A Study and analysis on Web Information Retrieval System for Distributed Environment

S. Meenakshi  
Associate Professor, R. M. K. Engineering College,  
Tiruvallur District, Kavaraipettai, Tamil Nadu, India.

Dr. R. M. Suresh  
Dean, Vel Tech Dr. RR & Dr. SR Technical University,  
Avadi, Chennai-600 062, Tamil Nadu, India.

Abstract
Although modern methods of information retrieval have enormously improved our ability to find relevant information for distributed environment, it remains the case for the foreseeable future that the best performance can only be obtained by some pre-processing of the documents to be searched. Web Information retrieval system assumes a noteworthy part in retrieving the information from a larger collection of data. Queries in the web will retrieve both relevant and irrelevant information. Users have to restructure using appropriate search terms to retrieve relevant information. There are many ways to retrieve relevant information. Some of them are based on service-based, Context based, User Profile, User Context based, Query based, domain specific based vertical search, Concept based, Query expansion and User Relevance, Comprehensive search service, Methontology based semantic web documents and Ontology based approach. Semantic Information Retrieval Comprehensive Domain Service Ontology based approach always improves the information retrieval performance as it could retrieve relevant and exact web documents and ignore irrelevant web documents even it contains search key words thereby providing energy efficiency, less effort but on more precise results. In this manuscript articulate the semantic web search approaches, methodologies and mechanism involved and study on many of the semantic enhanced web information retrieval system for distributed environment. Considering the time of Response, dead and redundant links, use of Ontology annotation knowledge representation methodology based on Concept Relevancy Ranking of Link and Page Contents algorithmic approach in Semantic enhanced information Retrieval is the approach of decision[14].

Keywords: Web Information Retrieval, Semantic Annotation, web service, ontology and semantic web.

Introduction
Web Information Retrieval (WIR) is quite an influential topic nowadays and quite near our daily life. It is the task of finding relevant information from a larger collection of unstructured data. WIR systems now become the main source of interaction with the internet. Search Engine is one of the ways to retrieve information from WWW in order to find all the documents relevant for a user query in a collection of documents.

Decades of research in WIR were successful in developing and refining techniques that are solely word-based. With the advent of the web new sources of information became available, one of them being the hyperlinks between documents and records of user behavior. The goal of WIR system in Figure 1, is to satisfy user’s information need. Unfortunately, characterization of user information need is not simple. User’s often do not know clearly about the information need. Query is only a vague and incomplete description of the information need.

![Figure 1. Web Information Retrieval Process](image)

This study and analysis manuscript is organized as follows: Section 2 gives the related work to this manuscript. Section 3 describes the manuscript goal and Section 4 gives the conclusion and the future work.

Semantic Enhanced Information Retrieval-Related Work
The traditional Web Information Retrieval system depends on the comparison of the keyword with the exact match of the words in the document. This method will retrieve the pages that are partial relevant to the user query. Web offers some opportunity to improve the traditional search. One of the ways is with the help of semantic search. Semantic Information Retrieval seeks to improve search accuracy by understanding searcher’s intent and the contextual meaning of terms as they appear in the searchable data space, whether on the Web or within a closed system, to generate more relevant results.
Semantic Enhanced Information retrieval is a solution to the Information retrieval problem because the main goal is to provide the relevant information according to user’s need and interest. Semantic Information Retrieval is a data-enabled process that is based on three types -- first-based on users, second-based on website usage, third-based on software and hardware. In this paper, gives a review of Semantic Information Retrieval based on data about the user and its process is explained as in Figure 2.

![Figure 2. Semantic Web Information Retrieval Process](image)

Semantic Information Retrieval can be applied to the many approaches based on their characteristics and features. We will discuss each of this survey details in summary as follows:

In service based approach, Semantic Information Retrieval facilitates the development of distributed, service oriented information retrieval system over the heterogeneous internet. It includes network topology structure, service calling process, task allocating and load balancing as well as time-threshold settings.

