

An Intelligent Meta Search Engine for Efficient Web Document Retrieval

Keerthana.I.P¹, Aby Abahai.T²

¹(Dept. of CSE, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India)

²(Dept. of CSE, Assistant Professor, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India)

Abstract: In daily use of internet, when searching information we face lots of difficulty due to the rapid growth of Information Resources. This is because of the fact that a single search engine cannot index the entire web of resources. A Meta Search Engine is a solution for this, which submits the query to many other search engines and returns summary of the results. Therefore, the search results receive are an aggregate result of multiple searches. This strategy gives our search a boarder scope than searching a single search engine, but the results are not always better. Because the Meta Search Engine must use its own algorithm to choose the best results from multiple search engines. In this paper we proposed a new Meta Search Engine to overcome these drawbacks. It uses a new page rank algorithm called modified ranking for ranking and optimizing the search results in an efficient way. It is a two phase ranking algorithm used for ordering the web pages based on their relevance and popularity. This Meta Search Engine is developed in such a way that it will produce more efficient results than traditional search engines.

Keywords: Metacrawler, Meta Search Engine, Ranking, Search Engine, Web Crawler

I. Introduction

The World Wide Web (WWW) is a collection of information linked together from all over the world. It is based on hypertext, which means that when the user is navigating on the ocean of information he can pick up an interesting word or expression within a text and request for more information about it. This can not apply to all words in a text, but only to those who have been properly designated as such by the producer of the information and which are displayed on screen example as underlined. The web is highly dynamic in nature i.e., new web pages are being added, some are removed and some others are modified. So it is very difficult to find relevant information from it. Search engines are used to extract valuable Information from the internet. A search engine [1] is a program that searches for and identifies items in a database that correspond to keywords or characters specified by the user, especially for finding particular sites on the WWW.

On the Internet, a search engine is a coordinated set of programs that includes: A spider (also called a "crawler" or a "bot") that goes to every page or representative pages on every web site that wants to be searchable and reads it, uses hypertext links on each page to discover and read a site's other pages. Search engine includes program that creates a huge index (sometimes called a "catalog") from the pages that have been read. It also contains a program that receives our search request, compares them to the entries in the index and then returns results.

A Meta Search Engine [2] is a search engine that queries other search engines and then combines the results that are received from all. Here, the user is not using just one search engine but a combination of many search engines at once to optimize web searching. Meta Search Engines don't create their own database of information. They search the other search engine's databases. The major advantage of a Meta Search Engine is that it allows the user to search several search engines simultaneously. So there is no need to search in each search engine separately.

In this paper a new Meta Search Engine is proposed, which is capable of producing relevant results for particular query string. It will display the results from different search engines as soon as those found by a crawler. Finally it will rank each and every web page based on its popularity and relevance. For that, this paper proposes a new ranking algorithm. This paper is organized as follows: in section II, the works related to search engines is discussed. Section III describes the Meta Search Engine implementation. In section IV the performance analysis of the proposed Meta Search Engine is discussed. Section V describes the advantages of the proposed system. Finally the section VI gives the conclusion for the work.

II. Related Works

2.1 Search Engines

A search engine is designed to search WWW for information about the given search query and returns links to various documents in which the search query's keywords are found. There are mainly three types of search engines-Web Search Engines, Meta Search Engines, Vertical Search Engines [3]. A web search engine

searches for information in WWW. A Vertical Search provides the user with results for queries on a particular domain. Meta Search Engines send the user's search queries to various search engines and combine the search results. It is important for search engine to maintain a high quality websites. A database is to be made in which following attributes should be there like length of title, keywords in title, number of back-links, in-links, length of the URL, ranking. WEKA and Tanagra a data mining tools can be used to extract meaningful knowledge from this large set of data [4].

The search engines used in this paper for the Meta Search Engine construction are Google, Yahoo, Ask and Bing [5]. Google is a web search engine owned by Google Inc. it is the most widely used search engine on the Web. The main purpose of Google Search is to search for text in WebPages, as opposed to other data, such as with Image Search. Google was developed by Larry Page and Sergey Brin in 1997. Yahoo is a web search engine, owned by Yahoo! Inc. the 2nd largest search engine on the web by query volume. It was founded in January 1994 by Jerry Yang and David Filo. Ask is a Search Engine, which is also known as Ask Jeeves. It is basically designed to answer the user's Queries in the mode of Q&A and is proved to be a focused search engine. Ask was developed in 1996 by Garrett Gruener and David Warthen in Berkeley, California. Bing is a Search Engine, which was formerly known as Live Search and MSN Search. It is a web search engine that was owned by Microsoft. It went fully online on 2009 June 3, with a preview version released on June 1, 2009.

