IJCTA, 9(27), 2016, pp. 347-352 ISSN: 0974-5572
© International Science Press

POBSIRS: Personalized Ontology Based Semantic Information Retrieval System

K. S. Ramanujam* **and K. David****

ABSTRACT

Searching and retrieving required information from the internet are very challenging task because of the huge amount of electronic data populated day by day. Most of the current web information retrieval (IR) systems brings out information based on user's query keywords. Currently, it is inadequate because of huge amount of online data and it has less accuracy since the system considers syntactic level search. Mostly the current IR systems give the same information to all kind of web users. So, to improve the information retrieval accuracy an IR system that search and retrieve information semantically with user personalization is required. In this paper, a novel approach to retrieve internet information is proposed using semantic concepts such as ontology that is designed to satisfy users requirement while extracting data from the web. Additionally, a new page ranking algorithm is proposed using the time spend on each website and the users past search history is utilized to give personalized information from the internet. Experiments are conducted to test the performance of the proposed semantics based information retrieval system. The results shows better accuracy comparing to state-of-the-art information retrieval systems.

Keywords: Information retrieval, semantic web, ontology, search engine, personalization, page ranking.

1. INTRODUCTION

Information from internet can be retrieved in two ways, Syntactic based and Semantic based search and retrieval. The technology giant google has introduced the keywords based information search and retrieval system few decades ago. From that time, many information retrieval systems have been developed and deployed to search and retrieve appropriate information from the internet. Syntactic level information retrieval system provides limited capabilities to capture the concepts of the user needs and the relation between the keywords. Also, the syntactic based search often does not satisfy the users' search queries. Thus to satisfy the users' requirement in information retrieval the idea of the concept or semantic based search is introduced.

This paper deals with the Semantic Based Information Retrieval System to extract information from the web in a more efficient way. The main contributions of this paper are as follows:

1. A new architecture for web information retrieval is proposed that takes list of plain keywords from the web users.
2. The query keywords taken from users are then converted into semantic query.
3. The semantic query is defined utilizing the domain concepts of the pre-existing domain ontologies and thesauri.
4. The web information are retrieved using semantic query and then ranked to show top ranked information to the users.
5. A new page-ranking algorithm is proposed using time duration that users spent on each website. Users past search histories are also considered to extract personalized information from the internet.

* Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India, *Email: loginprabu@gmail.com*

** Assistant Professor, Department of Computer Science, H.H. Rajah's College, Pudukkottai, Tamilnadu, India, *Email: jdbdavid@gmail.com*

The remaining sections of this paper is organized as follows: section 2 presents related works on semantic based information retrieval methods, section 3 provides a note on semantic based information retrieval system, section 4 gives detailed explanation of the proposed semantic web search technique. Experiments and results are presented in section 5 and conclusion of this paper is given in section 6.

2. RELATED WORKS

The world wide web (internet) is a widely distributed information center that serves information to the users. Day by day the size of the internet data is increasing very rapidly. Hence information retrieval from the internet is becoming tedious to retrieve desired content in desired time. Doing manual work to find required data is highly time consumable, so a system that extracts information automatically from the web is required. Information retrieval (IR) is an area that helps to search and retrieve documents from the world wide web (www).

The current IR systems deal with representing, storing and organizing the www content. In the web 2.0 standard finding unknown knowledge is very hard to achieve where there is no relationship is defined among www information. Hence, information needs to be connected with ontology or semantic relationships. To get desired information now users are heading towards web 3.0 version where data are represented and stored with ontological relations. Many information retrieval concepts are proposed to get information from the web.

The papers [1, 2, 3] are few of them. Recently, to extract appropriate information from the web the papers [4, 5, 6] are proposed in the information retrieval literature. They utilized machine learning techniques with ontology based knowledge representation to improve data retrieval accuracy. The ontology based IR systems exploits many existing resources such as texts, annotated contents, thesauri and indexing databases. Also, it uses machine learning techniques from information retrieval and agents [7] fields. The making of a web ontology is given in [8] that deals with rdf to owl ontology representation of web contents. A detailed analysis of the semantic web mining can be found in [9]. Generic web content or information retrieval is not suitable for all web users, so information retrieval methods using personalization [10, 11] are proposed in the information retrieval literature.

3. PROPOSED POBSIRS METHODOLOGY

The proposed information retrieval system contains six steps such as web spider, semantic annotator, semantic indexer, semantic query converter, semantic content retriever and content ranking. The architectural diagram of the proposed POBSIRS methodology is shown in figure 1.

