Semantic Based Efficient Retrieval of Relevant Resources and its Services using Search Engines

Pradeep Gurunathan[#1], Dr. Seethalakshmi Pandian[*2]

[#]Research Scholar, Anna University, Chennai, Tamilnadu, India

[1]pradeep_g8@yahoo.com

*University College of Engineering, BIT Campus, Tiruchirappalli – 620 024, Tamilnadu, India

Abstract - **The main objective of this paper is to propose an efficient mechanism for retrieval of resources using semantic approach and to exchange information using Service Oriented Architecture. A framework has been developed to empower the users in locating relevant resources and associated services through a meaningful semantics. The resources are retrieved efficiently by Modified Matchmaking Algorithm and dynamic ranking, which shows an improvement in search technique provided by the proposed search mechanism. The performance of retrieval of the proposed search mechanism is computed and compared with existing popular search engines like google and yahoo which shows a significant amount of improvement.**

Keywords: SOA, Semantic Web, Modified Matchmaking Algorithm.

I. INTRODUCTION

The current architecture of web search engines consists of various mechanisms such as crawling, indexing, retrieving, and ranking. In order to go beyond the current state of the art, next generation web search engine has to perform a deeper analysis of available content for presenting the user with a more structured search result than simply a list of links [1]. Instead of merely exploiting the syntactic structure of the Web and its documents, it is possible to leverage semantic information about Web resources. Thus, web search need not be stopped with documents. Rather, it is an interface for finding Web-mediated solutions and functionalities of any type based on user request.

The semantic web is an extension of the traditional web where information is not only human-readable but it is also machine-readable. The success of the semantic web depends mainly on two factors namely the existence of available semantic information on the web and the use of such semantic information. The semantic information encompasses ontologies to represent the knowledge and semantic data referred by such ontologies. Therefore, ontologies and semantic annotations are fundamental for the creation and publishing of semantic content. In fact, in semantic web, each search processes the metadata that has additional details concerning the web page itself. In this paper, an architectural framework has been introduced that identifies the functional and non functional components that are to be addressed by process mediation components. Specifically, we focus on search algorithms to retrieve relevant educational content in the formats such as document, images and videos along with discovery of external services. We present the Modified Matchmaking Algorithm for efficient retrieval of information is presented and the performance of retrieval such as precision and recall of the proposed search mechanism with other popular search engines like yahoo and google are calculated.

The rest of the paper is organized as follows. We briefly review related work in Section II. Motivation of the semantic search is discussed in Section III. Section IV devises content-based search. Section V discusses the experimental results of the work. Section VI evaluates the efficiency and effectiveness of retrieval in the proposed search mechanism over other search engines. Finally, the conclusion of the work is shown in Section VII.

II. RELATED WORK

In this section we briefly discuss some of the research work related to retrieve documents and locating Web services.

Current Information Retrieval (IR) approaches [4] do not formally capture the explicit meaning of a keyword query, but provide a comfortable way for the user to specify the information needs on the basis of keywords. Ontology-based approaches allow for sophisticated semantic search but impose a query syntax which is more difficult to handle. They presented an approach for translating keyword queries to Description Logic conjunctive queries using background knowledge available in ontology and implementation shows that this interpretation of keyword can then be used for both exploration of asserted knowledge and for a semantic based declarative query answering process.

Raymond Lau, Dawei Song, Yuefeng Li, Terence Cheung and Jin-Xing Hao [17] presented a fuzzy domain ontology extraction method for adaptive e-learning. Accordingly, instructors are often overwhelmed by the huge number of messages created by students through online discussion forums. It is quite difficult, if not totally impossible, for instructors to read through and analyze these messages to understand the progress of their students on the fly. As a result, adaptive teaching for a large class is handicapped. The main contribution of their work illustrates a novel concept map generation mechanism which is underpinned by a fuzzy domain ontology extraction algorithm. The proposed mechanism can automatically construct concept maps based on the messages posted to online discussion forums. Initial experimental results [3] reveal that the accuracy and the quality of the automatically generated concept maps are promising. The research work opens the door to the development and application of intelligent software tools to enhance e-Learning. To our best knowledge, the work presented in this paper demonstrates the first application of fuzzy domain ontology extraction method to facilitate adaptive e-Learning.

