Journal of Theoretical and Applied Information Technology
10th December 2013. Vol. 58 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RSBR: A PARADIGM FOR PROFICIENT INFORMATION RETRIEVAL USING QSPA AND RELEVANCE SCORE BASED RANKING

[1]SRIDHARAN. K, [2]M. CHITRA

[1]Department of Computer Science and Engineering,
Anna University, Chennai, India
[2]Department of Information Technology,
Sona College of Technology, Salem, India.

ABSTRACT

Due to the speedy growth of content volume over the internet, the required content that is relevant to the user's query is retrieved with difficulty by the common search engines. To overcome this limitation, semantic web search approaches are utilized. Many researches in semantic web depend on data search centered meaning. The general purpose of these researches is to enhance the current data search and retrieval techniques. An effective Relevancy-based Semantic Search Engine (RSSE) prototype that allows the users to determine relevant resources and services by semantics is proposed. The proposed approach uses Query Similarity Prediction Algorithm (QSPA) for efficient information retrieval with minimum processing time. The technique serves multiple remote users. The relevancy based ranking of documents depending on the occurrence of semantic terms is performed in QSPA. The experimental results show that the approach is efficient when analyzed with parameters like precision, recall, F-measure, and the time needed to obtain query results.

Keywords: - *Information Retrieval (IR), Service Level Agreement (SLA), Semantic web, Query Similarity Prediction Algorithm (QSPA), Ranking, Cache server.*

1. INTRODUCTION

Due to the constant and speedy growth of the data stored and that are shared on the web and other document repositories, different information retrieval approaches are widely adopted. This enlargement leads to difficulties like determination of correct results and to maintain all existing data contents in an efficient manner. Search engines are used to effectively maintain the Information Retrieval process. The Information Retrieval System, a main component of search engine performs important tasks like web pages collection and retrieval of suitable text documents answering a user query. The users need to be capable to obtain appropriate data to satisfy their appropriate information requirements. Figure 1 depicts the generic architecture of a search engine.

To extract relevant documents from the document corpus, the web crawler is included. The traditional web search engines are utilized for searching and obtaining the results. But the main drawback is that the keywords are used for retrieving document. The semantic similarity of the query is used to provide required information to the user query.

The semantic web, extension of the current web, provides efficient reuse, atomization and interoperability. Generally, the semantic web is known as the web of relations between resources that denotes real world objects. The retrieval on the semantic web improves the information search and retrieval results in two ways [21].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

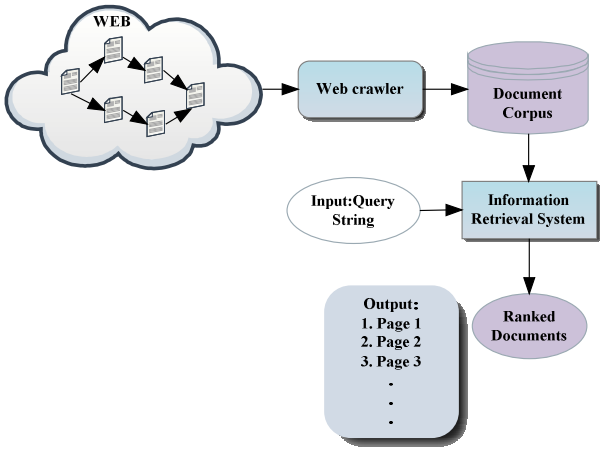

Figure 1: Generic Architecture of Search Engine

- Provides an easy approach to support the semantic search module for better understanding the query denotation
- Enhances the relevancy rate of the search results.

The semantic web is also used in developing a web of semantic documents. Handling semantic markup is an important task to be achieved during the Information Retrieval [22]. It is well known that a sort of semantic web query is given as input to the retrieval system. So there is need of semantic markup encoding. There are two types of searches on the internet [23]:

Navigation searches: In this search, the submission of query on the search engine is performed by the user to determine the documents. The search engine is utilized as a navigation tool in order to navigate to a specific required document.

