RSSE: A Paradigm for Proficient Information Retrieval using Semantic Web

Sridharan. K[1], M. Chitra[2]

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

Abstract: In today's scenario, due to the rapid growth of content volume over the internet, the conventional search engines hardly do afford the required content relevant to the user's query. This can be effectively solved by enforcing semantic web search methodologies. On addressing that, this paper proposed an efficient prototype Relevancy-based Semantic Search Engine (RSSE). Moreover, the framework enables the users to locate relevant resources and services through semantics and domain expertise. A novel algorithm called Query Similarity Prediction Algorithm (QSPA) has been developed for proficient information retrieval with minimized processing time consumption and simultaneously, serving multiple remote users. Ranking is also performed based on the relevance score of retrieved documents to aid users for finding which documents are most likely to be relevant documents to the given queries. The experimental results reveal the efficiency of the proposed work with respect to the parameters such as precision, recall, F-measure, and time required to retrieve the results for queries.

[Sridharan. K, M. Chitra. **RSSE: A Paradigm for Proficient Information Retrieval using Semantic Web**. *Life Sci J* 2013;10(7s): 418-425] (ISSN:1097-8135). http://www.lifesciencesite.com.

Keywords: Cache Server, information retrieval (IR), service level agreement (SLA), semantic web, query similarity prediction algorithm (QSPA), ranking.

1. Introduction

There is a wide adoption of various Information Retrieval techniques over the past decades due to the continuous and fast growth of the content stored and shared on the web and other document repositories. This enlargement results in well known problems and difficulties such as finding appropriate results and managing the contents of all existing information in an effective manner. Moreover, it is well-known that the process of Information Retrieval can be effectively managed by search engines. The main component of as search engine is the Information Retrieval System that performs the significant tasks such as collecting the web pages and retrieving appropriate text documents that answers a user query. It is obvious that there is a huge quantity of text, audio, video and other documents on the internet. The users need to be capable to retrieve the appropriate relevant information to satisfy their particular information requirements. The figure 1 shows the generic architecture of search engine. Web crawler is incorporated for extracting the relevant document from the document corpus.

The conventional web search engines are used for searching and retrieving the results. But the major problem is to be claimed that the document or content retrieval is performed with respect to the keywords. Perhaps, this may not afford the most relevant or required information to the user query. The solution for this is to consider the semantic similarity of the query.

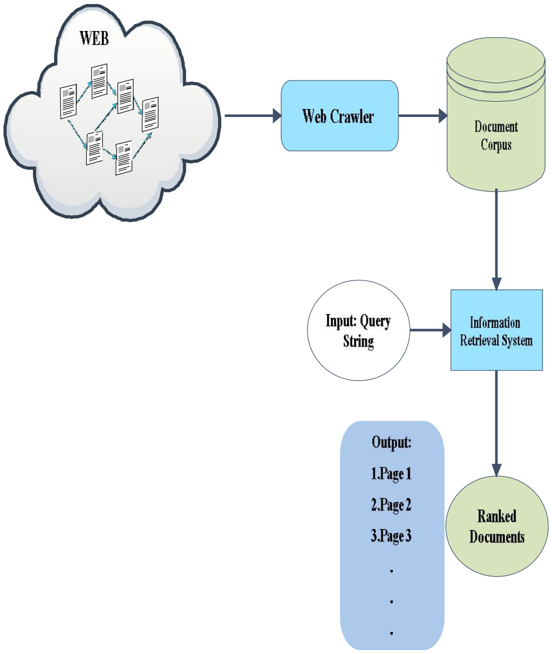

Figure 1: Generic Architecture of Search Engine

Semantic web can be stated as an extension of the current web. The main aim of semantic web is to provide better atomization, reuse and interoperability. Typically, semantic web is defined as a Web of relations between resources denoting real world objects. The process exploits the semantics of the queries or documents to be retrieved. Furthermore, retrieval on semantic web enhances the information search and retrieval results in two ways [24].

- Provides simple method to aid the semantic search module for better understanding of the denotation of the query
- Improves the relevancy rate of the search results

The semantic web is also involved in organizing and developing a web of semantic documents. Processing semantic markup is a significant task to be accomplished while Information Retrieval [16]. It is well-known that the input given to the retrieval system is a sort of semantic web query. Hence, there is a necessity of semantic markup encoding.