Typical pseudo-relevance feedback methods [7] assume the top retrieved documents are relevant and use this pseudo-relevant document to expand terms. The initial retrieval set can however, contain a great deal of noise. In this paper, a cluster based resampling method is discussed to select better pseudo-relevant documents based on the relevance model. The main idea is to use document clusters to find dominant documents for the initial retrieval set, and to repeatedly feed the documents to emphasize the core topics of a query. Experimental results on large-scale web TREC collections show significant improvements over the relevance model. For justification of the resampling approach, we examine relevance density of feedback documents. A higher relevance density will result in greater retrieval accuracy, ultimately approaching true relevance feedback. The resampling approach shows higher relevance density than the baseline relevance model on all collections, resulting in better retrieval accuracy in pseudo-relevance feedback. This result indicates that the proposed method is effective for pseudo-relevance feedback.

The usefulness of association rules is strongly limited by the huge amount of delivered rules. To overcome this drawback, several methods were proposed in the literature such as item set concise representations, redundancy reduction, and post processing. However, being generally based on statistical information, most of these methods do not guarantee that the extracted rules are interesting for the user. Furthermore, an interactive framework is designed to assist the user throughout the analyzing task [2].

In [1] Anne H.H. Ngu, Michael P. Carlson, Quan Z. Sheng, and Hye-young Paik proposed a novel application of semantic annotation together with the standard semantic web matching algorithm for finding sets of functionally equivalent components out of a large set of available non Web service based components. Once such a set is identified, user can drag and drop the most suitable component into an Eclipse based composition canvas. After a set of components has been selected in such a way, they can be connected by data-flow arcs, thus forming an integrated, composite application without any low level programming and integration efforts. One limitation with using SAWSDL or any other annotation techniques is that component must be annotated a-priori.

Ljiljana Stojanavic in his "e-learning based on semantic web" described an approach for implementing the e-learning scenario using semantic web technologies. It is primarily based on ontology-based description of content, context and structure of the learning materials and benefits the providing of and accessing to the learning materials [8].

Yanyan Li and Ronghuai Huang describes one solution to the problem of how to collect, organize and select web learning resources into a coherent, focused organization for instruction to address learners immediate and focused learning needs in "semantic based thematic search for personalized e-Iearning" [25].

Lora Aroyo and Darina Dicheva outline the state-of-the-art research on semantic e- learning and suggest a way towards the educational semantic web. They propose a modular semantic driven and service-based interoperability framework and related ontology-driven authoring tools in "The New Challenges for E-Iearning: The Educational Semantic web" [9].

III. MOTIVATION OF SEMANTIC SEARCH

The semantic search highly improves search accuracy of the query related data and the search engine delivers the exact content, the user intent to know. There’s no denying the power and popularity of the Google search engine. But there are other ways to search the web, using semantic search engines. By using semantic search engine we will ensure that it results in more relevant and smart results. For carrying out the exact search, word sense disambiguates could be used. This process involves the use of other information present in a semantic analysis system. And also it references takes help of other words presenting in the sentence and in the rest of the text. This paper discusses with proposed techniques available for semantic search and also tries to put focus on comparative analysis of them.