Research searches: This type of search includes denoting an object about which the user is trying to obtain data. In the semantic web, each page contains semantic metadata that performs the recording of additional data related to the web page itself. Further, the semantic web has benefits like obtaining the results in fast and effective manner, clustering the results of different search engines, etc. However, the semantic web is basically identical to the web of HTML documents.

An efficient caching mechanism is used for handling exponential progress and the changing environment of World Wide Web. This approach is also used to provide fast searching mechanism. Each and every web browser contains a built in local cache to store the user needed objects. So, if some other user browses the page in need of the same data, that information will be loaded quickly. Caching is the automatic impermanent copies of data stored on the host server for the easy availability of information.

The main contribution of this paper is to design a semantic based search engine that includes Query Similarity Prediction Algorithm (QSPA) to match the given query with the information stored on the cache. The method uses Service Level Agreement (SLA) for effectively caching the search data and user activities. Relevance Score Based Ranking (RSBR) is used for sorting the results depending on its relevancy rate regarding the user query. The main objective of this work is to obtain the search results by determining the context and semantics of the query with minimum time consumption and enhanced precision.

The remainder of this paper is framed as follows: Section 2 describes the related works, Section 3 describes the proposed method for RSSE. Section 4 provides the experimental results of the proposed approach. Finally, Section 5 describes conclusion and future work.

2. RELATED WORKS

A forwarded step for effective semantic we formation is proposed in the paper. SEAL (SEmantic PortAL) approach [1] is used for acquiring information at a portal along with its construction and conservation. The SEAL architecture consists of Ontobroker system and knowledge warehouse. The method includes software agents, community and general users. The personalization and semantic ranking is performed for achieving correct results. A wide difference between the conventional information retrieval and semantic web [2] was analyzed. The pillars of the semantic web are considered as markup languages, ontology and intelligent agents depending on the explanation of this paper. A unifying technique for semantic web knowledge with web usage mining was described [3]. The process achieves web personalization which defines the tracking of web experience to a specific user or a set of users. [4] discusses different semantic similarity algorithms

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

like MeSH and WordNet and their issues. The approach also includes query and term expansion for semantic similarity aggregation.

The building mechanism of web ontology based editor and browser is described [5]. The paper demonstrated about the annotation process, hypertextual navigation, views and display of a site for performing correct search. The Semantic Term Matching (STM) technique analyzes the semantics of queries and documents. The approach can be extended using some lexical resources like WordNet. Analysis of ontology and semantic web is provided [6]. The paper described the importance of ontology in the semantic web development. It focuses on the ontology-based data visualization. Different techniques for information retrieval and web search are described [7]. Various search engine development components like page repository, query module, web crawlers, indexing module, ranking module and pertinent pages are described.

The distributed data retrieval in semantic web [8] includes

- Selection of resource
- Reformulation of query
- Fusion of data and rank accumulation

Development of OntoLook system in relation-based semantic search [9] incorporates a key algorithm in order to generate concept-relation graph for the given query and stored documents. The priority based page ranking could be included as future work. The path from the conventional World Wide Web to the semantic web was described in [10]. The above two scenarios were distinguished and analyzed in the aspects like pattern of collaboration, message exchange, etc. The information about conventional Chinese medicine is extracted by employing knowledge discovery and information retrieval on semantic web [11]. The information retrieval process [12] supports multiple remote users in distributed computing environment. The query handling architecture was included in the approach.

A comparative study between Yahoo and Google [13] with respect to relative recall and precision of the search engines considers the comparison of retrieval efficiency of both search engines. The comparison is performed based on the simple one-word queries, simple multi-word queries and complex one-word queries. The results showed that the recall rate and precision of Google is relatively higher than Yahoo.

Personalized Semantic Search Engine (PSSE) [14] utilizes multi-crawlers for gathering information from web resources. PSSE involves three stages: Processing, searching and ranking stage. The user satisfaction is improved by including an efficient technique that minimizes the retrieval time. SPIRS [15] depends on agents and semantic web to support expressive queries. The approach also included a user model in order to improve the relevant documents ranking. Semantic information retrieval for obtaining relevant data from the web documents used crawler depending on domain ontology. The semantic information retrieval enhances the retrieval process than the traditional methods [16]. A Semantic Information Extraction in University domain (SIEU) [17] includes construction of ontology, refined query formation and ranking of obtained links.

