INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Information Retrieval Web Service in a Distributed Environment Using Ontology - a Novel Approach to Semantic Web

S. Meenakshi
Associate Professor, RMK Engineering College, Tamil Nadu, India
Dr. R. M. Suresh
Principal, Chennai Institute of Technology, Tamil Nadu, India

Abstract:
With the rising trend in Information Technology, Web Information Retrieval is picking up its vitality step by step. As have experience in making use of well-known search engines consistently, the Search Engine Results Page (SERP) returned is truly excessively large and nearly irrelevant. Need to continuously pass on to the "next page" to acquire the web pages which the users really need. The reason is that, when the user wants to search information in the Web, the search engine abstracts the information to the keyword combination and afterward submits it. The relationship between keywords is clear to the users but not to the search engines. The Semantic Web is an evolution of the current Web that represents information in a machine-readable format, while maintaining the human-friendly mark up language representation and thereby avoiding key word searching. This research work propose a Web Service Architecture to cater the need of distributed environment whereby the link content and page content of the SERP are checked for the given user query keywords, so that the more relevant pages are retrieved and then discusses key techniques about extracting domain concepts, interrelationships between concepts-keywords and automatically constructing classification system in ontology learning using semantic technologies such as annotation, RDF and SPARQL queries.

Keywords: *Semantic web search, domain concept ontology, semantic annotation, relation keyword.*

1. Introduction

Having substantial measure of information on world wide web, people suffer from spending much time and effort to analyze the results given by a search engine. This is principally because of searching without considering the users' preference and returns extensive measure of irrelevant and inaccurate results. If the Web page only includes the keywords and there is no relationship between keywords in the content of the Web page, the Web page does not give what the user needs and normally an incorrect and inadequate page.

In Fact keyword-based search engines make little provision for the formulation of very specific queries, especially those that make utilization of relationships between domain concepts and keywords. The proposed work permits semantic based search in domain situations. It incorporates two fundamental components:

1) A general framework for resource annotation focused around ontological model of the domain.

2) A user-friendly search interface that permits the formulation and execution of knowledge-based queries over the generated data.

Semantic web augments the current web by including facilities for machine - processable descriptions of meaning. In order for semantic exchanges of information to happen there requirements to be agreement on how to model meaning. Ontologies are the mechanism for representing formal and shared domain descriptions. Keeping in mind the end goal to support ontologies on the semantic, web standards need to be developed to help interoperability between machines.

There have been many standards for semantic web defined by W3C are RDF and OWL. RDF is to document a language for representing information about the resources on the web. OWL is to let applications process the content of information. The query interface in semantic search only helps formal queries viz. SPARQL. The proposed method utilizes all the above concepts to enhance the efficiency of the search.

1.1. Structure of the Paper

The rest of the paper is composed as follows. Section 2 portrays related work in both the IR and SW areas, and addresses a common understanding of what semantic search is, and where are standing in the progress towards semantic Information Retrieval. Section 3

portrays methodology of the proposed work. Section 4 portrays architectural design of the proposed work. Section 5 portrays algorithm for concept relevancy of ranking web pages. Experimental results and performance evaluation is reported in Section 6 and Conclusion and future enhancements are exhibited in Section 7.

2. Related Works

Having huge distributed environment of web resources-searching and retrieving the significant information is still a challenge facing research. Moreover, the semi-structured and unstructured nature of web resources makes the need for web content information retrieval.

In Paper [1,8], the author separates web content mining from two different perspective; Information Retrieval and Database and in paper [2] with the ranking algorithm best suitable for efficient semantic search in actuality Techniques for Effectively Searching and Retrieving Information from Internet are discussed. In Paper [3] an assortment of issues of identifying content such as a sequence labelling problem, a common problem structure in machine learning and natural language processing is recognized. Extracting Information from Web Data Sources in extension of Semi-Structured are tagged in [4]. A new method for relevance ranking of web pages with respect to given query was dead set in paper[5]. In Paper [6] a survey of web content mining plays as an prominent tool in extracting structured and semi structured data and mining them into useful knowledge is exhibited. A framework is proposed to give facilities to the user during search[7]. Design specification on SWISE: Semantic Web based Intelligent Search Engine introduced in [9].