IV. CONTENT BASED SEMANTIC SEARCH

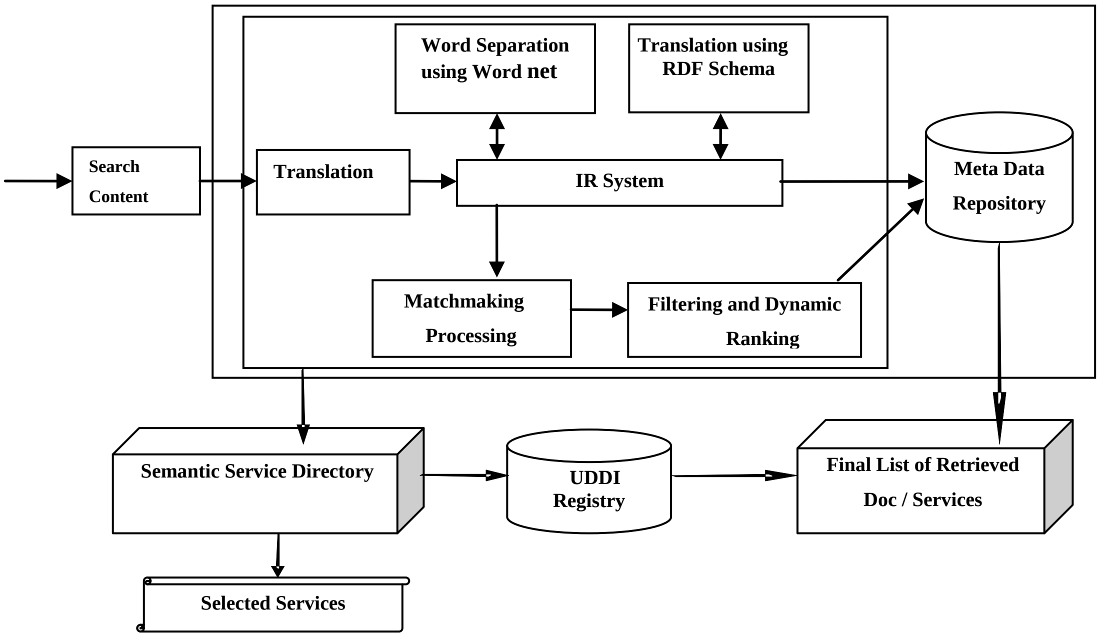

Word Separation using Word net
Translation using RDF Schema
Search Content
Translation
IR System
Meta Data Repository
Matchmaking Processing
Filtering and Dynamic Ranking
Semantic Service Directory
UDDI Registry
Final List of Retrieved Doc / Services
Selected Services

Fig.1. Architecture of Content based search

The architecture shown in Fig.1 consists of several modules which plays a unique role to refine the given text content. The user search text will pass through the following refinement processes to extract the annotations (nouns) that are used to discover the services.

- Noise Remover – To remove the noise, if any, present in the text document.
- Tokenizer – To separate each word and make as tokens.
- Filter – To filter and remove the “stop words”, if any, present in the document.
- WordNet – The lexical database to find the meaning for the words.

A. *Refinement Process*

The user given text document may contain several special characters like comma (,), dot (.), plus (+) etc. These special characters are, indeed, not useful for the retrieval of relevant information or service discovery. So these types of words are considered as a noise and these are to be removed from the content. The noise remover module removes this type of noise, if any, present in the text content. Once the noise is removed from the text document, the remaining words are to be tokenized. The tokenizer will analyze the noise free content and separate each word as a token. The result of the tokenizer is a set of words which may also contain “stop words”. Stop words are nothing but verbs in natural language, such as “want”, “which”, “is”, “what” etc. These stop words are also not useful for the retrieval of information or service discovery. So these types of stop words are also to be removed from the tokenized content. The tokenized content is then given into the WordNet and RDF to infer the semantic meaning to determine a more precise degree of similarity. The noise free tokenized content is fed to the filter module to filter the “stop words” and the filter removes all the “stop words” present in the content.

B. *Matchmaking Process*

Matchmaking is a process of finding the service provider that satisfies the server requester’s requests. Matchmaking is executed based on whether the web service request and web service provider match occurs or not. The match between requests and provider is determined based on whether the service input and output among the functional description match exists or not. [12].

The matchmaking framework includes a resource specification component, a request specification component, and matchmaking algorithms. A request specification includes a matchmaking function and possibly two additional constraints, a cardinality threshold and a matching degree threshold. The cardinality threshold specifies how many resources are expected to be returned by the matchmaking service. The matching degree threshold specifies the least matching degree of one of resources returned by the service. The matchmaking service executes a matchmaking algorithm for each request sent by the requester. The input of the algorithm is the request and the resource instances stored in the repository of the matchmaking service. The matchmaking algorithm evaluates the request function in the context of each resource instance in the repository.

The output of the algorithm is a number of relevant resources ranked according to their matching degrees. Let *n* denote the cardinality threshold specified by the request. The matchmaking algorithm returns the relevant resources that have the *n* largest matching degrees to the requester.