In Context based and user context based approaches, Semantic Information Retrieval can be implemented as a work that proposes a hybrid approach to content clustering that combines the best of web information retrieval methods and also used the personal preference information of the users modeling a wide range of contexts.

In User Profile approach, Semantic Information Retrieval can be done by making use of tag data for evaluating personalized retrieval system involving thousands of users. Analogously to studies involving implicit feedback mechanisms in IR, which have found that profiles based on the content of clicked URLs outperform those based on past queries along, it found that profiles based on the content of bookmarked URLs are generally superior to those based on tags alone.

In Query based approach, Semantic Information Retrieval can be implemented in a semantic web based personalized learning service for enhancing the quality of recommendations based on the underlying structure of a web site. It introduces usage-based page rank, a page rank-style algorithm that relies on the recorded usage data and line analysis techniques based on user interested domains and user query.

In domain specific based vertical search, use of Semantic Information Retrieval as model for the medical domain. The vertical search engine uses a web crawler to fetch the web pages from a seed URL provided. The web pages crawled is checked for relevance based on the domain chosen and indexed. An inverted index based on page ranking is created. In Concept based approach, Semantic Information Retrieval can be done towards identifying relevant pages through Semantic Information Retrieval path analysis and provides an effective personalized web search. This improves web search by providing content and individual based relation between the search query and its relevant web pages.

Comprehensive search service always provides Semantic Information Retrieval by considering the varying interests of the user. It provides navigation assistance and the position of each result set to aid the user in selecting the relevant required documents. This is at the cost of increased storage and computational complexity.

Use of Ontology annotation knowledge representation methodology and descriptive metadata repository mechanism based on Concept Relevancy Ranking of Link and Page Contents algorithmic approach in Semantic Information Retrieval as a layered model of semantic web provides solution to the problem by providing tools and technologies to enable machine readable semantics in current web contents. It describes the model through ontology development and metadata repository.

During the paper, contrasted all the approaches firstly with theoretically and then practically and discover advantages that lead to the Ontology annotation knowledge representation methodology, advantages is:

Users could perform more accurate queries while being eliminating more than 90 percent of the irrelevant documents.

Semantic Information Retrieval performs better than traditional searching methods in case of semantically meaningful sentences or phrases but will fall short for keyword based search.

**Manuscript Goal**

**Proposed Approach**

In the Semantic Web, Ontological models permit the annotation of Web documents (modeling the representation of information contained in them) and therefore the formulation of more exact queries to retrieve documents. Annotation typically includes instances of concepts from the ontology to represent specific entities recognized in the resources, and afterwards linking this metadata to the resource as its description, a new methodology namely – ontology annotation knowledge representation is introduced to rank the relevant pages based on the domain concepts and keywords rather than keyword.
Proposed approach is the descriptive metadata repository mechanism based on concept retrieval algorithm in Semantic Information Retrieval as a layered model of semantic web provides solution to the problem by providing tools and technologies to enable machine readable semantics in current web contents. It describes the model through ontology development and metadata repository.

In this approach using Ontology based approach in Semantic Information Retrieval; it describes two ontology-domain ontology and service ontology, in order to improve the relevance of the results presented to the user. This approach is applied to Enterprise domain and noted that it makes it possible to improve the precision of research and thus the relevance of the documents returned to the user. Now we describe the components of a semantic enhanced WIR System functions in the following steps as shown in Figure 3.

**Data Set**
In this the input data set is Extracted Search Engine Results Page from search engines viz. Google, Yahoo, Bing, Ask And Aol, Etc and output data set is Re-ranked Search Engine Results Page based on Concept Relevancy Ranking of Link and Page Contents Algorithm and initially SERP’s are extracted based on the user query. Pre-process both user query and SERP for domain ontology and semantic annotation.

