Architecture Based Study Of Search Engines And Meta Search Engines For Information Retrieval

A. Madhavi[1],K. Harisha Chari[2]
[1]Asst.Professor, Matrusri Institute of PG Studies, Hyderabad, India.
[2]Asst.Professor, K.V Ranga Reddy College, Hyderabad, India.

Abstract

WWW is a huge repository of information. The complexity of accessing the web data has increased tremendously over the years. There is a need for efficient searching techniques to extract appropriate information from the web, as the users require correct and complex information from the web.

This paper analyses the architectures and features of metasearch engines for searching and receiving documents on single and multiples domains on the web. A web search engine searches for information in WWW.

1. Introduction

Web search engines have been helping users find content online for a decade. Today, as then, an individual search engine indexes only a portion of the available content. Metasearch services, introduced a year later, send a user's query to multiple search engines, thus providing the means for a user to search a broader set of documents and potentially get a better set of results. Building a good metasearch engine can be difficult because different query languages are needed to access various engines and the engines use undisclosed ranking algorithms. Popular metasearch engines additionally need to pay for bandwidth, and negotiate with the primary engines for continued high volume access.

Current metasearch engines make several decisions on behalf of the user, but do not consider the user’s complete information need when making these decisions. A metasearch engine must decide which sources to query, how to modify the submitted query to best utilize the underlying search engines, and how to order the results. Some metasearch engines allow users to influence one of these decisions, but not all three.

The primary advantages of a metasearch engine over a permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Single search engine are increased coverage and a consistent interface. A recent study by Lawrence and Giles estimated the size of the web at about 800 million indexable pages. This same study concluded that no single search engine covered more than about sixteen percent of the total. By searching multiple search engines simultaneously via a metasearch engine, coverage increases dramatically over searching only one engine. Lawrence and Giles found that combining the results of 11 major search engines increased the coverage to about 42% of the estimated size of the publicly indexable web.

A consistent interface is necessary for a metasearch engine to be useful. Such an interface ensures that results from several places can be meaningfully combined, while insulating the user from the specifics of the underlying search engines.

Through automatic information retrieval existed before WWW, post-Internet era has made it indispensable. IR is sub field of computer science concerned with presenting relevant information, gathered from online information sources to users in response to search queries. Various types of IR tools have been created, solely to search information on Internet. Apart from heavily used search engines other useful tools are deep-web search portals, web directories and meta search engines. Among various IR tools available, an index is searched rather than entire web. Index is created and maintained through on going automated web searching by programs commonly known as spiders. Web-directories are databases of web sites compiled and maintained by humans. Since web-directory content is hand picked by humans, results have high frequency but a typical web directory’s index size will be only a fraction of that of the search engine and content can easily become outdated. Open Directory Project is biggest web-directory project available today.

Meta Search Engines are derived from general Search Engines which indexes the content available in the entire World Wide Web. A spider explores hyper-linked documents of web, searching and gathering web pages to index. Index and copy of documents themselves are stored in a database. SE accepts a query and creates a list of links to web documents marching query and presents it to the user. Though this logical architecture is basically same as Search Engines, almost all modern Search Engines use computer cluster and massive parallelism to handle heavy loads and to provide fail-safe functionality.

2. Metasearch Engines Architecture

In this section, we describe several architectures used by the metasearch engine.Meta Search Engines can be classified into two types. a) General purpose metasearch engine and b) Special purpose Meta Search Engines. The former aims to search the entire Web, while the latter focuses on searching information in a particular domain (e.g., news, jobs).

- Major Search Engine Approach**:** This approach uses a small number of popular major search engines to build a metasearch engine. Thus, to build a general-purpose metasearch engine using this approach, we can use a small number of major search engines such as Google, Yahoo!, Bing (MSN) and Ask. Similarly, to build a special purpose Meta Search Engine for a given domain, we can use a small number of major search engines in that domain.
- Large scale Metasearch engine approach**:** In this approach, a large number of mostly small search engines are used to build a Metasearch engine. For example, to build a general-purpose metasearch engine using this approach, we can perceivably utilize all documents driven search engines on the Web. Such a metasearch engine will have millions of component search engines. Similarly to build a special purpose metasearch engine for a given domain with this approach, we can connect to all the search engines in that domain. For instance, for the news domain, tens of thousands of newspaper and news-site search engines can be used.