Another work described four perspectives of users and designers [18] such as high recall, static knowledge structure, lack of experimental tests and low precision. A Smart Web Query Method (SWQ) is used for performing semantic web search [19]. The technique is used to formulate the correct query by using domain similarities depending on context ontologies. The semantic search filtering process in incorporated to perform relevance ranking of web pages. The Information Retrieval [20] described the analysis of lowest-level word semantics. The Word Semantic Model (WS) is used for keyword based matching. The part of speech, word stems, searching and semantic indexing are considered. A forwarded step for effective semantic we formation is proposed in the paper.

3. PROPOSED WORK

A novel method is proposed here for searching is almost done on the basis of word occurrences in the document. Typical search engines improve this in the context of the web with information about the hyperlink structure of the web. Further, the accessibility of large volume of structured data about a extensive range of objects on the semantic web provides some criteria for improving the traditional search models. Figure 2 shows the general proposed flow of information retrieval from web server. The proposed approach depends on content based information retrieval with SLA accountability. This enhances the accuracy of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

information retrieval with minimum time consumption than utilizing other traditional search models.

3.1 Relevancy-based Semantic Search Engine (RSSE)

3.1.1 Enhancement of SLA based Accountability

The primary process for RSSE is to create accountability for users those are going to access the search engine to retrieve the necessary documents. For that concern, Service Level Agreement (SLA) has been framed for enabling the RSSE to users. The default gateway of the organization is used by RSSE for SLA constraints and user based levels. The constraints are provided for deferent levels of user such as normal user , knowledge user and experts with corresponding their domain of knowledge. The block diagram of the proposed RSSE is shown in figure 2. A default gateway is the node on the computer network which is used by the network software when an IP address does not match with any other routes in the routing table. It is known as the IP address of the router to which the PC network is connected. It is used for security purpose.

3.1.2 User registration

Each user in the organizational network will be provided with a unique ip for network access after registered with necessary details. The registered information is verified and validated based on the levels of users automatically. That unique ip is the unique login id for RSSE. This login id is used for both validation and agreement process. Based on the id and default gateway, history details are stored and aids for the proposed algorithm as it contains what type of query was processed for which ip from which network. The accessibility for the registered user is used to retrieve the required information with more precision.

The Search will be enabled only when the user creates appropriate accountability for their access. In addition, the SLA comprises the norms such as the openness of data that are acquired by a specific registered user and stored in cache can be accessible to other users who are the authorized members of proposed scheme for searching the contents they require.

3.2 ENABLING RSSE

My proposed search engine is activated automatically for registered accountability based on the service agreement to retrieve the available relevant data in various formats such as audio, video or text for a given user query. In additionally proposed Relevancy based Similarity Search Engine, process of data retrieval is accomplished in two ways.

- Cache server based retrieval
- Online based retrieval

Tracking process is performed in this step. The query submitted by the user is retrieved and checked whether there is existence of that query in the gateway. If the requested query already exists, the Query Similarity Prediction Algorithm (QSPA) is used for relevant information retrieval. By using cache server based retrieval both speeds up access to data retrieval, reduces demand on an enterprise's bandwidth and remote access performance is significantly improved compare with other similar models.

3.3 New QSPA Model

Semantic terms are discovered from the query and it is utilized to match with documents stored in the cache server for relevancy. When the semantic terms in the query matches with the cache server document with specified constraints, then it is known as relevant documents. These relevant documents are stored as match set.

