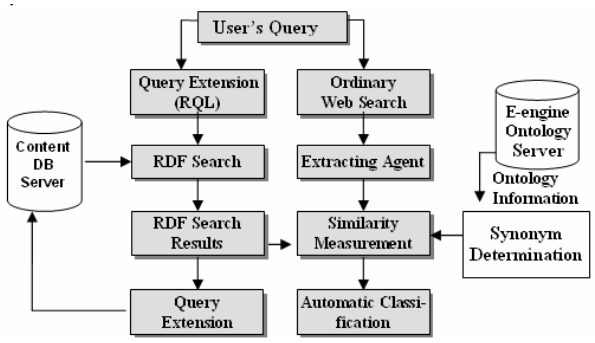

Figure 8. Step-by-step searching method using ranking algorithm

3.2 Comparison and Analysis

For performance evaluation, experiments were carried using XML and RDF documents from the pages found with the search word "the book which the author is burners lee." The resulting documents of the web page search are given in [3] and the numbers attached to each document indicate the document number. The performance evaluation is done according to a new step-by-step searching method as illustrated in Figure 8. and the detailed procedures are given below.

Step 1: The top ten documents are searched through the ordinary search engine [Google]. The unnecessary web pages are extracted from the searched pages by the extraction agent. Here, documents number 3, 9 and 11 are removed from the web page list because there are broken links.

Step 2: The RDF document is searched by using RQL. The search results suggested person_book.rdf document [4].

Step 3: The vector based cosine similarity is calculated for the seven web documents and two XML and RDF documents, which were obtained through the extraction agent.

Step 4: The HTML documents are converted to XML documents by using the HTMLtoXML Wrapper in order to measure the relationship between terms. Then using the new ranking algorithm, the similarity cosine is measured for the XML documents and the RDF document. As observed in Table 1, the ranking was different from the ranking results using the cosine similarity. Based on the difference, the ranking for each document is readjusted and the documents are automatically classified.

The currently used search engine does not take the weight of words and synonyms and the relationships into account at all. Furthermore, as observed in Table 1., the RDF document was ranked at 4, which is relatively lower than the rankings of ordinary web documents, when the vector based cosine similarity was used. This is because in the case when the vector model's cosine similarity is used, no application are made of the ***term relationship variable***, which measures the synonymous closeness between terms (***i***), and the ***Semantic Distance variable***, which measures the relationship between terms, that is the proximity measured by the distance between two terms. Meanwhile, when the new ranking algorithm, an improvement of the current vector model, is applied for measuring similarity, the RDF document that had been ranked at 4 rose to rank 1 and document number 0, which had been ranked at 1 with the vector model, was ranked lower at 3. The new ranking algorithm performed higher preciseness and recalls, which can be relied more by the user.

4. Conclusion and Further Studies

This paper proposed SW-IQS composed of Semantic Management Module, Content DB Server, E-engine Ontology Server and Interface Management Module. As a solution to the current search engine model, an integrated information searching system is suggested to enhance efficiency and preciseness of searching by information extraction and automatic classification, and to maximize the processing of both semi-structured and unstructured documents.

Table 1. Similarity measurement results

Document No.	Term weight				Cosine Similarity	**Rank**	kj	Similarity	**Rank**
	book	author	berners Lee	**book author berners Lee**					
0	0	0	0	0	0	8	0	0	**8**
1	0	0.0011745	0.0003355	0.00151011	0.00075505	4	0.002064926	0.001409991	**1**
3	0.00695933	0.000466	0.0023297	0.00975511	0.00487755	0	0.001553185	0.00321537	**0**
5	0.00067344	0.000505	0.0001893	0.001367778	0.00068388	5	0.000399741	0.000541815	**6**
6	0.00153377	0.00089855	0.0008985	0.003330889	0.00166544	1	0.000994426	0.001329935	**3**
7	0.00054144	0.00008	0.0001522	0.000769778	0.00038488	6	0.001152593	0.000768741	**5**
8	0.00165611	0.00058988	0.00052777	0.002773778	0.00138688	3	0.000910704	0.001148796	**4**
9	0.000043	0.00004	0.00004	0.000122	0.000061	7	0.000053	0.000057	**7**
11	0.00037555	0.00223922	0.0000689	0.003295778	0.00149486	2	0.001205704	0.001350284	**2**

The searching system using the integrated information uses semantic web factors, which are emerging as the next generation web. With the proposed system, semantic data search and integration are possible through establishment of ontology, data standardization, data integration and semantic connection methods. Through the establishment of ontology, data standardization, data integration and the semantic connection method, the semantic web search model is capable of semantic data search and integration.