3.1. User profiling

Generic information retrieval systems provide same results for all web users. Different users has different expectations, likes and dislikes. So it needs personalized information retrieval system. Users details such as age, gender, education, IP addresses, location etc and past search history has some advantages in providing best information to the users. The past history reveals what a particular user likes or dislikes. The webpages which have seen long time are treated as users interest and the websites which the users does not see or spent very less time are treated as uninterested webpages. It will be used to provide personalized information to the web users. The processing of collecting past history of the user is called user profiling that will be used in the next step for web page ranking.

3.2. Web spider

The web spider is an internet technology that systematically extracts information from web pages from various domains which will be saved in a web database for web indexing (or web spidering). Web search engines usually utilize the web spidering (or crawling) approach to update their web content for indexing purpose. The web crawlers also track the pages based on users search history for later processing by a search engine. That will provide an appropriate and personalized information to the web users.

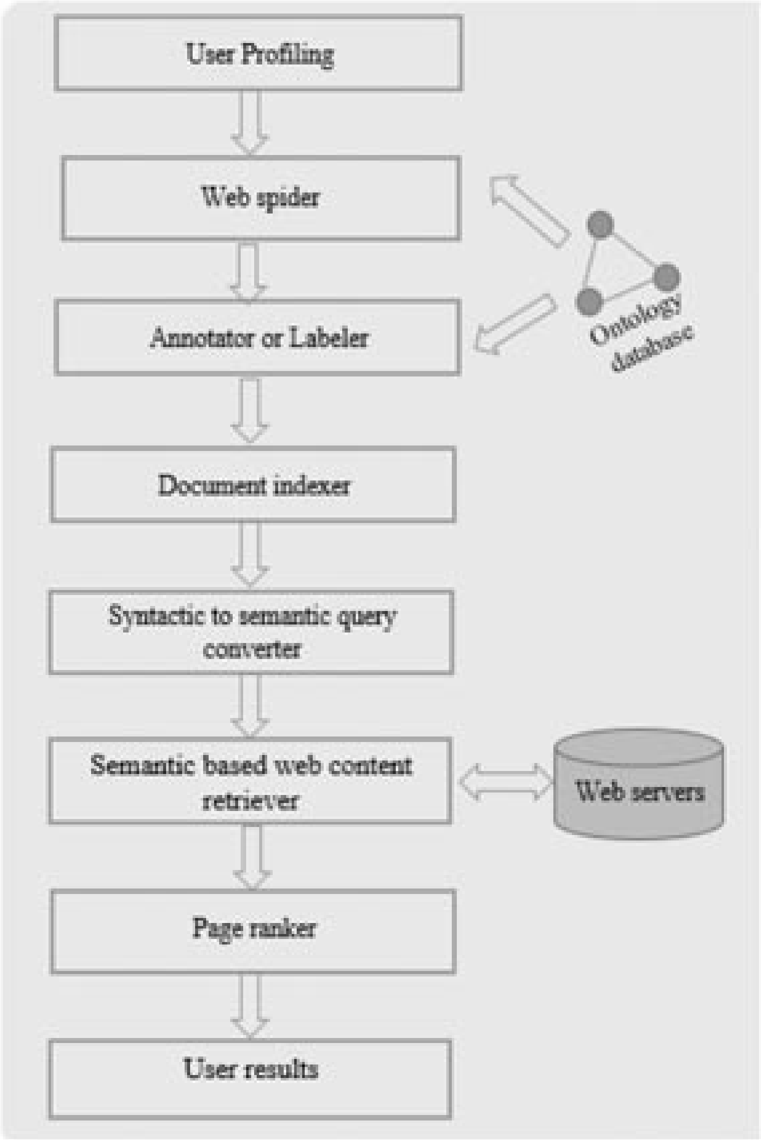

Figure 1: The architectural diagram of the proposed POBSIRS methodology

3.3. Semantic content labeler

Web documents need to be annotated in order to search them easily. The web content annotation gives a meaningful index for web documents and their associated websites to make them easily accessible. The content generated from this semantic annotation stage will be given to next stage i.e., semantic indexer stage.

3.4. Semantic based document indexer

The annotated content from the previous step is collected and converted into a knowledge base. Using that knowledge base the concepts are indexed for further information processing. The indexing is computed by applying the tf - idf model.

3.5. Syntactic to Semantic query converter

The user query keyword will be converted into three different ways. First the keyword matching with domain ontology, second keywords linking to websites are taken and finally the thesaurus are used to get more meaningful words for semantic search. Based on this the user query will be submitted to a search engine to retrieve documents from the web servers. Here, both syntactic and semantic of the user query are used for searching the web content.