Root words are extracted from the user query to form a repository. Here the link content and page content of the SERP’s are checked with repository so that the more relevant pages are retrieved.

**Pre-processing**
Pre-process user query and extract root words, which are considered for constructing Repository and it is built along with its domain ontology and semantic annotation.

**Link content and page content determination**
Pre-process and extract the link content[9] and page content keywords for the search engine result pages and compared against the Repository. If match found then corresponding strength is granted each word.

**Relevancy Calculation**
The relevancy is calculated based on how well the results matches the query in addition to how related the retrieved index items of the results to the query.

After finding the web pages on the proposed approach[2] relevancy for the particular Search Engine Results Pages against user query is computed by summarizing all the strength of the link contents and page contents by use of damp factor d.

The search result page’s total relevancy is ranked in increasing order.

**Re-ranking**
Finally re-rank the search results[15][20] on Total relevancy in increasing order. The Top Search Result is the most relevant and bottom is the least relevant for the User query.

The following Table 1. Summarized the current approaches, methodology and mechanism of semantic enhanced web information retrieval with contrast to the proposed approach.

**Conclusion and Future Ahead**

**Conclusion**
Web Information Retrieval is an important and overwhelming research problem. It provides solutions related to service based, Context based, User Profile, User Context based, Query based, domain specific based vertical search, Concept based, Query expansion and User Relevance, Comprehensive search service, Ontology annotation knowledge representation methodology based on Concept Relevancy Ranking of Link and Page Contents algorithmic approach.

The proposed approach gives far better results when compared with the specified traditional search. From the experimental results, it is proved that using the semantic search with text content, most relevant pages are retrieved. This approach was evaluated and verified with different queries, and the relevancy was proved.

In evaluating the performance of WIR for distributed environment, it is observed that by using concept relevancy ranking of Link and Page Content algorithm, users could perform more accurate queries while being eliminating more than 90 percent of the irrelevant documents.

Due to ever-increasing complexity and voluminous of data, there is needly and certainly a room for further advancements. Further research is going on to refine.
Table 1: Current approaches, methodology and mechanism of Web Information Retrieval with Proposed Approach