The technique is applied for all relevant documents in the match set and semantic terms in the query. The match set consists of set of relevant documents for the user query.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

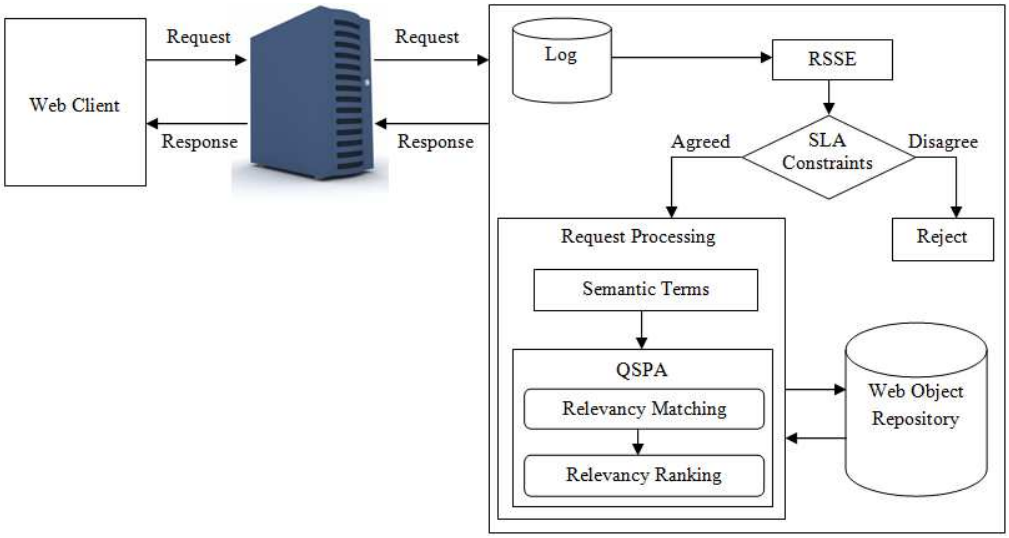

Figure 2: Proposed Diagram for IR from Web Server

3.3.1 Ranking

There are three types of document ranking

- Page ranking
- Structure ranking
- Content based ranking

In page ranking, the ranking of documents is performed based on the links in the page. In structure ranking, tag structurization is considered. The proposed approach uses content based ranking of documents. The relevant documents are ranked by the occurrences of semantic terms. The QSPA algorithm is shown below

```
bestrankmatch outputs (List I, List O, split-seq-Node N,Query Q)
if O is empty then
  return true
end if
o1  head(O)
for all k to N children do
  k.matchSet  =  k.matchSet {o1}
  if matchOutputs(I, k.matchSet, k) then
    if matchOutputs(I, tail (O), N) then
      return true
    end if
 end if
  k.matchSet =  k.matchSet {o1}
end for
for all k to N children do
   K.similarmeasure=k.similarmeasure(Q,K)
   Ranking;
end for
   k.matchset=toprankoutputs();
return false
```

List I defines the Semantic words from the user query, List O characterize the set of matching terms for the semantic words of the user query from the data dictionary. In addition, Split–seq-Node N – XML exemplifies the child nodes of the semantic description fields like RDF and OWL files.

If the option list is empty, it will not enter into the searching loop, which means there are no relevant documents for the given query.

Otherwise, each option in the option list is compared to measure the weight of all documents, for each child node in the ontological file. Moreover, weight of each option list is computed.

Document weight is computed based on the number of occurrence of each word in the option list in each document using the following formula.

$$R_c(i) = \sum_{s \in S} W\left(S(i)\right)$$

Where

$R_c(i)$ - Refers relevancy of document – (i)

$W\left(S(i)\right)$ - Refers to Weight of the term s from S

s – Each term in the option list

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

S – All the terms in the option list.

- Average document relevancy is computed using relevancy measure of all retrieved documents for document ranking.

$$Mean = \frac{\sum_{i=0}^{n} R_c(i)}{n}$$

Where

i - Refer document (i)

$R_c(i)$ - Relevancy measure of document (i)

n - Number of documents.

- Mean value is the threshold value of retrieved documents
- The documents which have relevancy measure less than threshold value is removed from the relevant documents.
- The relevant documents are matched with the user query (Q) to compute similarity measure of the document for document ranking.

$$S_m(i) = \sum_{i=0}^{n} M_i(Q)$$

Where

$S_m(i)$ - Similarity measure of document (i)

$M_i(Q)$ - Matching term (i) in the query (Q)

- Based on the similarity measure of each document in the relevant documents, the top scored 10 documents are selected and provided to the user.