2.2 Web Crawlers

All search engines internally use web crawlers to keep the copies of data a fresh. A search engine is divided into different modules. Among those modules crawler module is the module on which search engine relies the most because it helps to provide the best possible results to the search engine. Crawlers or spiders are small programs that browse the web on the search engine's behalf, similarly how a human user would follow links to reach different pages. Google crawlers run on a distributed network of thousands of low-cost computers and can therefore carry out fast parallel processing. This is the way Google returns results within fraction of seconds. It has three main components [6]: a frontier which stores the list of URL's to visit, a Page Downloader which download pages from WWW and Web Repository receives web pages from a crawler and stores it in the database. The repository stores only standard HTML pages. The working of a web crawler is as follows [6]:

1. Initializing the seed URL or URLs
2. Adding it to the frontier
3. Selecting the URL from the frontier
4. Fetching the web-page corresponding to that URLs
5. Parsing the retrieved page to extract the URLs
6. Adding all the unvisited links to the frontier
7. Again start with step 2 and repeat till the frontier is empty.

Based on different strategies employed in web crawling there are four types of web crawlers: focused crawler, distributed crawler, incremental crawler, and parallel crawler [6]. A focused crawler is the Web crawler that tries to download pages which are related to each other. It collects documents which are specific and relevant to the given topic. An incremental crawler is a traditional crawler which refreshes its collection periodically and replaces the old documents with the newly downloaded documents. It also exchanges less important pages by new and more important pages. A distributed Crawler uses a distributed computing technique. Many crawlers are working to distribute in the process of web crawling to get most coverage of the web. A central server manages the communication and synchronization of the nodes, as it is geographically distributed. In the case of parallel crawler multiple crawlers are often run in parallel, which are termed as Parallel crawlers. A parallel crawler consists of multiple crawling Processes called as C-procs which can run on network of workstations. The Parallel crawlers depend on Page freshness and Page Selection. A Parallel crawler can be on local network or be distributed at geographically distant locations.

Popular examples of web crawlers are Yahoo Slurp was the name of the Yahoo search crawler, Googlebot is the name of the Google search crawler, Bingbot is the name of Microsoft's Bing web crawler, etc.

2.3 Web Page Elements

The main web page elements that will be extracted by a crawler are given below [7].

Title Tag: Each web page has a unique title tag, which helps search engines to know how the page is distinct from the others on the site. Titles can be both short and informative. If the title is too long, then search engines will show only a portion of it in the search result. Page titles are an important aspect of search engine optimization.

Description Meta Tag: A page's description meta tag gives search engines a summary of what the page is about. A page's title may be a few words or a phrase, but a page's description meta tag might be a sentence or two or a short paragraph. Fig.1 shows an example for title and description tags.

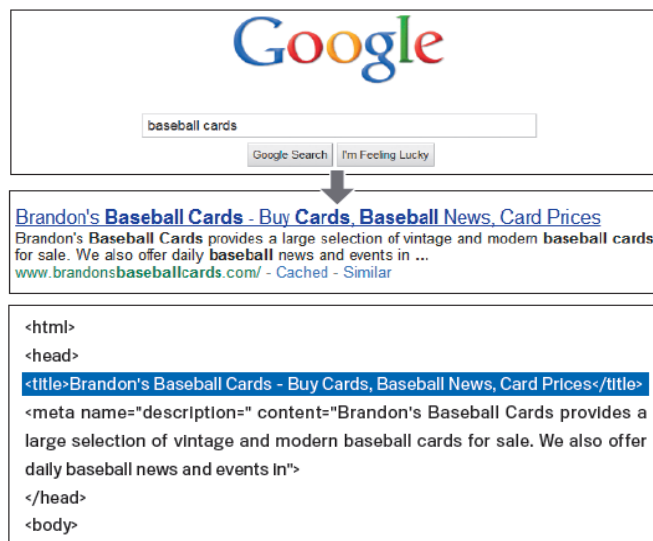


Fig.1: Title and Description Tags

URL (Uniform Resource Locator): The URL to a document is displayed as part of a search result in search engines, below the document's title and snippet. Search engines are good for crawling all types of URL structures, even if they are quite complex, but spending the time to make the URLs as simple as possible for both users and search engines can help. Fig.2 shows a sample URL.