Before getting on to the detailed process of semantic web contribution on web search, it is to be stated about the kind of searches on Internet. Generally, there are two types of searches [7]:

1. ***Navigational searches:*** In this type of searches, the user submits the query on the search engine to find documents. The search engine will be used as a navigation tool to navigate to a particular required document.
2. ***Research searches:*** This type of search involves in denoting an object about which the user is trying to acquire information.

In semantic web, each page holds semantic metadata that record additional information pertaining to the web page itself. Further, the semantic web has the advantages such as clustering the results of multiple search engines, acquiring results in fast manner, etc. However, the semantic web is fundamentally similar to the web of HTML documents.

In order to handle the exponential progress and dynamic environment of World Wide Web, an effective caching mechanism is needed. This technique is incorporated to provide fast searching mechanism for the users. Every web browser has a built in local cache that stores the objects that the user needs with a specific motive that if any other user come back to browse the page in need of same information, it will load quicker. Caching can be stated as the automatic impermanent copies of information will be stored on host server or user's system for making the availability of information more facile.

The major contribution of this paper is to design a semantic based search engine that incorporates a novel algorithm called Query Similarity Prediction Algorithm (QSPA) for matching the given query with the stored information on cache. The framework is developed on the basis of Service Level Agreement (SLA) for caching the user activities and search data in a perfect manner. Furthermore, ranking is also enforced for sorting the results based on its relevancy rate concerning the user query. The ultimate goal of this work is to retrieve the search results by analyzing the context and semantics of the query with less time consumption and more precision.

The remainder of this paper is structured as follows: Section 2 confers about the related works; Section 3 summarizes the proposed method for RSSE. Section 4 reveals the experiments and results achieved. Finally, Section 6 concludes the paper with pointers to future work.

2. RELATED WORKS

Myriad researches are carried out for retrieving more relevant results from the World Wide Web. A Smart Web Query Method (SWQ) has been developed for semantic web search in [2]. The approach involved in formulating appropriate query using domain similarities based on context ontologies. Moreover, the work included the process of semantic search filtering that helps in identification and relevance ranking of web pages. Another work in [17] described the word semantics for Information Retrieval. The lowest-level word semantics was analyzed in this paper. The Word Semantic Model (WS) proposed in that paper keyword based matching in accordance with the consideration of word stem, part of speech, semantic indexing and searching. The concept of the paper provided a forwarded step for efficient semantic web formation.

SEAL (SEmantic PortAL) approach is given in [14]. The approach is developed for providing and accessing information at a portal along with its construction and maintenance. The SEAL architecture comprised knowledge warehouse and Ontobroker system. The authors have made the architecture in such a way of determining the types of users include software agents, community users and general users. The approach also performed the semantic ranking and personalization for attaining more appropriate results.

Wide difference between the traditional informational retrieval and semantic web was analyzed and discussed in [1]. Based on the description of this paper, the pillars of the semantic web are considered as ontology, markup languages and intelligent agents. An integrative approach for semantic web knowledge with web usage mining was discussed in [4]. The process accomplished for web personalization which defines the action that tracks the web experience to a particular user or a set of users.

The work in [10] described about the building process of Web ontology based browser and editor. The paper comprised a clear demonstration on hypertextual navigation, views and display of a site and the annotation mechanisms for exact search. There is an approach proposed for analyzing the semantics of queries and documents instead of analyzing its matching terms [5]. The process was named as Semantic Term Matching (STM). The work could be further extended using some lexical resources such as WordNet.