In Paper [10] an approach for Identifying URLs Based on Division Score and Link Score in Focused Crawler being discussed. Work moves from[11], the consideration above and relies on the assumption need to know the structure of the underlying ontology and of the Web page to be ranked in order to compute the corresponding relevance score. Enhancing the quality of search results by eliminating web outliers using chi-square [13]. In Paper[14] a brief Survey in term of techniques focused around semantic search engine depicted. For web content mining, relevance ranking and evaluation of search results through experimental study in[16]. Semantic searching IT careers concepts focused on ontology is thorough portrayed in [17].

3. Methodology

In the Semantic Web, Ontological models permit the *annotation* of Web documents (modelling 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.

In this approach 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.

3.1. SERP Extraction

Based on the user query, Search Engine Results Page (SERP) are retrieved. Pre-process both User Query and Search Engine Results Pages individually based on domain ontology and semantic annotation.

3.2. Preprocessing

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.

3.3. Link content and Page Content Determination

Pre-process and extract the link content 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.

3.4. Relevancy Calculation

The relevancy is calculated based on how well the results matches the query plus how related the retrieved index items of the results to the query.

After finding the web pages on the proposed approach 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. The search result page's total relevancy are ranked in increasing order.

3.5. Re-ranking

Finally re-rank the search results on Total relevancy in increasing order. The Top Search Result is the most relevant and bottom is the least relevant for the User query.

4. Architecture

Architecture of the proposed work takes the advantage of domain ontology and semantic annotation against Repository.

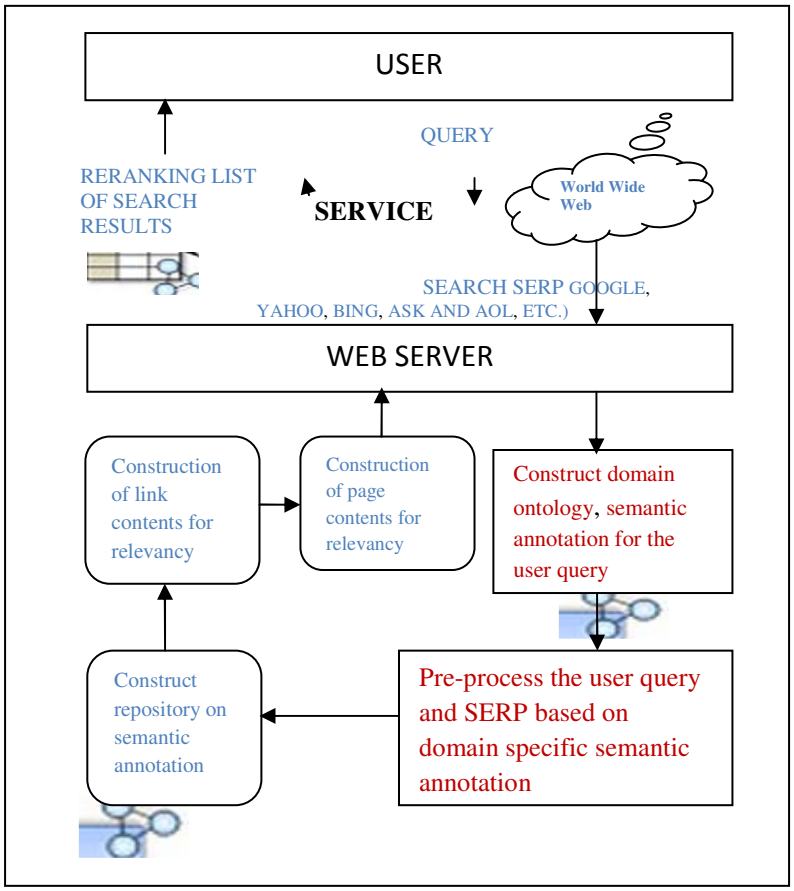

Figure 1: Architecture design

The Fig.1. describes a high level overview of the components used in semantic search architectural design. This structure of the main components, according to, includes: SERP extraction, pre-processing, link cont and page content determination, relevancy calculation and re-ranking. In this figure, documents have their own indexes which mean that they have semantic structures, and queries have reference to the concepts and their relationships. Also the retrieval documents have semantic concept. Moreover, the re-ranking can be improved on refinement of the results.

In this architecture, which is proposed, there is a possibility to have similar search processes at semantic layers as well. User queries are pre-processed which means they have semantic structure, and therefore every user query is understandable by the computer. However, the queries, which are used for searching documents, will be understandable and meaningful by the computer. Also, the relationship between key words and concepts is considered in this structure. The communication between every user query and their relationships are based on ontology and semantic annotations, which are used to facilitate this possibility for the semantic web.