The matchmaking service executes a matchmaking algorithm for each request sent by the requester. The input of the algorithm is the request and the grid resource instances stored in the knowledge base of the matchmaking service. The matchmaking algorithm evaluates the request function in the context of each resource instance in the knowledge base.

The reason behind of using match making algorithm is to get accurate similarity value and therefore increases relevance in the search results. For each content-based query or published web service, similarity checking algorithm computes the similarity between inputs of such web services and the inputs of desired web service in user's request. Finally, computes the similarity of services according to the threshold value as shown in Fig. 3. The proposed Modified Matchmaking algorithm with dynamic ranking shown in Fig. 2 improves the overall search and produces the most relevant services or query at the top of the page.

```
Modified Matchmaking
Input request req, a finite set f resource instances rs
Output a finite set of candidate resource instance cs
begin
            $cs = \phi$
            n=req.CardinalityThreshold
            m=req.MatchingDegreeThreshold
            for each resource r in rs from a random beginning position
                md=evaluate req.RequestFunction in the context of r
                if $\left((md > 0) \text{ AND } (md \geq m)\right)$
                    add r into cs
                end if
                if ( the size of the candidate set > k * n )
                    break
                end if
            end for
            sort items in cs according to their matching degrees
            keep the items in cs that have the highest n matching degrees and remove the rest
end
```

Fig 2. Modified Matchmaking Algorithm

```
Similarity Checking
   input $\sigma_q$ ;  //receive the query
   input thsim; //threshold similarity
   Let $\sum_{\sigma_q} = \emptyset$ ; //initialize the result
   for (p=0;p< |∑|;p++)
     sim = $fsim(\sigma_q , \sigma_p)$;
     if(sim>=thsim) then
      add($\sum_{\sigma_q}$ , $\sigma_p$ );
     end if;
   end for;
   Output $\sum_{\sigma_q}$
```

Fig.3. Similarity Checking

A matchmaking service maintains a relevant knowledge base with a large number of resource instances. Performing an exhaustive matchmaking involving all resources in the knowledge base is very expensive for large knowledge bases. In a modified matchmaking algorithm shown in Fig. 2, the algorithm finishes searching the knowledge base when *k***n* (where *k* is a constant) resources are found with the required matching degrees (not less than the matching degree threshold).

Pseudo Code for Service Selection using Matchmaking and Re-ranking

```
Input: Request for Web service
Output: Composed Service Results
User requests a desired service from Database or performs Crawling;
If Crawling is requested
    For each input
            Request looks in to Appropriate Search Engines through search engine API.
            Matchmaking, re-ranking process takes place and engine produces results.
            Results are parsed to human readable format.
            Only service name and related wsdl links are extracted from the results.
            Quality of Service for a search is performed
             Search results are displayed
Results are stored in backup database.
    If no result is found for user's query word
        Message dialogue is displayed to enter synonym query word, or to scale the engine to more links
User selects a service from list;
For each Selected Service
```

Fig.4. Service Selection using Matchmaking and Re-ranking

V. Experimental Results

An experiment is conducted by creating proposed browser to search the content through search engines like yahoo, google and proposed search engine in order to measure the performance of the proposed architecture. The result shows the most relevant contents are at the top of the web page for the requested query and calculates the time interval between request and response for the search query. Meanwhile, a comparative analysis has been conducted with popular search engines like yahoo and google and proves that the proposed search mechanism produces the most appropriate results Fig 5.

VI. Performance Evaluation

In the real scenario, it is impossible to get all retrieved documents that can be considered as totally relevant, means a fraction of retrieved documents does not seem to be relevant to the users. Also, on the other hand it is also impossible to retrieve all set of documents that can be relevant for the users. Thus, depending upon the relevancy and precision required by the users, performance evaluation parameters such as precision, recall has been widely used.