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Web Information retrieval approach</th>
<th>Methodology/Algorithm/Mechanism</th>
<th>Domain/Implementation</th>
<th>Dataset</th>
<th>Publications</th>
<th>Dead and redundant links (out of top 10 links)</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Service based approach</td>
<td>Network topology structure, service calling process and task allocation</td>
<td>Heterogeneous</td>
<td>Heterogeneous data</td>
<td>Jian Meng, Zhao Yan, Ji Li, “A Web services-based distributed information retrieval model”, <em>IEEE</em>, 2008</td>
<td>NA</td>
<td>Threshold are briefly addressed</td>
</tr>
<tr>
<td>4</td>
<td>User Profile approach</td>
<td>Tag data and contents</td>
<td>NA</td>
<td>Tag data</td>
<td>Mark J. Carman, Mark Baillie, Fabio Crestani, “Tag Data and Personalized Information retrieval”, <em>SSM’08, ACM, October 30, 2008</em></td>
<td></td>
<td>Restrictions for profiles based on the content of bookmarked URL’s</td>
</tr>
<tr>
<td>5</td>
<td>User Context based approach</td>
<td>Semantic concepts, local information space</td>
<td>Online learning</td>
<td>Linked pages</td>
<td>Dr. Ritu Soni, Mrs. Preethi Bakshi, “Context Based Personalized Search Engine for online learning”, <em>IJCSRIS, International Journal of Computer Science and Information security</em>, Vol. 8-No. 7, October 2010</td>
<td>5</td>
<td>Queries are restricted to context</td>
</tr>
<tr>
<td>6</td>
<td>Query based approach</td>
<td>Usage based Page Rank Algorithm</td>
<td>Online course resource system</td>
<td>C-logs and the document cluster</td>
<td>Mahendra Thakur, Yogendra Kumar Jain, Geetika Silakari, “Query Based Personalization in semantic web mining”, <em>IJACSA, International Journal of Advanced Computer Science and Applications</em>, Vol. 2-No. 2, January 2011</td>
<td>6</td>
<td>Indexing and Searching large instances of data is still a challenge</td>
</tr>
<tr>
<td>Query formulation and Relation based Page Rank Algorithm</td>
<td>Homogenoues data</td>
<td>NA</td>
<td>It does not work well for long queries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meta data creation and resource annotation</td>
<td>Domain-task data</td>
<td>Fabrizio Lamberti, Andrea Sanna and Claudio Demartini, “A Relation-Based Page Rank Algorithm for Semantic Web Search Engines”, IEEE on Knowledge and Data Engineering, Vol. 21-No. 1, January, 2009</td>
<td>Recorded usage data and line analysis techniques based on user interested domains and user query.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inverted index algorithm for keyword search</td>
<td>Medical domain, Jdk 1.6, MySQL Server</td>
<td>Association rules Education</td>
<td>The presented method improves the precision and recall metrics of web retrieval system.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person data, train heading</td>
<td>Semantic decision tree</td>
<td>Web Data</td>
<td>D. Jeon and W. Kim, “Development of Semantic Deci-sion Tree,” Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications, Macau, 24-26 October 2011, pp. 28-34.</td>
<td>NA</td>
<td>The presented algorithm has more complex and rich expression, has a number of limitations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Association rules</td>
<td>NA</td>
<td>DBpedia data set</td>
<td>Z. Abedjian and F. Naumann, “Context and Target Configurations for Mining RDF Data,” Proceedings of the 1st International Workshop</td>
<td>NA</td>
<td>Further research is required</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No.</td>
<td>Topic</td>
<td>Methodology</td>
<td>Resource</td>
<td>Additional Information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------------</td>
<td>----------</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Query expansion and User Relevance</td>
<td>Pseudo local , Feedback (PLF)</td>
<td>User Query in text file</td>
<td>Ashok S.O., Vikas Kumar, and Sanjay K. Jena, “An Extended approach to Personalize the web search by measuring the user relevance” 2010</td>
<td>More on relevance data and precision and relative recall values on time for search results are not considered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Comprehensive search service</td>
<td>Collaborative web search</td>
<td>Result Set</td>
<td>Maurice coyle and Barry Smyth, “Supporting Intelligent Search”, ACM Transactions on Internet Technology, Vol 7-No. 4, October 2007</td>
<td>It concentrate at the cost of increased storage and computational complexity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Ontology based approach</td>
<td>Query processing and reformulation</td>
<td>User Text Query data</td>
<td>Marie-Aude, Aufaure Rania Soussi, Hajer Baazaoui, “SIRO-On-Line semantic Information retrieval using Ontologies”, IEEE, 200</td>
<td>Tested on small data sets and querying is in general slower</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Java (jdk 1.6)</td>
<td>Data mining as dataset1, SE as dataset2 and mobile communication as Dataset3.</td>
<td>Poonam Yadav,R.P. Singh, “An Ontology-Based Intelligent Information Retrieval Method For Document Retrieval” , Poonam Yadav et al. / International Journal of Engineering Science and Technology (IJEST) , 2012</td>
<td>Enhancements to the system can be done by improving the calculation strategy of the mutual association value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------------------------</td>
<td>--------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed Approach</td>
<td>Future Ahead</td>
<td>Acknowledgments</td>
<td>References</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------</td>
<td>----------------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
for Semantic Web Search Engines”, IEEE Transactions on Knowledge and Data Engineering, Vol. 21-No. 1, January, 2009


[38] G.Poonkuzhali, R.Kishore Kumar, R.Kripa Keshav, K.Thiagarajan K. Sarukesi, “Statistical Approach for Improving the Quality of Search Results”, Recent Researches in Applied Computer and Applied