In order to evaluate the performance of the proposed integrated searching system, some of the previous methods [2] were improved and a new performance-evaluating algorithm has been suggested to verify the efficiency and preciseness of the system. In other words, a new semantic vector model capable of cosine similarity measurement of non-binary weight as an improvement of the formerly proposed searching method using RDF semantic meta-information. The suggested system was tested and analyzed by applying the new ranking algorithm. Results showed that the ranking of the documents extracted from the web and documents stored in the Content DB Server is more efficient and precise compared to the former method.

References

[1] The RDF Query Language (RQL), http://139.91.183.30:9090/RDF/RQL/.

[2] Okkyung Choi, Seokhyun Yoon, Myeongeun Oh. Sanygoung Han, "Semantic web Search Model for information retrieval of the semantic data", The Second HSI Conference, June. 2003. 6 , pp. 588-593

[3] Web test data, http://ec.cse.cau.ac.kr/okchoi/test_webdata.html

[4] RDF document, http://ec.cse.cau.ac.kr/okchoi/person_book.rdf

[5] Tim Berners-Lee, work in progress, October 1998, Why RDF model is different from the XML model, http://www.ilrt.bris.ac.uk/discovery/rdf/resources/#sec-examples

[6] Gomez-Perez, A.; Corcho, O. "Ontology languages for the semantic web**",** IEEE Intelligent Systems , Volume: 17 Issue: 1 , Jan/Feb. 2002 p.54 –60

[7] Hendler, J. "Agents and the Semantic Web**",** IEEE Intelligent, Volume: 16 Issue: 2 , March-April 2001 p.30 –37

[8] Mihalcea, R.F.; Mihalcea, S.I., "Word semantics for information retrieval: moving one step closer to the Semantic Web**"**, Tools with Artificial Intelligence, Proceedings of the 13th International Conference on , 2001, p.280 –287

[9] Sheth, A.; Bertram, C.; Avant, D.; Hammond, B.; Kochut, K.; Warke, Y., "Managing semantic content for the Web", IEEE Internet Computing , Volume: 6 Issue: 4 , July-Aug. 2002, p. 80 –87

[10] Forum on the Standardization of Electronic Commerce, "Electronic Commerce Standard Operation System Applying the Semantic Web Technology", ECIF, 2001

[11] Lassila, O.; van Harmelen, F.; Horrocks, I.; Hendler, J.; McGuinness, D.L. , "The semantic Web and its languages", IEEE Intelligent Systems , Volume: 15 Issue: 6 , Nov.-Dec. 2000 Page(s): 67 –73

[12] Decker, S.; Mitra, P.; Melnik, S., "Framework for the semantic Web:an RDF tutorial", IEEE Internet Computing , Volume: 4 Issue: 6 , Nov.-Dec. 2000 Page(s): 68 –73

[13] Decker, S.; Melnik, S.; van Harmelen, F.; Fensel, D.; Klein, M.; Broekstra, J.; Erdmann, M.; Horrocks, I., "The Semantic Web:the roles of XML and RDF", IEEE Internet Computing , Volume: 4 Issue: 5 , Sept.-Oct. 2000 Page(s): 63 –73

[14] http://139.91.183.30:9090/RDF/RQL/ , Forth Institution of Computer Science

[15] Kim Yeong-cheon et al, "Research on Enhancing Information Search by Regiving Term Weight", Korea Fuzzy Logic and Intelligence Systems Society, Vol. 11 No. 9, 2001, pp. 811-816