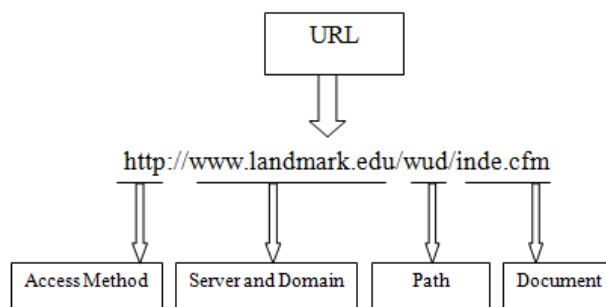


Fig.2: URL

Anchor Text: Anchor text (See Fig.3) is the clickable text that users will see as a result of a link, and is placed within the anchor tag ``. This text tells search engines something about the page you're linking to. It may be internal (pointing to other pages on the same site) or external (leading to content on other sites).

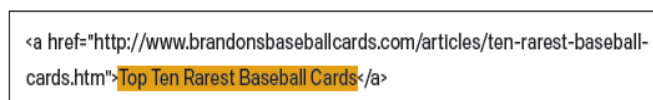


Fig.3: Anchor Text

2.4 Searching Techniques & Strategies

Several searching techniques and strategies are essential for narrowing search results and will assist you with locating valuable resources via the Internet. These techniques and strategies include wildcard, plus and minus signs, quotation marks and brackets, pipe symbol, Boolean operators, and nesting [8].

Wildcard: A wildcard is a special character that can be added to a phrase while searching and the search engine or subject directory looks for all possible endings. The results will provide all possible documents in their database that have those letters.

Plus and Minus Signs: The plus sign used before a keyword or phrase should retrieve results that include that specific keyword or phrase. The minus sign used before a keyword or phrase should retrieve results that exclude that specific keyword or phrase.

Quotation Marks and Brackets: Quotation marks and brackets assist with narrowing the search results from the search tools. When they are used, the search engine will only retrieve documents that have those key terms appearing together.

Pipe Symbol (|): The pipe (|) symbol is used for narrowing down results within a broad category.

Boolean Operators: Working of AND operator is similar to the plus sign and the NOT operator is similar to the minus sign. The OR operator tells the search engine to retrieve one term or the other.

Near: The NEAR phrase indicates to the search tools that those terms must be located within a certain number of words. The results may vary depending on the search tool. To illustrate, some search tools may try to locate the terms within 2, 10 or 25 words of each other. The command to use is NEAR/#.

Nesting: Nesting allows the user to perform complex search. Here the parentheses are used to group the key words and Boolean operators together.

2.5 Challenges Faced By Search Engines

Search engines are the major tools for retrieving information from the web. But the list of retrieved documents contains a high percentage of redundant and near document result. So there is the need to improve the performance of search results. Most of the current search engine use data filtering algorithm which will eliminate near duplicate and duplicate documents to save the users time and effort [9]. The identification of near-duplicate pairs or similar documents in a large collection is a major problem with many applications.

The problem with most of the search engines are [5], most people don't take advantage of the offered tools, but instead, they just type a few keywords for a query. Thus, search engines have to find a way how to improve basic queries so that they can provide users successful research at the same time. Using a search engine, an index is searched rather than the entire Web. An index is created and maintained by automated web searching by spiders. Plain search engines prove to be very effective for certain types of search tasks, such as retrieving of a particular URL and transactional queries. However, search engines can't address informational queries, where the user has information that needs to be satisfied. A Meta Search Engine overcomes the above by virtue of sending the user's query to a set of search engines, collects the data from them displays the relevant records by using a clustering algorithm [10].

III. Implementation

The new Meta Search Engine has the following components: Meta Search Engine GUI, Query Formulator, Metacrawler, Redundant URL Eliminator, Modified Ranking and Result Generation. The new architecture is built by using four search engines Google, Yahoo, Bing and Ask. The proposed architecture is shown in Fig.4. It shows the interaction between various components of the Meta Search Engine.