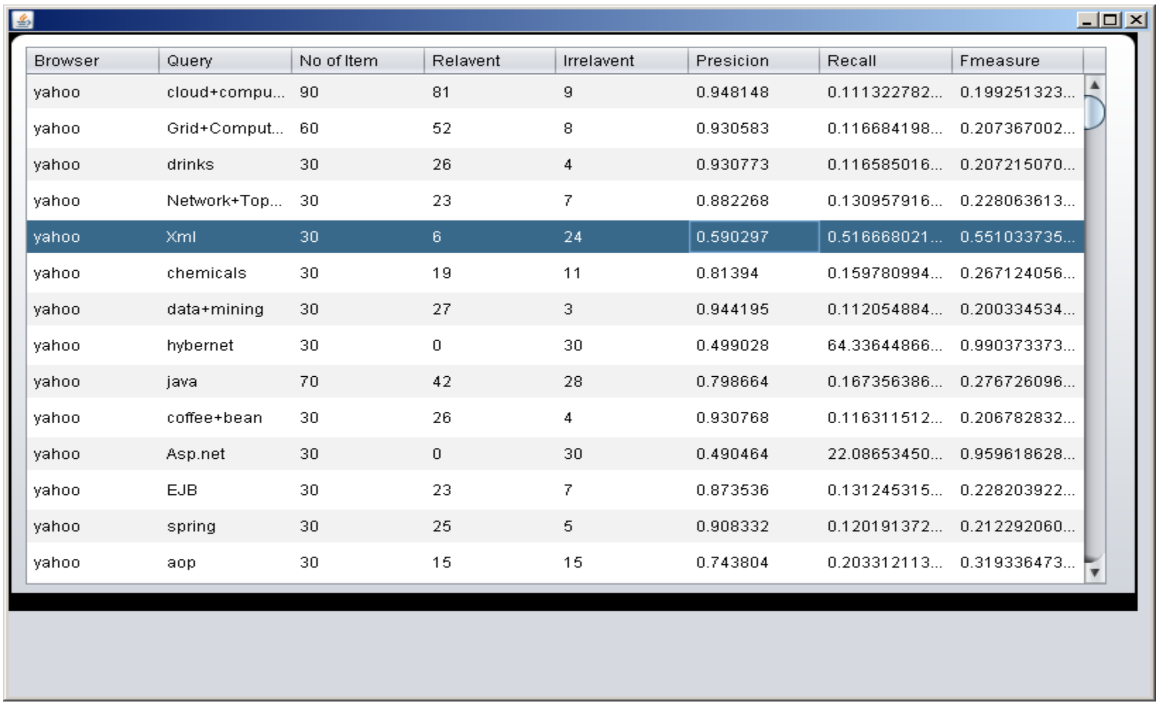
The following sets of notations are required to compute the precision, recall and F-measures.

retD	: Set of retrieved documents;
_retD	: Set of non-retrieved documents;
relD	: Set of relevant documents;
_relD	: Set of non-relevant documents;

A. Precision

Precision is one the most commonly used metrics for information retrieval. It basically measures how precisely the system picks the related documents among all documents. More specifically, it is the proportion of the related documents in the retrieved documents to the total number of retrieved documents.

$$P = \frac{|ret_d \cap \ rel_d|}{|ret_d|}$$



Browser	Query	No of Item	Relavent	Irrelavent	Presicion	Recall	Fmeasure
yahoo	cloud+compu...	90	81	9	0.948148	0.111322782...	0.199251323...
yahoo	Grid+Comput...	60	52	8	0.930583	0.116684198...	0.207367002...
yahoo	drinks	30	26	4	0.930773	0.116585016...	0.207215070...
yahoo	Network+Top...	30	23	7	0.882268	0.130957916...	0.228063613...
yahoo	Xml	30	6	24	0.590297	0.516668021...	0.551033735...
yahoo	chemicals	30	19	11	0.81394	0.159780994...	0.267124056...
yahoo	data+mining	30	27	3	0.944195	0.112054884...	0.200334534...
yahoo	hybernet	30	0	30	0.499028	64.33644866...	0.990373373...
yahoo	java	70	42	28	0.798664	0.167356386...	0.276726096...
yahoo	coffee+bean	30	26	4	0.930768	0.116311512...	0.206782832...
yahoo	Asp.net	30	0	30	0.490464	22.08653450...	0.959618628...
yahoo	EJB	30	23	7	0.873536	0.131245315...	0.228203922...
yahoo	spring	30	25	5	0.908332	0.120191372...	0.212292060...
yahoo	aop	30	15	15	0.743804	0.203312113...	0.319336473...