3.4 Modified Online Based Retrieval

This modified version of online based retrieval activated automatically when the agreed users make new request if it is not get relevant search result from our corpus. The searching process will be done on the web server instead of cache server , If a new request of authorized user that does not have any match on the catch server corpus, the request will be thrown or forwarded for online processing i.e., the information is retrieved from yahoo or google . Semantic similarity between the given query and the documents on the web server is computed. The relevant documents are retrieved. Then, HTML parsing is performed.

3.5 Performance Evaluation

3.5.1 Precision

The Precision value can be determined using the equation 1.

$$Precision = \frac{P_{RES} \cap R_{RES}}{R_{RES}} \quad (1)$$

In equation (1), P_{RES}, and R_{RES} denotes the relevant result and the retrieved result derived from a single query.

3.5.2 Recall

Similarly the recall value is measured using the equation 2.

$$Recall = \frac{P_{RES} \cap R_{RES}}{P_{RES}} \quad (2)$$

Range of values for Precision, Recall and F-Measure is 0 to 1.

3.5.3 F - Measure

$$F = \frac{2.precision.recall}{(precision + recall)} \quad (3)$$

3.5.4 Time (in Milliseconds)

$$T(Q) = T_r - T_q \quad (4)$$

Where,

$T(Q)$ - Time taken to provide results for the given query

T_r - Results provision time

T_q - Query given time

4. EXPERIMENAL RESULTS

We gathered initial set of documents for our experimentation from the online search engine. We are considering the sample size of documents for results discussion here because other documents are less informative then the considered top ranked documents. The proposed method is evaluated using the various metrics, namely precision, recall, f-measure and time consumption. The proposed search engine is also compared with the existing search engines such as Yahoo, and Google. Google

Journal of Theoretical and Applied Information Technology
10th December 2013. Vol. 58 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

describes for a given query, set of documents are retrieved from Google and our QSPA algorithm is applied for those documents. Yahoo defines for a given query, set of documents are retrieved from yahoo and our QSPA algorithm is applied for those documents.

Precision is the measure used to determine the fraction of retrieved results that are pertinent to the input query. Figure 4 shows the precision efficiency comparison of Google, Yahoo and RSSE for various numbers of documents. The analysis shows that the proposed approach achieves high precision efficiency when compared with other search engines. Table 1 depicts the precision comparison of proposed and other search engines. The proposed approach works upto 50000 documents with 2GB RAM space. When the RAM space and the system configuration are enhanced, the proposed technique can be applied for high number of documents.

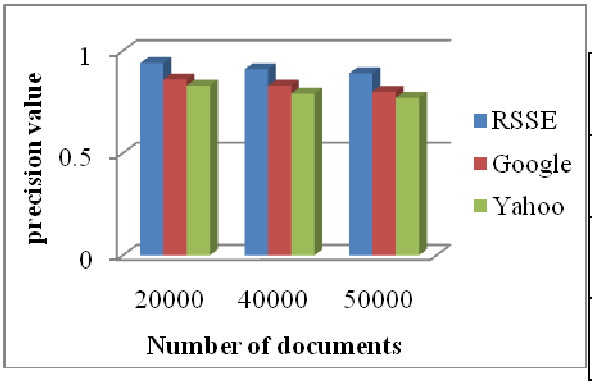

Figure 4: Precision Analysis Of RSSE, Google And Yahoo For Various Numbers Of Documents

Table 1: Precision analysis of RSSE, Google and Yahoo for various numbers of documents

Number of documents	RSSE	Google	Yahoo
20000	0.94	0.86	0.83
40000	0.91	0.83	0.79
50000	0.89	0.8	0.77