Each of the above two approaches has its advantages and disadvantages. An obvious advantage of the major search engine approach is that such a metasearch engine is much easier to build compared to the large-scale metasearch engine approach because the former only requires the metasearch engine to interact with a small number of search engines. Almost all currently popular metasearch engines, such as Dogpile, Mamma and MetaCrowler, are built using the major search engine approach, and most of them use only a handful of major search engine. One example of a large-scale special-purpose metasearch engine is AllInOneNews, which uses about 1,800 news search engines from about 200 countries/regions. In general, more advanced technologies are required to build large-scale metasearch engines. As these technologies become more mature, more large-scale metasearch engines are likely to be built.

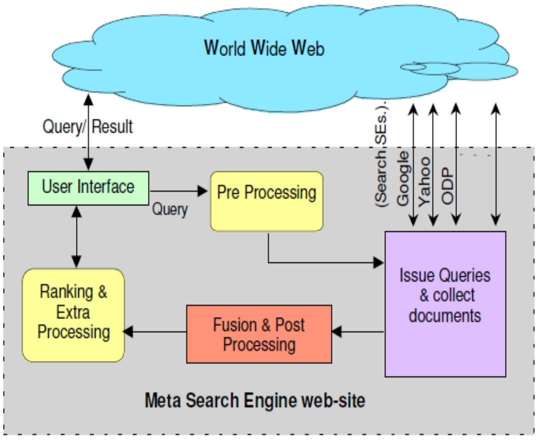

Figure1: Architecture of a typical Meta search Engine.

3. Information Retrieval using MSEs

Metasearch Engines on internet have improved continually with applications of new technologies and methodologies. Understanding and utilizing of MSEs are valuable for computer scientists and researchers, for effective information retrieval.

Meta Search engine receives request from the user and sends the request to various search engines. The search engines check their indices and extract a list of web pages as links and pass the result to the Meta Search Engine. The Meta Search Engine receives the links, applies few algorithms, ranks the results and finally displays the result. The Meta Search Engine architecture is shown in Fig.2.

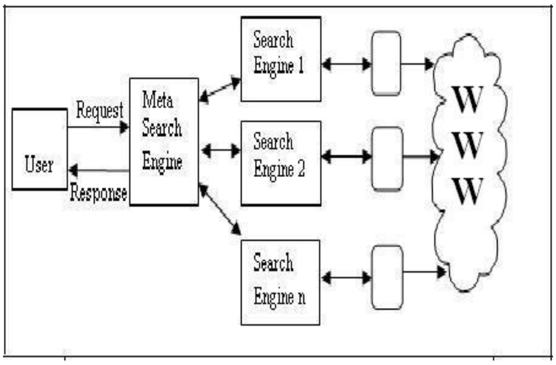

Figure 2: General architecture of Metasearch Engine

We studied a user-friendly Multi-Domain Meta Search Engine that sends search queries to various search engines and to retrieve results from them. Various cases of selection of search engines are proposed in this paper. The basic architecture of the Multi-Domain Meta Search Engine is given in Fig. 3.

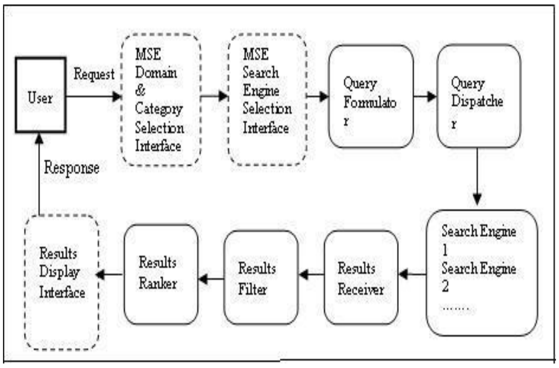

Figure 3: Multi Domain Meta Search Engine Architecture

In the first model the Meta Search Engine was designed to send queries to Search Engines such as Google, Yahoo, AltaVista and AskJeeves. The queries were formed for a specific domain by adding the domain name as part of the search query. An interface for the Meta Search Engine was formed consisting of push buttons to select the domain and text box to enter the query string. The query string to be sent to the search engines are formed with the + symbol applied on the string entered and the selected domain name.

The second model of Multi Domain Meta Search Engine aims at providing an efficient information retrieval on a particular domain for the user by accessing various Vertical Search Engines.

The result of the query can be shown as a list of results from each search engine in individual windows, or as a list of results in different frames of a same window or as a single ordered list of results.

The Multi-Domain Meta Search Engine consists of a user interface for domain selection, a user interface for search engine selection, query formulator, dispatcher, results receiver, filter, ranker and display of results.

- User Interface – Domain Selection
- User Interface – Search Engine Selection
- Query Formulator and Dispatcher
- Results Receiver
- Filtering and Ranking results
- Results generation in different windows
- Result generation in multiple frames of same window
- Result generation in same window.

