Improving the Efficiency of Semantic Web with Meta Crawler

S. Raja Ranganathan[1] , Prabakar D[2],Dr. M. Marikkannan[3] ,Dr. S. Karthik[4]

[1]*Assistant Professor, Dept of CSE, SNS College of Technology-Coimbatore.*

[2]*Assistant Professor, Dept of CSE, SNS College of Technology-Coimbatore.*

[3]*Assoicate Professor, Dept of CSE, Instuite of Road and Transport Technology-Erode.*

[4]*Professor and Dean, Dept of CSE, SNS College of Technology-Coimbatore.*

Abstract

The search engines play a major role for the people to collect resources they need. Still now most of the users using a single crawler search engines by ongoing development of web resources there are millions of web pages added by the programmers throughout the world daily hence the single crawler search engines efficiency is not sufficient. There are more search engines evolved with different search services,each with a unique interface and a database covering a different portion of the Web.As a result, users are forced to search repeatedly use their queries across different services. In some cases these services return many responses that are irrelevant, outdated, or unavailable, forcing the user to manually sift through the responses searching for useful information. In Most of the times the user retrieve result mislead the user towards the wrong area. In this paper we going to propose an enhanced Metacrawler for semantic web search engine. The Metacrawler provides a centralised interface for web resource search. This enhanced crawler help to provide access to users to search his query in multiple resources and get a most relevant and enhanced results.

Keywords: *MetaCrawler, WWW, World Wide Web, Semantic Web Search.*

1. Introduction

The Metacrawler is otherwise called as Metasearch engine it forms the combination of multiple crawlers search results.Most of the popular Web search and resource services such as Ask, Lycosand WebCrawler have proven most useful for user queries. As the Web resources grows, the number and variety of search services is increasing as well. Examples include: the Yahoo ``net directory''; the Harvest home page search service [7]. Since each service provides an incomplete snapshot of the Web, users are forced to try many times their queries across different search engine crawlers until they find appropriate and relevant responses. The process of querying multiple services is quite disturbing becauseeach service has its own unique methodology interface which the user is forced to learn. Further, the services return poor responses that are irrelevant, outdated, or unavailable, forcing the user to manually sift through the responses searching for useful information. Most cases blind links are the huge pain for the users who search for resources in World Wide Web.

The Metacrawler was created to solve most of the above mentioned problems outlined. Metacrawler is a Artificial intelligent based software robot that cumulates multiple web crawlers. While the users enter queries, and Metacrawler forwards those queries in parallel to the multiple search services. Metacrawler then collates the results and ranks them into a list of efficient resources, returning to the user the sum of knowledge from the best Web search services from the database. The key idea is that the Metacrawler allows the user to express *what* to search for and frees the user from having to remember *where* or *how*. Users can provide command to the the Metacrawler to find pages with either *all* of the words in their users query, *any* of the single particular words in their query, or all of the words in their query as a *complete sentence*.

Here we are going to enhance the performance of users query with the help of advanced Metacrawler and the crawler is implemented in Semantic web the next generation metadata based search strategy. Then results obtained are to be stored inside a knowledge database and finally the most efficient result for the

user query is extracted from knowledge database. The results are published to the user effectively.

2. Literature Survey

The extraction of information from World Wide Web is not a new mechanism but we have to face challenges in information retrieval in many ways. There is different kind of search engines available in World Wide Web each search engine follows a separate and unique mechanism of indexing and processes of search of its own so the information extraction as well as the result produced by these search engines are not the same. Some of the popular search engines such as GOOGLE, YAHOO, BING and ALTA VISTA produce results based on their uniqueness of crawlers after the keyword are processed. They only search information available on the web page, recently updated, some research group's such as OWL based SWOOGLE produces search results from their semantics based search engines, and however most of them are in their initial stages they face certain problems in matching ontology and combining keywords in RDF. The major problems facing by the search engines are they not able to gather content whole indexing in entire internet.