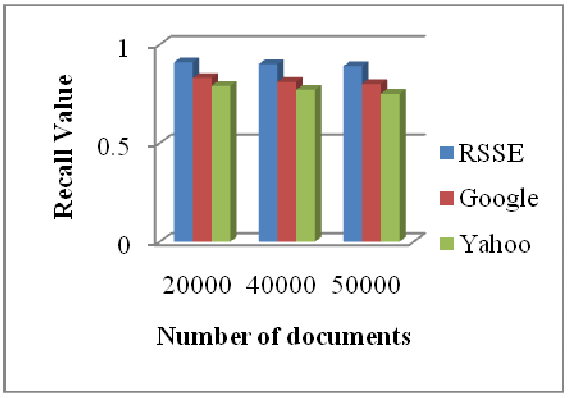

Figure 5: Recall Analysis

Figure 5 and table 2 depicts the recall analysis of Google, Yahoo and the proposed RSSE for different number of documents. The experimental results shows that the proposed approach outperform other methods.

Table 2: Recall Analysis Of RSSE, Google And Yahoo For Different Number Of Documents

Number of documents	RSSE	Google	Yahoo
20000	0.91	0.83	0.79
40000	0.9	0.81	0.77
50000	0.89	0.8	0.75

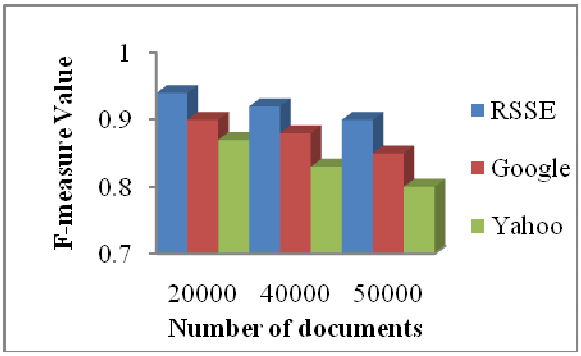

Figure 6: F-measure analysis

Figure 6 and table 3 represents the F-measure values for the three search engines namely Google, Yahoo, and RSSE, which expresses explicitly that the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

proposed technique retrieves more accurate results for a query than the other two techniques.

Table 3: F-measure analysis of RSSE, Google and Yahoo for different number of documents

Number of documents	RSSE	Google	Yahoo
20000	0.94	0.9	0.87
40000	0.92	0.88	0.83
50000	0.9	0.85	0.8

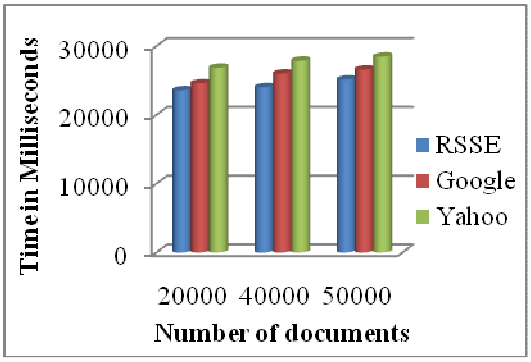

Figure 7: Time analysis

Time is another important factor, which decides the efficiency of the search engine. This paper also discusses the time factor of the three search engines. The analysis and its results for time factor are presented in the Figure 7. It explicitly expresses that the time taken from the proposed RSSE search engine is lesser than existing methods. This paper measures the time factor is measured in milliseconds. Table 4 shows the time analysis of the proposed technique and other search engines.

Table 4: Time analysis of RSSE, Google and Yahoo for various numbers of documents

Number of documents	**RSSE**	**Google**	**Yahoo**
20000	23568	24587	26850
40000	24015	25998	27854
50000	25145	26584	28457

5. CONCLUSION

This paper described a new search engine RSSE for obtaining most applicable documents for user queries. The approach allows the users to find the location of pertinent services and resources by using semantic expertise. The proposed technique uses Query Similarity Prediction Algorithm (QSPA) to serve multiple remote users in an efficient way. The method uses content based retrieval of information along with SLA accountability based on the different user levels. It achieves effective retrieval of information with less time consumption when compared with the conventional search models. The relevancy based ranking method is used to arrange the obtained results based users. The experimental results showed the adeptness of proposed method with respect to parameters like recall, precision, time required to obtain query results and F-measure. As a future work, the development of secure and improved trust based retrieval process for the different type of users based is considered.