5. Algorithm for Concept Relevancy Ranking of Link and Page Content

Input : Extracted Search Engine Results Page

Output : Re-ranked Search Engine Results Page

- Step 1:/* In our example, the user query is "company cts Chennai taramani "*/

For the user query extract SERP's of Top-K results.

- Step 2: /* pre-process based on domain ontology and semantic annotation.*/

pre-process the user query and SERP's based on domain ontology and semantic annotation .

- Step 3:/*construction of repository*/

Pre-process the user query and Extract root words RW and construct a repository without duplications of root words RW.

- Step 4: /*link content computation*/

Pre-process and extract the link contents words for the SERP's and calculate link content keyword strength. Compare each link keywords against Repository. if match found grant the keyword strength to the specific link content keyword Else grant 0. Calculate Total Strength for link content Keyword by summarizing strength of all link content keywords.

- Step 5: */page content computation/*

Pre-process and extract the page contents words for the SERP's and calculate page content keyword strength. Compare each page content keyword against Repository. if match found grant the keyword strength to the specific page content keyword Else grant 0. Calculate Total Strength for page content Keyword by summarizing the strength of all page content keywords.

- Step 6: Compute total relevancy for the particular SERP using damping factor d..

TR_i = total strength of link content keywords * d + total strength of page content keywords * (1- d)

- Step 7 Repeat the Step 4 through 6 for all SERP's
- Step 8 Rerank the result based on TR in increasing order.

The Topmost Search Result SRi is the most relevant and bottom most search result is the least relevant for the User query whereby display the retrieved documents according to the re-rank.

6. Experimental Results and Performance Evaluation

Since there is no standards metrics to measure the quality of ranking ontologies or instances in the semantic at present; evaluate the accuracy of our proposed ranking; in comparison with search engine ranking and procedure based manual ranking.

Now we compare the rankings of the various search engines Google, Yahoo, Bing, Ask, AOL on the domain specific user query ("company cts Chennai taramani") on the same day – Table 1. As examining all the results in Table 1, out of which search engine ranking Google Vs proposed and procedure based manual ranking give more closure results.

SERP ID	PROPOSED RANKING APPROACH ON					PROCEDURE BASED MANUAL RANKING
	YAHOO	BING	ASK	AOL	GOOGLE	
SERP1	9	6	9	3	10	10
SERP2	6	4	5	10	5	4
SERP3	5	3	6	6	4	5
SERP4	4	10	4	9	2	2
SERP5	1	9	2	7	1	1
SERP6	2	1	1	8	9	9
SERP7	8	8	8	5	8	7
SERP8	10	7	7	4	3	3
SERP9	3	2	3	1	6	6
SERP10	7	5	10	2	7	8
Contd ●●●						

Table 1: Comparison of Multiple Search Engines – Relevancy Ranking

In this Section experimentally evaluate the proposed methodology. The goal of the experiments are : (1) sample Dataset is consider for evaluation purpose and Top-k search engine results that are more relevant to the user – TABLE 2; (2) to compare the performance of our ranking with search engine ranking - TABLE 3; and (3) to compare the search engine ranking with proposed and procedure based manual ranking to demonstrate the effect of the proposed methodology-TABLE 4.

Here in detail will summarize the methodology adopted and the results obtained for a domain specific user query ("company cts Chennai taramani") against specific search-engine namely Google; whereby Top-k web pages from that search-engine are taken as an input dataset as in Table 2.

Did you mean: company cts chennai ***taramani***

SERP ID — Search Engine Results Pages

SERP1 **Cognizant Technology Solutions** India Private Limited, **Taramani ...**
*www.asklaila.com › **Chennai › IT Companies***
IT **Companies**, Airtel Payment Dropbox: **Cognizant Technology Solutions** India Private Limited, **Taramani**, **Chennai**, Tamil Nadu – Get contact address, mobile **...**

SERP2 **Cognizant** in Jobs, recruitment in **Taramani**, Tamil Nadu | Indeed.**co**.in
*www.indeed.**co**.in/ **Cognizant** -in-jobs-in-**Taramani**,-Tamil-Nadu*
Jobs 1 - 10 of 38 – 38 **Cognizant** in Jobs available in **Taramani**, Tamil Nadu on Indeed.com. one search. all **... Cognizant** IN 340 reviews - **Chennai**, Tamil Nadu **...**