The Metacrawler along with the Semantic Web is a collaborative construction movement led by the World Wide Web Consortium (W3C) [1] that is privileged with common formats for the information that is available on the World Wide Web. The inclusion of semantic content in web pages, the Semantic Web aims at converting the current web of unstructured documents into a "web of Information". It builds on the W3C's Resource Description Framework (RDF).[4] According to the W3C, The Semantic Web provides a common framework that allows information to be shared and reused across application, enterprise, and community boundaries.[4] The term was coined by Tim Berners-Lee,[1] the inventor of the World Wide Web and director of the W3C, which oversees the development of proposed Semantic Web standards. He defines the Semantic Web as a web of data that can be processed directly and indirectly by machines. As with the WWW, the growth of the Semantic Web will be driven by applications that use it. Semantic search is an application of the Semantic Web to search. Search is both one of the most popular applications on the Web and an application with significant room for improvement. We believe that the addition of explicit semantics can improve search. Semantic Search attempts to augment and improve traditional search results (based on Information Retrieval technology) by using data from the Semantic Web. Traditional Information Retrieval (IR) technology is based almost purely on the occurrence of words in documents. Search engines like Google [4], augment this in the context of the Web with information about the hyperlink structure of the Web. The availability of large amounts of structured, machine understandable information about a wide range of objects on the Semantic Web offers some opportunities for improving on traditional search. Before getting into the details of how the Semantic Web can contribute to search, we need to distinguish between two very different kinds of searches.

3. Metasearch Architecture

Normal search services extracts, creates and stores an index of the Web as well as retrieve information from that index. Unlike these services, the Metacrawler is a *Multi-based service* which not uses its own database, it relies on other external search services(such as semantic web knowled database) to provide the information necessary to satisfy user queries. The insight here is that by separating the retrieval of pages from indexing and storing them, a lightweight application such as the Metacrawler can access multiple databases both semantic search database and other web search engine databases and thus provide a larger number of potentially higher quality references than any search service tied to a single database.

One of the main advantage of using Metacrawler is it does not depend upon the implementation or existence of any one search service.Some indexing mechanism is necessary for the Web. Typically, this is done using automated web robots or web spiders,which may not necessarily be the best choice [11]. However, the underlying architecture of the search services used by the Metacrawler is not that much important. As long as there is no central complete search service and several partial search services exist, the Metacrawler can provide the benefit of accessing them simultaneously and collating the results. The resources in the World Wide Web rapidly growing hence the need for semantic based Metacrawler also increased. The popularity of the Metacrawler also increasing among people. The below mentioned diagram shows the work flow of information in Metacrawler.

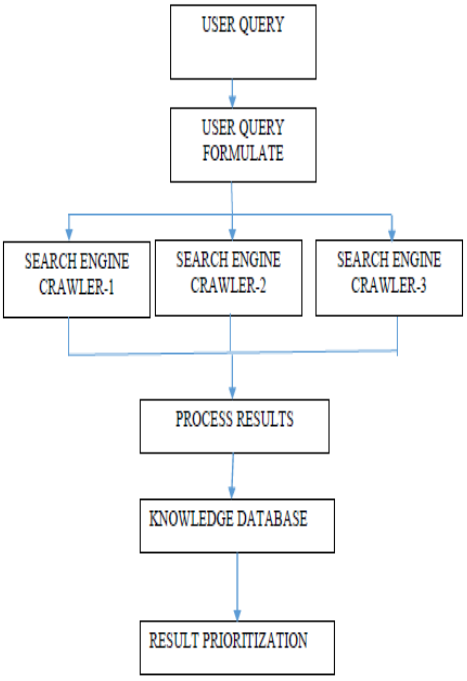

Figure-1 Information flow in Meta crawler

4. Interfacing Metacrawler in semantic web Search

The Semantic Web is able to describe things in a way which the computers can understand. This technology is refered to be as next generation compared with current searching methods [4] that provides the meaning of information in well-defined format that allows user to process the contents and retrieve the information in well understand manner. In semantic web the solution for the problems is effectively overcome by its architecture itself. One of the main components in semantic web is (RDF) Resource Description Framework a new standard of W3C the search efficiency has been improved by multiple combinations made for user's keywords the RDF looks subject, Predicate and Object for each statement the user intend to search. The RDF is purely an XML language and RDF enables exchange and reuse of structured metadata. The second important component in semantic web is Ontology [5] this helps to make the relation among the successful concepts. The ontologies use OWL web ontology language in different levels we can express they are OWL Lite, OWL DL and OWL Full ordering by increasing level. The Semantic Web will support more efficient routing, expertise decision, integration and reuse of data and provide support for interoperability problem which cannot be resolved with current web technologies

The single crawler is not sufficient for semantic annotated results so in order to improve the efficiency of search we going to search the content in Metacrawler. The results what extracted should be stored inside knowledge database finally prioritized and sent to results page for the user all the above tasks done within seconds. The ontology which is implemented inside knowledge database would have enormous resource of mapping thanks to the Metacrawler this only responsible for to retrieve millions of indexed contents.the efficiency has been proved in following table the comparison has been made for single crawler and multiple crawler and successful information'sretrieved under various circumstances.