3.6. Semantic based web content retriever

This step does the job of retrieving and projecting relevant content or document of information to the users' query. The semantic query produced in the previous step (semantic query converter) is matched with the ontologies to search for the content. Based on the users' query terms i.e. semantic query terms, the web documents are searched and retrieved then given to the ranking stage of the information retrieval system

3.7. Personalized page ranking algorithm

The contents i.e., web documents retrieved in the previous step are ranked here. The ranking will be generated comparing the given semantic query terms and each document indices. The top ranked documents are assumed as the most relevant documents that are shown to the users. In the page ranking algorithm, the number of times the websites clicked is considered for relevancy checking. High in clicks will have higher priority and low in clicks will give lower priority in page ranking. The original page rank algorithm [12] is given by

$$PR(A) = (1-d) + d\ (PR(T_i) / C(T_i) +...+ PR(T_n) / C(T_n)) \qquad (1)$$

where PR(A) denotes the PageRank of page A, $PR(T_i)$ is the PageRank of pages that (incoming) link to page A, $C(T_i)$ is the number of outbound (outgoing) links on page T_i and d_i is a damping factor which can be set between 0 and 1 as said in [12]. In conventional page ranking algorithms stated above, the websites or documents are ranked based on number of incoming and outgoing links the website holds. In addition to that the time spent on each website is taken in the proposed page ranking algorithm. It is inferred that the user spending more time in reading a webpage is more important than the webpages that users spent less time. Thus, the proposed time based page ranking has an advantage to improve information retrieval accuracy. Hence the equation 1 can be rewritten as follows,

$$PR(A) = (1-d) + d\ (PR(T1) / C(T1) +...+ PR(T_n) / C(T_n)\ ) + V_t \qquad (2)$$

Here, V_t is the time spent on each web site by a user. The performance of this proposed personalized page ranking algorithm is evaluated and results are observed. The proposed POBSIRS system is an improved system of the SBIRS [12]. The results are given in the following experiments section.

4. EXPERIMETAL RESULTS

In this section, the experimental setup and the implementation of the proposed POBSIRS are explained in detail. The experimental results of the proposed POBSIRS are compared with two recently prosed information retrieval systems.

4.1. Experimental setup and implementation

The experiments are conducted offline mode. For that, 2500 webpages of various domains such as sports, politics, research and development, Healthcare and education. The proposed POBSIRS approach is implemented using C#.NET as a web-based system following [13]. Protégé tool is used to develop domain ontologies like [14] and SPARQL is used for querying the system to retrieve relevant documents. The web contents are manually annotated by human experts in the annotation domain.

4.2. Results

The experiment initiates with the syntactic query to semantic query conversion. Then the semantic query is fed into the annotator that matches the query with the web documents annotated content. Then the web documents are ranked based on the proposed page ranking algorithm. The effectiveness of the information retrieval system is

Table 1
Performance evaluation of the information retrieval systems

Search query	*KBIRS*		*SBIRS*		*Proposed POBSIRS*	
	Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
1	1.00	0.55	0.95	0.61	0.96	0.65
2	0.36	0.49	0.86	0.74	1.00	0.77
3	0.61	0.61	0.74	0.62	1.00	0.63
4	0.24	0.52	0.79	0.66	0.92	0.80
5	0.47	0.66	0.67	0.66	0.79	0.97
6	0.41	0.59	0.85	0.67	0.84	0.75
7	0.36	0.61	0.86	0.71	1.00	0.75
8	0.29	0.58	0.66	0.68	1.00	0.84
9	0.37	0.54	0.69	0.69	0.96	1.00
10	0.49	0.68	0.72	0.73	0.98	1.00

usually measured by the ratios Precision and Recall. In this experiment, the standard performance evaluation metric precision and recall are computed to test the performance of the proposed POBSIRS retrieval system.

Precision = number of relevant documents retrieved / total number of retrieved documents

Recall = number of relevant documents retrieved / total number of relevant documents

In this experiment, 10 different keywords are used to query the information retrieval system. The results are compared with two other existing information retrieval systems called KBIRS [14] and SBIRS [14]. The observed performance are provided in the table 1 where it is obvious that the proposed system outperforms other two existing information retrieval systems. The retrieval time comparison among the SBIRS, KBIRS and the proposed POBSIRS is shown in figure 2, where it is noticeable that the response time of the proposed POBSIRS is higher than the comparative methods. This is due to the additional steps of user profiling and time parameter in the proposed approach. However, the performance of the proposed POBSIRS is higher than other comparative methods which is shown in table 1.