Fig.5. Retrieval of resources from google, yahoo search engines and proposed search using semantic approach

The Precision is calculated for google, yahoo and proposed search engines. The result shows that precision value is comparatively more in the proposed search mechanism than the other two popular search engines Fig.6. It is also represented in the form of graph.

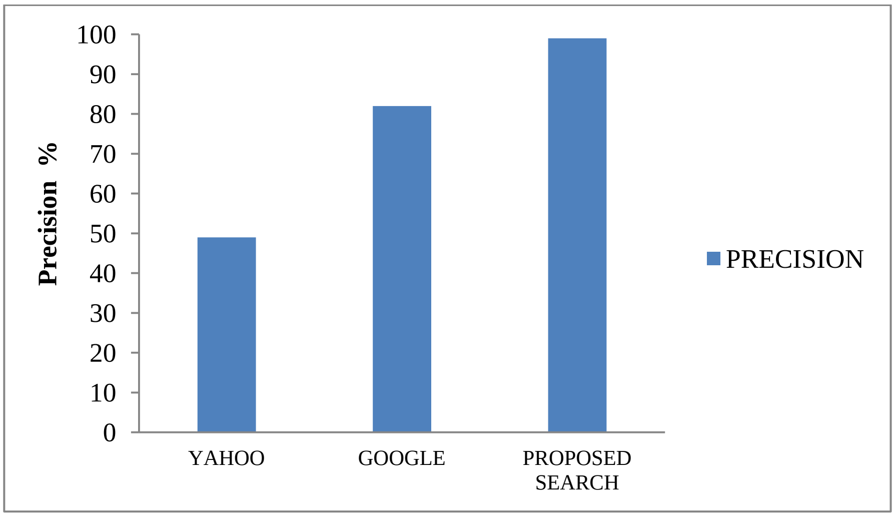


Fig.6. Precision Graph for Yahoo, Google Search Engines and Proposed Search

B. Recall

Recall is another widely used information retrieval metric. It is the proportion of the retrieved related documents to the total number of related documents that should have been retrieved. Similar to precision, it is not much meaningful on its own, because it does not takes into account the unrelated documents retrieved.

$$R = \frac{|ret_d \cap \ rel_d|}{|rel_d|}$$

The Recall calculated here is the relevant recall in which performance is compared in relevance to the google, yahoo and proposed search engines. When the proposed system is tested, there is a significant improvement in performance is observed for most of the queries as a result of semantic analysis, although a small fraction of them had negative and similar performance with a generic search engine. The recall values of the retrieved relevant documents are shown in the graph Fig. 7.

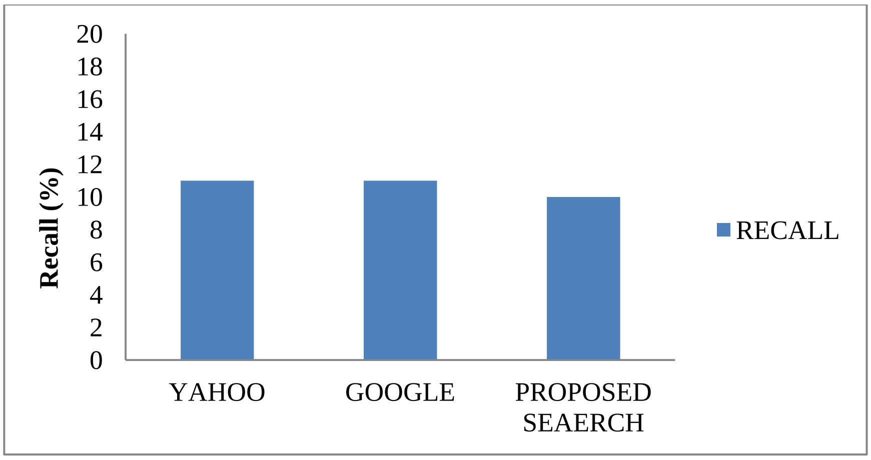


Fig.7. Recall Graph for Yahoo, Google Search Engines and Proposed Search

A comparative analysis is carried out for popular search engines such as yahoo, google and proposed search mechanism. The resultant graph proves that precision and recall are inversely proportional. When the precision value for a search increases, recall value decreases and vice-versa as shown in Fig.8.