SERP3 **Cognizant Technology Solutions** Jobs, recruitment in **Taramani ...**
*www.indeed.**co**.in/**Cognizant**-Technology-Solutions-jobs-in-**Taraman**...*
Jobs 1 - 10 of 31 – 31 **Cognizant Technology Solutions** Jobs available in **Taramani**, **... Cognizant** IN 340 reviews - **Chennai**, Tamil Nadu **...** Sr. **Business** Analyst **...**

SERP4 **Cognizant Technology Solutions**
www.**cognizant**.com/contactus/office-**locations**
Score: **24 /** 30 · 22 Google reviews

SERP5 **Cognizant Technology Solutions**
*www.**cognizant.**com/*

SERP6 Cognizant Technology Solutions Ltd. in Tharamani ...
yellowpages.sulekha.com › ... › Software Companies in Tharamani
Cognizant Technology Solutions Ltd. in **Tharamani**, **Chennai** - 600113 – Get **Cognizant Technology** ... Fill this Form and Software **Companies** will call you now.

SERP7 Cognizant in Jobs, recruitment in Taramani, Tamil Nadu ...
*www.indeed.**co**.in/**Cognizant**-in-jobs-in-**Taramani**,-Tamil-Nadu*
Jobs 1 - 10 of 50 - 50 **Cognizant** in Jobs available in **Taramani**, Tamil Nadu on Indeed.com. one search. all jobs. ... Advanced Job Search. job title, keywords or **company**, city or state ... **Cognizant** IN 2,095 reviews - **Chennai**, Tamil Nadu ...

SERP8 Cts Jobs, recruitment in Taramani, Tamil Nadu | Indeed.co.in
*www.indeed.**co**.in/**Cts**-jobs-in-**Taramani**,-Tamil-Nadu*
Jobs 1 - 10 of 53 - 53 **Cts** Jobs available in **Taramani**, Tamil Nadu on Indeed.com. one search. all jobs. ... **CTS**, SITEL, SUTHERLAND – **Chennai**, Tamil Nadu ...

SERP9 Cognizant Technology Solutions India Pvt Ltd in Tharamani ...
*www.justdial.com/**Chennai/Cognizant**...**Tharamani/** 044P7011372_Q2hlb...*
Rating: 4.3 - 172 votes
Cognizant Technology Solutions India Pvt Ltd in **Tharamani**, **Chennai** listed ... your friends rating SEBA BUISNESS SOLUTIONS ENCHANTER **CORPORATION** ...

SERP10Cognizant Technology Solutions India Pvt Ltd in Perungudi ...
*www.justdial.com/**Chennai/Cognizant**..**Tharamani**.../044PF005719_Q2h...*
No 1, Veeranam Road, Perungudi,**Chennai** - 600096 Opp To **Tharamani** Railway Station | Map ... We found the **company** to be good in payment and service.
Contd. ●●

Table 2: Input Data Set

Proposed algorithm - Concept relevancy ranking is applied to above SERP and results are listed in Table 3.

Sl. No.	SERP ID	TOTAL RELEVANCY	RANK
1	SERP5	0.1	1
2	SERP4	0.1	2
3	SERP8	2.35	3
4	SERP3	2.45	4
5	SERP2	2.45	5
6	SERP9	3.0	6
7	SERP10	3.0	7
8	SERP7	2.45	8
9	SERP6	2.45	9
10	SERP1	3.8	10
Contd. ●●●			

Table 3: Relevancy Ranking

From the TABLE 3 results, it is understood that if the total relevancy value is less, it is ranked as first and vice-verse.
Now the same relevant dataset is evaluated against retrieved dataset. Comparison results of the proposed approach against search engine ranking and procedure based manual ranking are given in the TABLE 4.

SERP ID	SEARCH ENGINE RANKING	PROCEDURE MANUAL RANKING	PROPOSEID RANKING APPROACH
SERP1	1	10	10
SERP2	2	4	5
SERP3	3	5	4
SERP4	4	2	2
SERP5	5	1	1
SERP6	6	9	9
SERP7	7	7	8
SERP8	8	3	3
SERP9	9	6	6
SERP10	10	8	7
Contd. ●●●			

Table 4: Comparison of – Relevancy Ranking

TABLE 4 represents the matching of procedure based manual ranking against proposed approach ranking. Document SERP2, SERP3 represents the mismatching of procedure based manual ranking against proposed approach. As can observe from the experimental results; proposed methodology outperforms existing ranking results.