In an alternative way, definition for semantic similarity is given as the computation of similarity in a conceptual manner rather than the consideration of textual information [9]. Moreover, the paper comprised the discussion on various semantic similarity algorithms such as WordNet and MeSH and their issues. Query expansion and term expansion is performed for semantic similarity accumulation. Various methods for the process of Information Retrieval and Web Search have been effectively described in [15]. The authors described about the components involved in search engine development such as web crawlers, page repository, indexing module, query module, pertinent pages and ranking module. An overview for ontology and semantic web was given in [19]. It was claimed in the paper that ontology plays a vital role in the development of semantic web. Ontology-based information visualization was being the major core of their work. Following that, distributed information retrieval process in semantic web has been narrated by the authors of [23]. The process composed three steps:

1. Resource selection
2. Query reformulation
3. Data fusion and rank aggregation

The path from traditional World Wide Web to semantic web was effectively described in [3]. The difference between the above two scenario was analyzed in the aspects such as channels of message exchange, pattern of collaboration, etc. OntoLook system was developed for relation-based semantic search [13]. The key algorithm was incorporated for making concept-relation graph for the stored documents and given query. The priority based page ranking could be incorporated for the future studies. In application point of view, information retrieval and knowledge discovery on semantic web has been employed for extracting the data about traditional Chinese medicine [25].

A comparative study between Google and Yahoo was made in [12] with respect to the precision and relative recall of the search engines. The retrieval effectiveness of both search engines was compared on the basis of simple multi-word queries, simple one-word queries, complex one-word queries and simple one-word queries. The results of the study showed that the precision and recall rate of Google is comparatively higher than Yahoo. In distributed computing environment, support for multiple remote users in retrieving information on demand was discussed by the authors of [22]. Architecture for query handling was incorporated for distributed context management.

Personalized Semantic Search Engine (PSSE) architecture given in [20] is a crawler-based search engine that uses multi-crawlers for collecting information from web resources. There are three stages in PSSE, namely: Processing stage, searching stage and ranking stage. The user satisfaction could be further enhanced with the incorporation of effective mechanism that reduces the retrieval time. In [21], a tool was developed to handle the semantic data. This was developed on focusing two types of users: web designers and web application developers.

Semantic Information retrieval for extracting relevant data from the web documents focused crawler based on domain ontology. The authors claimed that the use of semantic information retrieval improves the retrieval performance than the conventional methods [8]. Following that, in [6], a system called SPIRS has been proposed based on semantic web and agents that supports expressive queries. A user model has also been incorporated to enhance the ranking of relevant documents. Another work [11] focused on four perspectives of designers and users such as static knowledge structure, high recall, low precision and lack of experimental tests. With an application point of view, Semantic Information Extraction in University Domain (SIEU) enclosed for a University domain has been proposed in [18]. The process comprised ontology construction, refined query formation and ranking of retrieved links.

3. PROPOSED WORK

In order to overcome the inefficiency of conventional Information Retrieval methods in extracting most relevant data from the web, a novel method is proposed here. In general, searching is almost done on the basis of word occurrences on the document. Typical search engines enhance this in the context of the web with information about the hyperlink structure of the web. Further, the availability of large volume of structured data about a wide range of objects on the semantic web provides some criteria for improving the traditional search models.

With that note, the proposed Relevancy-based Semantic Search Engine (RSSE) comprises a cache server that stores all the requests made by a specific category of users. There is a proposed algorithm called Query Similarity Prediction Algorithm (QSPA) involves in verifying the existence of results on the cache server. In such a way, it reduces the processing time of retrieving relevant documents and also enhances both the precision and recall rate. Figure 2 depicts a generic flow for Information Retrieval from web server.

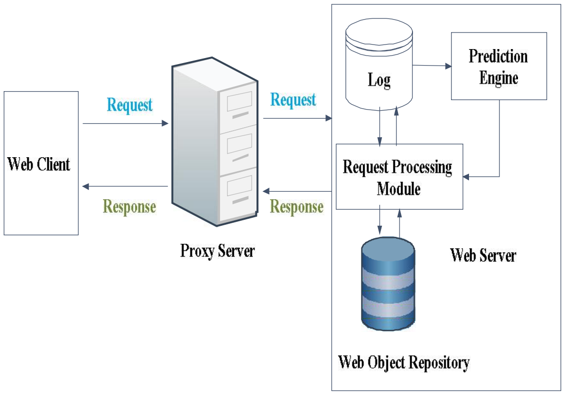

Figure 2: Flow for IR from Web Server

It is obvious from the figure that when a web client made a request, it will be processed through the proxy server on the web server. The prediction engine is responsible for finding the relevant information from the web database and extracts those documents. The block diagram of the proposed RSSE is given in Figure 3.