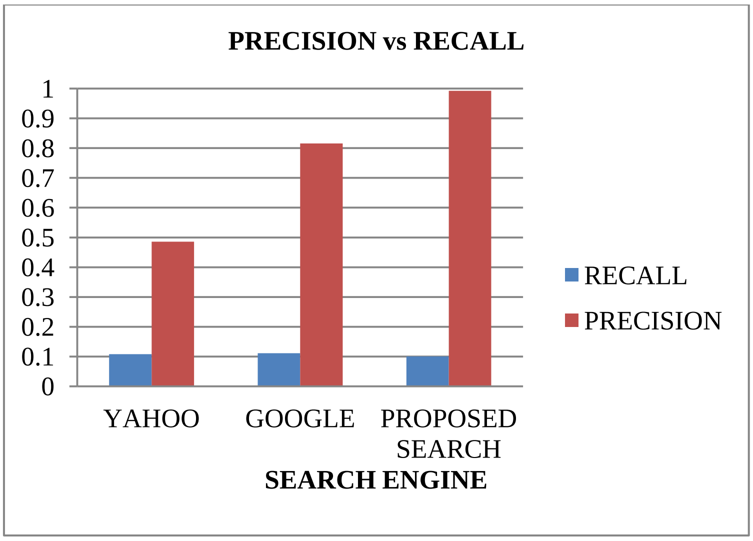


Fig.8. Precision VS Recall

VII. CONCLUSION

This paper has described the ontology based resource description language framework to retrieve the relevant contents as well as web services efficiently from the registry using modified match making algorithm. The Modified Matchmaking algorithm that takes as input a service request (or query) and an ontology of services and finds a set of services whose descriptions contain as much common information with request (or query) as possible and as little extra information with respect to request (or query) as possible. Thus, we have presented a ontology based architectural framework and semantic retrieval mechanism which includes all the aspects of Semantic Web, namely, ontology development, information extraction, dynamic ranking and performance metrics such as precision and recall. We compared our findings with other popular search engines and we observed that the proposed system provides better performance than the existing search engines. In the future, we extend the work based on user preferences and Quality of services to provide the exact content or services.

REFERENCES

[1] Anne H. H. Ngu, Michael Pierre Carlson, Quan Z. Sheng, Hye-young Paik: Semantic-Based Mashup of Composite Applications. IEEE T. Services Computing 3(1): 2-15 (2010)

[2] Claudia Marinica and Fabrice Guillet-Knowledge-Based Interactive Post mining of Association Rules Using Ontologies ,JUNE 2010

[3] Dijkman, R., Dumas, M., Garcia-Banuelos, L.: Graph Matching Algorithms for Business Process Model Similarity Search. 5701 (2009) 48–63

[4] S. Gauch, J. Chaffee, and A. Pretschner, ―Ontology-Based Personalized Search and Browsing, Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp. 219-234, 2003.

[5] Jon Lathem, Karthik Gomadam and Amit P. Sheth," SA-REST and (S)mashups : Adding Semantics to RESTful Services," In International conference on semantic computing, 2007, pp. 469 – 476.

[6] J.D. King, Y. Li, X. Tao, and R. Nayak, ‚Mining World Knowledge for Analysis of Search Engine Contenť,Web Intelligence and Agent Sys-tems,vol.5,no.3,pp.233-253,2007

[7] Kyung Soon Lee W. Bruce Croft James Allan- A Cluster-Based Resampling Method for Pseudo-Relevance Feedback, 2011

[8] Ljiljana Stojanovic, Steffen Staab and Rudi Studer "E-Learning based on the Semantic Web"

[9] Lora Arayo and Darina Dicheva, (2004) The New Challenges for E-learning: The Educational Semantic Web. Educational Technology & Society, 7 (4), 59-69.