Finally, performance and evaluation of a metasearch engine is based on many factors like

- Performance reports
- Popularity Statistics

and certain algorithms like

1. Algorithm 1: Use Top Document to Compute Search Engine Score (TopD)

2. Algorithm 2: Use Top SRRs to Compute Search Engine Score (TopSRR)

3. Algorithm 3: Compute Simple Similarities between SRRs and Query (SRRSim)

4. Algorithm 4: Rank SRRs Using More Features (SRRRank)
5. Algorithm 5: Compute Similarities between SRRs and Query Using More Features (SRRSimMF**)**

4. Findings

4.1. Testbed

The purpose of this work is to evaluate and compare different result merging algorithms under the context of metasearch over the general-purpose search engines. So we select 10 most popular general-purpose search engines as the underlying component search engine. They are: *Google*, *Yahoo*, *MSN*, *Askjeeves*, *Lycos*, *Open Directory*, *Altavista*, *Gigablast*, *Wisenut*, and *Overture*. The reasons these search engines are selected are:

(1) They are used by nearly all the popular general- purpose metasearch engines;
(2) Each of them has indexed a relatively large number of web pages; and
(3) They adopt different ranking schemes. Even though we focus our work in the context of general-purpose search engines.

4.2 Evaluation Criteria

Because it is difficult to know all the relevant documents to a query in a search engine, the traditional *recall* and *precision* for evaluating IR systems cannot be used for evaluating search/ metasearch engines. A popular measure for evaluating the effectiveness of search engines is the *TREC-style average precision* (TSAP). TSAP at cutoff *N*, denoted as *TSAP@N*, will be used to evaluate the effectiveness of each merging algorithm:

$$TSAP@N = (\sum_{i=1}^{N} r_i)/N$$

where ri = 1/i if the i-th ranked result is relevant and ri = 0 if the i-th result is not relevant. It is easy to see that TSAP@N takes into consideration both the number of relevant documents in the top N results and the ranks of the relevant documents. TSAP@N tends to yield a larger value when more relevant documents appear in the top N results and when the relevant documents are ranked higher. For each merging algorithm, the average TSAP@N over all 50 queries is computed and is used to compare with other merging algorithms.

5. Future Work

We need in-depth study regarding the functionality and implementation of search engines and metasearch engines The Web Search Engine, Vertical Search Engine and Meta Search Engine features are necessary to make a deep study. Multi-Domain Meta Search Engines that provide an efficient information retrieval on various domains using various Search Engines are also presented. Few Search Engines are tested for the relevancy, reliability, redundancy and availability of search results for few topics. Domain selection and search engine selection user interfaces are also presented. A query interface is designed to send user's search query to the various search engines and the resultant links are consolidated to display the results to the user. The system can be extended to process the content of the web pages in addition to processing of the links.

6. Conclusion

Analysis and implementation of collection fusion in MSEs reported, seems to be promising in improving information retrieval. Research into MSEs has been mainly focused on source engine selection, re-ranking and integration of multiple search engines. Apart from design and evaluation of full MSEs, research in distributed search, database selection and results merging combination techniques have a direct bearing on the domain of meta-searching. Recent smarter metasearch technology includes clustering and linguistic analysis that attempts to show themes within results, textual analysis and display that can help to dig deeply into a set of results. Rapid growth and evaluation of web is posing new challenges to MSEs. Emerging of services like social-networking, weblogs, RSS feeds, and increase in non textual information like podcasing, online-videos and convergence of non-conventional forms of communication such as mobile phones with internet will require easy IR in these areas.

References

[1]. Brian D. Davison - The potential of the metasearch engine, "Proceedings of the Annual Meeting of the American Society for Information Science and Technology, Providence, RI, November 2004."

[2]. Eric J. Glover, Steve Lawrence', William P. Birmingham, C. Lee Giles - Architecture of a Metasearch Engine that Supports User Information Needs.

[3]. Manoj M and Elizabeth Jocob - Architecture of a Metasearch Engine that Supports User Information Needs, Journal of Scientific and Industrial Research, volume 67, October 2008, pp.739-746.

[4]. Anwar A. Alhenshiri - Web Information Retrieval and Search Engines Techniques, Al-Satil journal, pp.55 – 92.

[5]. D.Minnie, S.Srinivasan – "Metasearch Engine with an Intelligent Interface for Information Retrieval on Multiple Domains", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.4, October 2011.

[6]. Eui-Hong (Sam) Han and George Karypis, Doug Mewhort and Keith Hatchard – "Intelligent Metasearch Engine for Knowledge Management", *CIKM'03,* November 3–8, 2003, New Orleans, Louisiana, USA.