3.1 SLA based Accountability

The initial process for RSSE is to create accountability for users those are going to access the search engine to retrieve the required documents. For that concern, Service Level Agreement (SLA) has been framed for enabling the RSSE to users. As per the statements and constraints given on the agreement, the user has to register on the engine. The accessibility for the registered user will be provided in the form of unique log in ID. Through that ID, the RSSE is accessible for the corresponding user to retrieve the required information with more precision. The process also aids for the proposed algorithm for tracking the search history of users to retrieve of match with similar search queries. The Search Engine will be enabled only when the user creates proper accountability for their access. Moreover, the SLA comprises the norms such as the accessibility of data that are acquired by a specific registered user and stored on cache can be accessible to other users who are the authorized members of RSSE for searching the contents they require.

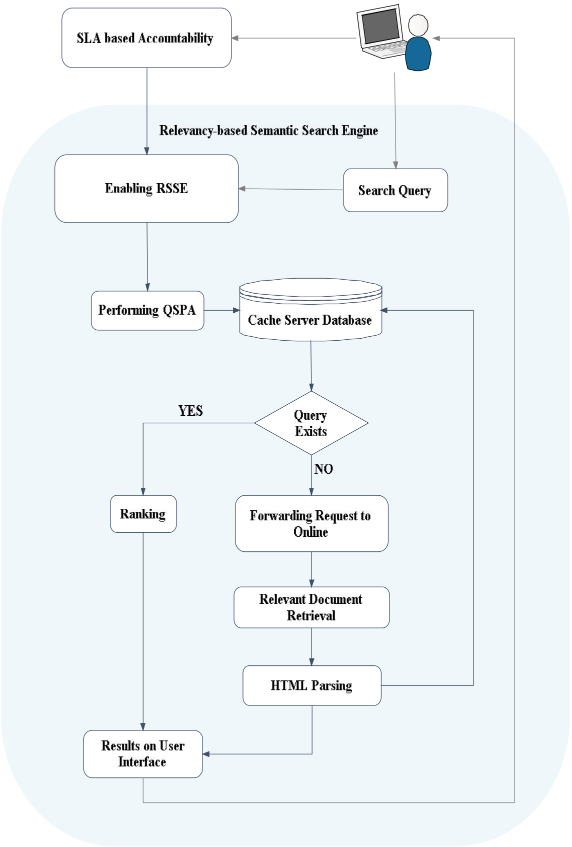

Figure 3: Block Diagram of RSSE

3.2 QSPA

Query Similarity Prediction Algorithm has been framed for making matches for the input query against the stored query results on cache server. Once the user requests a search engine for acquiring information, the request will be passed to the cache server and the process of query similarity prediction will be performed in an effective manner as per the algorithm shown below.

Based on SLA, the query given by an authorized user is checked against the data stored on cache for matching. The similarity measure will be computed with respect to the semantics of the given query. Semantic based similarity determination has been performed to enhance the precision rate of the retrieval results. Moreover, the results in the cache server are ranked as per the relevancy rate of the semantic terms of given query.

```
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
```

Algorithm 1: Query Similarity Prediction Algorithm

3.3 Retrieval of Results

The work of a search engine is to retrieve the available relevant data in various formats such as audio, video or text for a given user query. In the proposed Relevancy based Similarity Search Engine, process of relevant data retrieval is accomplished in two ways.

- Cache server based retrieval
- Online based retrieval

3.3.1 Cache Server based Retrieval

The QSPA algorithm explained above comes into effect only when the query is matched with any of the terms in the corpus of cache server database on the basis of semantic analysis. The semantics of word sequences are compared with the documents in the cache database. The relevant documents are ranked based on its relevancy score. The ranked documents are further displayed on the user interface.

3.3.2 Online based Retrieval

In another case that if an authorized user of RSSE gives a new request that does not have any match on the catch server corpus; the request will be thrown or forwarded for online processing. The searching process will be done at the web server. Semantic similarity between the given query and the documents on the web server is computed. The similar documents are retrieved and then that have to be stored on the local cache server for further performance of QSPA. For that, HTML parsing should be done.

4. EXPERIMENAL RESULTS

The proposed method is evaluated using the various metrics, namely precision, recall, and f-measure. The proposed search engine is also compared with the existing search engines such as Yahoo, and Google. The graph results depicted in the following figures are for the query XML process. Precision is the measure used to determine the fraction of retrieved results that are pertinent to the input query. The Precision value can be determined using the equation 1.