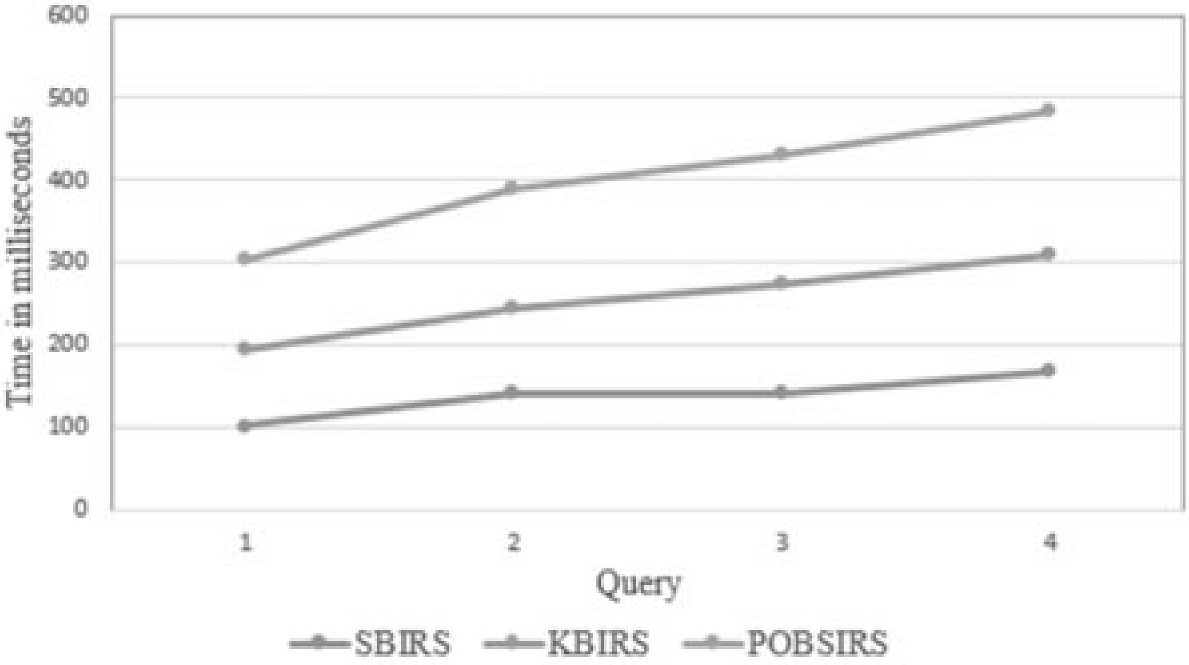

Figure 2: Information retrieval response time comparison among SBIRS, KBIRS and the proposed POBSIRS

6. CONCLUSION

In this paper, a novel information retrieval system denoted POBSIRS is proposed that bridges the gap between IR and SW in the understanding and realization of semantic search. It address the issues in syntactic based information retrieval and provides a novel semantic based information retrieval system. Also this paper proposed a new page ranking algorithm using time duration spent on each websites. The experimental results shows that the proposed POBSIRS is highly preferable than other IR systems. In future, the document ranking and semantic query conversion will be focus to further improve retrieval accuracy.

REFERENCES

[1] M. Hepp, “Semantic Web and semantic Web services: father and son or invisible twins?. “ IEEE Internet Computing **10.2** (2006): 85-88.

[2] L. Berners, et al., “Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor”, HarperInformation, 2000.

[3] W. Yong-gui and J. zhen, “Research on semantic web mining 2010”, International conference on computer design and application (ICCDA), vol, 1, pp. **67-70**, 2010.

[4] S. Dumais, et al. “Stuff I’ve seen: a system for personal information retrieval and re-use.” ACM SIGIR Forum. Vol. 49. No. 2. ACM, 2016.

[5] T. H. Haveliwala, M. J. Glen, and D. K. Sepandar, “Variable personalization of search results in a search engine,” U.S. Patent No. **9, 058, 364**. 16 Jun. 2015.

[6] S. Fitchett, and C. Andy, “An empirical characterisation of file retrieval.” International Journal of Human-Computer Studies **74** (2015): 1-13.

[7] A. Seaborne and E. Prud’ hommeaux, SPARQL query language for rdf, W3C Recommendation, W3C January, 2008.

[8] P. F. P. Horrocks, F.V. Schneider Harmeten, “from rdf to owl: the making of a web ontology language”, journal of web semantics 2003.

[9] A. Chakravarthy, “Mining the semantic web”, In proceedings of the first AKT Doctoral Colloquim, 2005.

[10] S. Chawla, A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search. Applied Soft Computing, 46, **90-103**, 2016.

[11] L. Shou, H. Bai, K. Chen, and G. Chen, “Supporting privacy protection in personalized web search”, IEEE transactions on knowledge and data engineering, **26(2), 453-467**, 2014.

[12] L. Page, et al. “The PageRank citation ranking: bringing order to the web.” (1999).

[13] M. Thangaraj, and G. Sujatha. “An architectural design for effective information retrieval in semantic web.” Expert Systems with Applications **41.18** (2014): 8225-8233.

[14] N.F. Noy, et al., Creating semantic web content with Protégé-2000, in protégé 2000, IEEE Intelligent system, vol 16, no. 2, 2001, pp. **60-71**.