REFRENCES:

[1] A. Maedche, S. Staab, N. Stojanovic, R. Studer, and Y. Sure, "Semantic portal-the seal approach," *Spinning the Semantic Web*, pp. 317-359, 2001.

[2] M. Benton, E. Kim, and B. Ngugi, "Bridging The Gap: From Traditional Information Retrieval To The Semantic Web," *AMCIS 2002 Proceedings*, p. 198, 2002.

[3] H. Dai and B. Mobasher, "Integrating semantic knowledge with web usage mining for personalization," *Web Mining: Applications and Techniques, Anthony Scime (ed.), IRM Press, Idea Group Publishing*, 2005.

[4] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Information retrieval by semantic similarity," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 2, pp. 55-73, 2006.

[5] A. Kalyanpur, B. Parsia, and J. Hendler, "A tool for working with web ontologies," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 1, pp. 36-49, 2005.

[6] L. Reeve, "Information retrieval on the semantic Web using ontology–based visualization," ed, 2006.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[7] Z. Markov and D. T. Larose, "Information Retrieval and Web Search," *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*, pp. 3-46, 2006.

[8] U. Straccia and R. Troncy, "Towards distributed information retrieval in the semantic web: Query reformulation using the oMAP framework," *The Semantic Web: Research and Applications*, pp. 378-392, 2006.

[9] Y. Li, Y. Wang, and X. Huang, "A relation-based search engine in semantic web," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, pp. 273-282, 2007.

[10] C. L. Chou, "From World Wide Web to Semantic Web," 2007.

[11] Z. Wu, T. Yu, H. Chen, X. Jiang, Y. Feng, Y. Mao, H. Wang, J. Tang, and C. Zhou, "Information retrieval and knowledge discovery on the semantic web of traditional chinese medicine," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 1085-1086.

[12] I. Roussaki, M. Strimpakou, C. Pils, N. Kalatzis, and N. Liampotis, "Distributed context management in support of multiple remote users," *Context-aware mobile and ubiquitous computing for enhanced usability. IGI Publishing Hershey, PA*, pp. 84-113, 2009.

[13] B. T. S. Kumar and J. Prakash, "Precision and relative recall of search engines: A comparative study of Google and Yahoo," *Singapore Journal of Library & Information Management*, vol. 38, pp. 124-137, 2009.

[14] A. M. Riad, H. K. El-Minir, M. A. ElSoud, and S. F. Sabbeh, "PSSE: An Architecture For A Personalized Semantic Search Engine," *International Journal on Advances in Information Sciences and Service Sciences*, vol. 2, pp. 102-112, 2010.

[15] K. M. Fouad, A. R. Khalifa, N. M. Nagdy, and H. M. Harb, "Web-based Semantic and Personalized Information Retrieval," *International Journal of Computer Science*, vol. 9, 2012, pp. 266-276.

[16] H. M. Harb, K. M. Fouad, and N. M. Nagdy, "Semantic Retrieval Approach for Web Documents," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 2, pp. 11-75, 2011.

[17] S. Rajasurya, T. Muralidharan, S. Devi, and S. Swamynathan, "Semantic Information Retrieval Using Ontology In University Domain," *arXiv preprint arXiv:1207.5745*, 2012.

[18] R. Khatri, K. S. Dhindsa, and V. Khatri, "Investigation and Analysis of New Approach of Intelligent Semantic Web Search Engines." 2012.

[19] R. H. L. Chiang, C. E. H. Chua, and V. C. Storey, "A smart web query method for semantic retrieval of web data," *Data & Knowledge Engineering*, vol. 38, pp. 63-84, 2001.

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

[21] W. Wei, P. M. Barnaghi, and A. Bargiela, "Semantic-enhanced information search and retrieval," in *Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007. Sixth International Conference on*, 2007, pp. 218-223.

[22] J. Mayfield and T. Finin, "Information retrieval on the Semantic Web: Integrating inference and retrieval," in *Proceedings of the SIGIR Workshop on the Semantic Web*, 2003.

[23] R. Guha, R. McCool, and E. Miller, "Semantic search," in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 700-709.