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

In equation (1), P_{RES}, and R_{RES} denotes the relevant result and the retrieved result derived for a single query. Figure 4, represents the precision values while executing the query ***"XML process"*** in Yahoo, Goolge, and our proposed search engine.

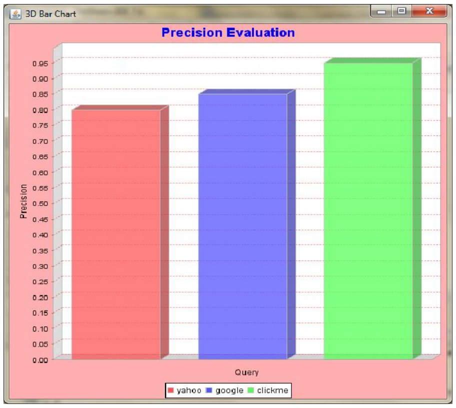

Figure 4: Precision Analysis

It also express that the proposed search engine outperforms the existing Google and Yahoo. The proposed RSSE has 0.95 as its precision value, whereas the Yahoo and the Google has 0.8 and 0.85 precision value respectively. Similarly the recall value is measured using the equation 2.

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

Figure 5 shows the recall value for the given query while executed in Google, Yahoo and the proposed scheme.

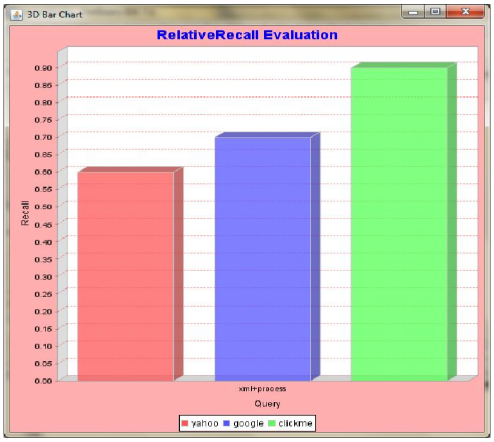

Figure 5: Recall Analysis

From figure 5, it is clear that the proposed technique has 0.9 as its recall value whereas the Yahoo and the Google has 0.6 and 0.7 as its recall value for the given input query ***"XML process".***

The accuracy of the proposed system is portrayed from the Figures 6, 7, 8, and 9 using the F-measure. The F-measure evaluation parameter considers both the recall and precision values.

The below figure depicts that the number of retrieved document retrieved is lesser than the irrelevant document retrieved for the input query using Google search engine.

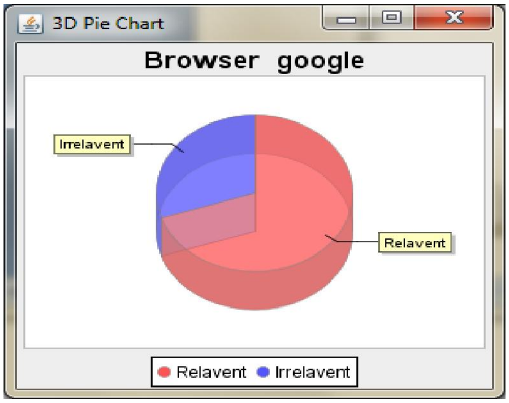

Figure 6: Number of relevant and irrelevant document retrieved using Google

Likewise, Figure 7 and Figure 8 depict the proportion of related documents retrieved for given query using Yahoo and Proposed scheme respectively.