Topics	Single Crawler	Multi Crawler
Home	4535	98654
India	3227	43234
crawler	1223	5432
Mobile Phone	7225	253421

Table-1: the comparison of successful count information between single crawler and Meta crawler for most relevant search topics in semantic web search

5. Performance of Services

- Web Coverage: How many hits will be returned on average result cumulated?
- Web Relevance: Are hits returned actually followed by users query?
- Web Performance: How much time taken to complete the users query process? Either it is successful or unsuccessful process of query?

A. Web Coverage

Measures the maximum number of hits required to process each service, the measurement is mainly based on the percentage references returned as well as some references exactly matches of users query where it returned them

Thus, 70% returned with 65% unique shows that on average a service returns 70% of its maximum allowed, with 65% of those hits being unique matching reference.

Web Search CRAWLER USED	Percent of Max Hits Average	Hits Returned Returned / Maximum Allowed
Open Text	80	8/10
LYCOS	76	15/17
WEB CRAWLER	70	21/25
GALAXY	68	12/16

Table-1 Percent of Maximum Hits Average and hits returned and allowed for various crawlers

The first column in the above table shows the percentage of the maximum hits allowed for each service returned.Each percentage was calculated by dividing the average hits returned by the maximum allowed for that service AVEREGA/MAXIMUM ALLOWED, as shown in the next column. This percentage is a measure of how many hits a service will provide given a pre-set maximum. The Metacrawler used different maximum values for services, as some had internal maximum values, and others would either accept only certain maximum values that produced in query processing services.

B. Web Relevance

for to calculate the web reverence the two methods were adopted the first one is service returned most references the users follows and the second one is the cumulative level of percentage obtained by each services The first is that the relevant information for people may be the list of references itself. For example, people who wish to see how many links there are to their home page may search on their own name just to calculate this number. The second process is that these numbers may be skewed by the number of hits returned by each service.

C.**Web Performance**

The final thing is to measure various performance of responce time of each services. Theresponce time vary time to time based on service load conditions. One explanation for the length of times taken by these services is that the majority of requests are during peak hours. Thus, results are naturally skewed towards the times when the services are most loaded. Times during non-peak hours are much lower. Hence the combination of all crawlers needed for the future web to enhance the performance.

6. Conclusion

Metacrawler presents users with a single artificialintelligent based interface that controls multiple powerful resources. The user's queries for each service it uses, collects the references obtained from those services, and optionally downloads those references to ensure availability and quality. It then removes duplicate references and collates the rest into a single list for the user. The user need know only *what* he or she is looking for the Metacrawler takes care of *how* and *where*. We have paid special attention to performance, making it a practical tool for Web searching.The semantic web technology is future web search methodology the enables the user to extract most precise information. By adding Metacrawler to semantic web enables the user to access most efficient interface that extracts also selects efficient precise data to the user finally the results are prioritized with page ranking strategy and most reliable results were published to the user.

10. REFERENCES

[1] Berners-Lee, T., Hendler, J. and Lassila, O. “The Semantic Web”, ScientificAmerican, May 2001.

[2] Deborah L. McGuinness. “Ontologies Come of Age”. In Dieter Fensel, J imHendler, Henry Lieberman, and Wolfgang Wahlster, editors. Spinning the Semantic Web:

Bringing the World Wide Web to Its Full Potential. MIT Press, 2002.

[3] Ramprakash et al "Role of Search Engines in Intelligent Information Retrieval on Web", Proceedings of the 2nd National Conference; INDIACom-2008.

[4] T.Berner-Lee and M. Fishetti, Weaving the web "chapter Machines and the web,"Chapter Machines and the web, pp. 177-198, 1999.

[5] D.Fensal, W. Wahlster, H. Lieberman, "Spanning the semantic web: Bringing the worldwide web to its full potential, "MIT Press 2003.

[6] G. Bholotia et al.: "Keyword searching and browsing in database using BANKS," 18th Intl. conf. on Data Engineering (ICDE 2002), San Jose, USA, 2002.

[7] D. Tümer, M. A. Shah, and Y. Bitirim, An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia, 2009 4th International Conference on Internet Monitoring and Protection (ICIMP'09) 2009.

[8] "Top 5 Semantic Search Engines".http://www.pandia.com/.

[9] H. Dietze and M. Schroeder, GoWeb: a semantic search engine for the life science web. BMC bioinformatics,Vol. 10, No. Suppl 10, pp. S7, 2009.

[10] Fu-Ming Huang et al. "Intelligent Search Engine with Semantic Technologies"

[11] S. A. Inamdar1 and G. N. Shinde "An Agent Based Intelligent Search Engine System for Web mining" Research, Reflections and Innovations in Integrating ICT in education. 2008.