7. Conclusion and Future Enhancements

Our system demonstrated that mature IR algorithms can be effectively turned into web services. The Proposed approach gives obviously better results compared with search-engine ranking. In evaluating the performance of the search system it is observed that by ontology-based annotations users could perform more accurate results while being returned up to maximum percent fewer results than with a keyword-based search engine in the best cases eliminating more percent of the irrelevant documents. Further research is going on as semantic web search still in its infant stage to refine theses deployments and are planning more industrial deployments in the near future.

8. Acknowledgment

I thank the journal publisher for their helpful comments and suggestions that greatly improve this work. Nevertheless, we express our gratitude toward our families and colleagues for their kind co-operation and encouragement which help us in completion of this work

9. References

i. Cheng Wang, Ying Liu, Liheng Jian, Peng Zhang, A Utility based Web Content Sensitivity Mining Approach, International Conference on Web Intelligent and Intelligent Agent Technology (WIIAT), IEEE/WIC/ACM 2008.

ii. Hongqi li, Zhuang Wu, Xiaogang Ji, Research on the techniques for Effectively Searching and Retrieving Information from Internet, International Symposium on Electronic Commerce and Security, IEEE 2008.

iii. Kshitija Pol, Nita Patil, Shreya Patankar, Chhaya Das, "A Survey on Web Content Mining and Extraction of Structured and Semi structured data", First International Conference on Emerging trends in Engineering and Technology, 2008.

iv. Mahmoud Shaker, Hamidah Ibrahim, Aida Mustapha, Lili Abdullah, "A Framework for Extracting Information from Semi-Structured Web Data Sources," iccit, vol. 1, pp.27-31, 2008 Third Intl. Conference on Convergence and Hybrid Information Technology, 2008

v. Tao Huang, Qingtang Liu, Sanya Liu, Shengming Wang, Yong Yang, (2008), Design and Implementation of Semantic Query System based on Ontology Context',Vol.3,pp 331-362.

vi. Shohreh Ajoudanian, and Mohammad Davarpanah Jazi," Deep Web Content Mining", World Academy of Science, Engineering and Technology,49 2009.

vii. Sungrim Kim and Joonhee Kwon, 'Information Retrieval using Context Information on the Web 2.0 Environment', IJCSNS International Journal of Computer Science and Network 62 Security, VOL.9 No.10, October 2009.

viii. Fabrizio Lamberti, Andrea Sanna, Claudio Demartini, "A Relation-Based Page Rank Algorithm for Semantic Web Search Engines," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 1,pp. 123-136, Jan. 2009

ix. FaizanShaikh, Usman Siddiqui.A, IramShahzadi, (2010), SWISE: Semantic Web based Intelligent Search Engine', Vol.10,pp 247-271

x. Debashis Hati, Amritesh Kumar, An Approach for Identifying URLs Based on Division Score and Link Score in Focused Crawler, Intl. Journal of Computer Applns., 2010

xi. G.Sudha Sadasivam; C. Kavitha; M. Saravana Priya; "Ontology based Information retrieval for E-Tourism" International Journal of Computer Science and Information Security Vol.8 No.2 May 2010.

xii. Aidan Hogan, Andreas Harth, JurgenUmbrich, Sheila Kinsella, Axel Polleres, Stefan Decker, (2011), _Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine', Vol.3, pp 365-401.

xiii. G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav, "Improving the quality of search results by eliminating web outliers using chisquare",Published in Lecture notes in CCIS – Springer, Vol. 202, pp.557-565, 2011.

xiv. G. Madhu, Dr. A.Govardhan, Dr.T.V. Rajinikanth, Intelligent Semantic Web Search Engines: A Brief Survey, International Journal of Web and Semantic Technology, 2011

xv. G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar G.V. Uma K. Sarukesi "Relevance ranking and evaluation of search results through web content mining", Proceedings of the Intl. multi conference of engineers and computer scientist IMECS Hong Kong Vol. 1 Mar 14-16 2012

xvi. Akshata D. Deore, Prof. R.L. Paikrao, Ranking Based Web Search Algorithms, International Journal of Scientific and Research Publications, vol 2, 2012

xvii. Paksakorn Singto; Asnirach Mingkhwan; "Semantic searching IT careers concepts based on ontology"; Journal of Advanced Management Science; Vol. 1; No.1; March 2013.