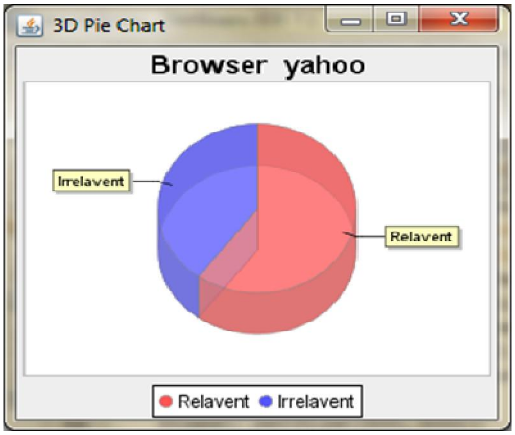

Figure 7: Number of relevant and irrelevant document retrieved using Yahoo

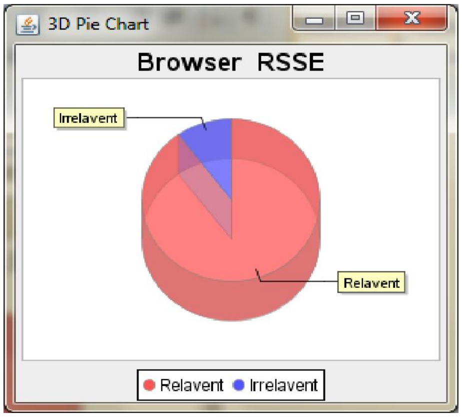

Figure 8: Number of relevant and irrelevant document retrieved using RSSE

The graph present in Figure 9 represents the F-measure values for the three search engines namely Google, Yahoo, and RSSE, which expresses explicitly that the proposed technique retrieves more accurate results for a query than other two techniques.

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 10.

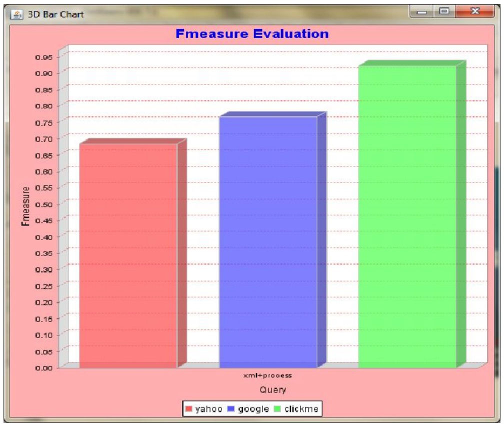

Figure 9: F-measure Analysis

It explicitly expresses that the time taken for the proposed RSSE search engine is lesser than existing methods. This paper measures the time factor is measured in milliseconds.

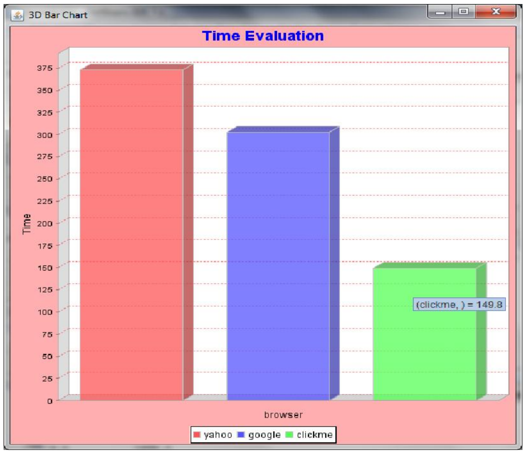

Figure 10: Time analysis

5. CONCLUSION

In order to retrieve the most pertinent documents for user query, this paper designed a new search engine RSSE, which enables the users to determine the location of pertinent services and/or resources through semantic expertise. This framework is designed along with an algorithm named Query Similarity Prediction Algorithm (QSPA), which efficiently serves multiple remote users simultaneously. The retrieved results are arranged through the ranking method. The relevancy score is used for ranking that is determined using the weight of the results. Experimental results in section 4 show that the proposed RSSE engine performs better than the existing search engines. The results are measured based on recall, precision, F-measure, and time taken to retrieve the results. The proposed method retrieves appropriate documents more exactly in less time than the Google and the Yahoo. In future the authors of this paper decided to secure retrieval process for their valuable customers.

REFERENCES

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

[2] 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.

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

[4] 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.

[5] H. Fang and C. X. Zhai, "Semantic term matching in axiomatic approaches to information retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 115-122.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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

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

[20] 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.

[21] M. Rico, Ó. Corcho, J. A. Macías, and D. Camacho, "A Tool Suite to Enable Web Designers, Web Application Developers and End-users to Handle Semantic Data," *International Journal on Semantic Web and Information Systems (IJSWIS),* vol. 6, pp. 38-60, 2010.

[22] 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.

[23] 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.

[24] 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.

[25] 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.