[10] M. Lennon, D. Pierce, B. Tarry and P. Willett, “An Evaluation of Some Conflation Algorithms for Information Retrieval” J.Information Science, vol. 8, no. 3, pp. 99-105, 1988.

[11] R.Navigli, P.Velardi, and A.Gangemi, ―Ontology Learning and Its Application to Automated Terminology Translation, IEEE Intelligent Systems, vol. 18, no. 1, pp. 22-31, Jan./Feb. 2003.

[12] Pathak, P., Gordon, M., & Fan, W. (2000). Effective information retrieval using genetic algorithm based matching function adaptation. In Proceedings of 33rd Hawaii international conference on system sciences-2000 (pp. 1–8).

[13] M. Paolucci et al. Semantic Matching of Web Service Capabilities. Springer Verlag, LNCS, International Semantic Web Conference, 2002.

[14] T. Pedersen, S. Patwardhan, and J. Michelizzi, “WordNet::Similarity— Measuring the Relatedness of Concepts,” Proc. Nat’l Conf. Artificial Intelligence, pp. 1024-1025, July 2004.

[15] Pradeep Gurunathan, Dr. Seethalakshmi Pandian and Balamohan Somasundaram,”Design and Composition of e-Learning and Research Resources using Service Oriented Architecture”International Journal of Computer Applications 12(5):19–25, December 2010, Published By Foundation of Computer Science

[16] Pradeep Gurunathan Dr. Seethalakshmi Pandian,” A New Tool for Web-based Educational System”, Fourth International Conference on Natural Computation, 978-0-7695-3304-9/08 $25.00 © 2008 IEEE, DOI 112.

[17] Raymond Lau, Dawei Song, Yuefeng Li, Terence Cheung and Jin-Xing Hao, “A Fuzzy Domain Ontology Extraction Method”.

[18] Rembert, A.J.: Comprehensive workflow mining. In: Proceedings of 44-th ACM-SE, USA, ACM (2006) 222–227

[19] N. Seco, T. Veale, and J. Hayes, “An Intrinsic Information Content Metric for Semantic Similarity in Wordnet,” Proc. European Conf. Artificial Intelligence (ECAI ’04), pp. 1089-1090, Aug. 2004.

[20] P. Shvaiko and J. Euzenat, “A Survey of Schema-based Matching Approaches,” Journal on Data Semantics, Vol. IV, pp. 146-171, 2005.

[21] Trajkova and S. Gauch, ―Improving Ontology-Based User Profiles, Proc. Conf. Recherche d„Information Assistee par Ordinateur (RIAO „04), pp. 380

[22] Van der Aalst, W., Vandongen, B., Herbst, J., Maruster, L., Schimm, G., Weijters, a.: Workflow mining: A survey of issues and approaches. Data & Knowledge Engineering 47 (2003) 237–267

[23] Xuanzhe Liu, Yi Hui Wei Sun, Haiqi Liang,“ Towards Service Composition Based on Mashup”IEEE Conference paper., 2007, pp.332- 339.

[24] Yanyan Li and Ronghuai Huang"Semantic-Based Thematic Search for Personalized E-Learning" ,Publisher Springer Berlin / Heidelberg ,pages 354- 357.

[25] Yanyan Li, Mingkai Dong, "Towards a Knowledge Portal for ELearning based on semantic Web", Eighth IEEE International Conference on Advanced Learning Technologies, 2008, 217

AUTHOR(S) BIOGRAPHY

1. Pradeep Gurunathan received the Master Degree in 1998 and obtained Master of Technology in Information Technology in Manonmaniam Sundaranar University in 2004. Currently he is working as a professor in the Department of Information Technology, A.V.C College of Engineering, Mannampandal, Mayiladuthurai, Tamilnadu, India.

2. Dr. Seethalakshmi Pandian received her B.E. degree in Electronics and Communication Engineering in 1991 and M.E. degree in Applied Electronics in 1995 from Bharathiar University, India. She received the Doctorate in Information and Communication Engineering in the year 2004 from Anna University, Chennai. Her Area of interest includes Distributed Systems, Multimedia Databases and currently, she is working in the Department of Electronics and Communication Engineering, University College of Engineering, BIT Campus, Tiruchirappalli Tamilnadu, India.