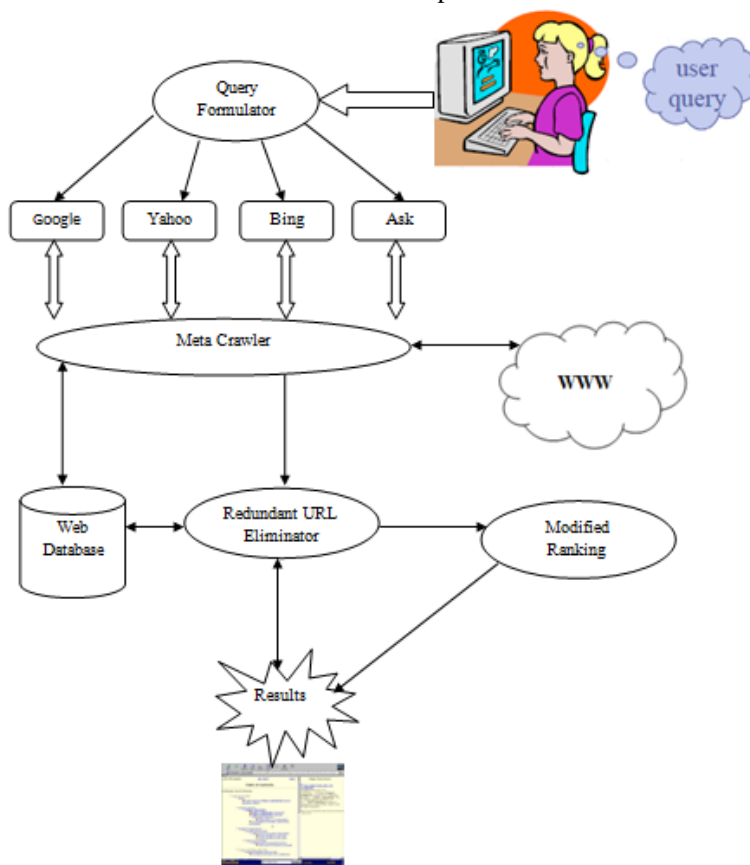


Fig.4: Meta Search Engine Architecture

3.1 Meta Search Engine GUI

In the proposed architecture, user gives the keyword to be searched through the Meta Search Engine Graphical User Interface (GUI). Here user has the choice to select the search engines he wants to search (See Fig.5). The query is the given to all selected search engines. The results corresponding to the query are then stored into the web database. The new Meta Search Engine retrieves important results from four different traditional search engines available. The retrieved results are then extracted by using the crawler. We have developed this application using Adobe Dreamweaver, using PHP for front end designing and My SQL Server as back end. We have used XAMPP Apache distribution to link front end and back end in our application.

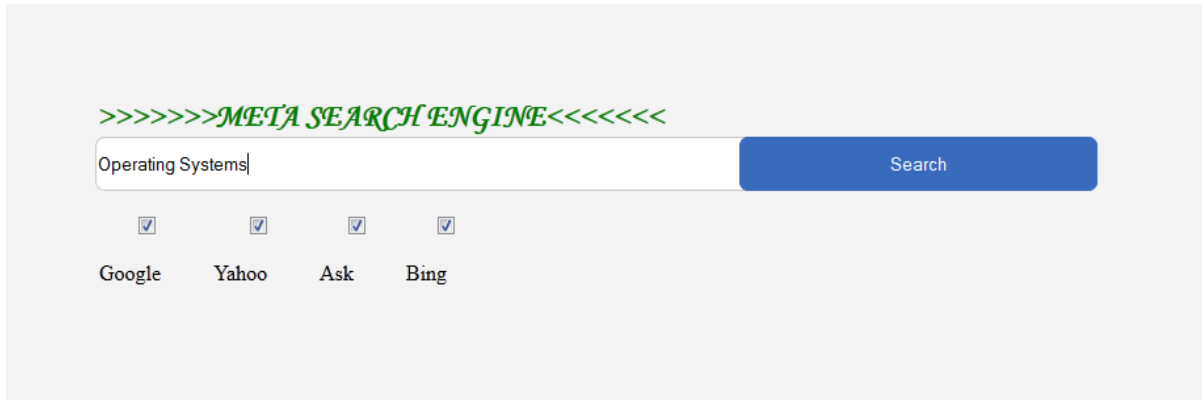


Fig.5: Meta Search Engine GUI

3.2 Query Formulator

As different search engines follow different styles for the representation of the query search string, different search query strings are generated for a given user input. The query strings are then sent to various search engines to extract desired results from the search engines. The query string formats for each of the search engines are given below.

For Google: <http://www.google.co.in/search?q=sword>

For Yahoo: http://in.search.yahoo.com/search;_ylt=?p=sword

For Bing: <http://www.bing.com/search?q=sword>

For Ask: <http://www.ask.com/web?q=sword>

Here sword is the user input. For example if we want to search “data mining” then the query string corresponding to this for Google is:

<http://www.google.co.in/search?q=data+mining>.

In the new Meta Search Engine we are focusing in crawling the first result page suggested by every search engine. Because important and highly relevant results are located in the first page of the results and majority of the users are focusing on those pages.

3.3 Metacrawler

The web crawler used in the new Meta Search Engine is termed as metacrawler. Because it helps to crawl web pages through multiple search engines. We developed this using PHP curl. Curl is a library that helps us to make HTTP requests in PHP. Using curl we can set the various parameters for HTTP requests like user agent, follow location details, HTTP version, header details, certificate verification details etc. It will help to handle authentication. The metacrawler will crawl through each and every search engine to find the web pages relevant to a query string. When crawling a web page it will remove scripts, style sheets and other information from the web page source code. It will only extract html code. Then it will convert the HTML code into DOM (Document Object Model) structure.

Compared with traditional plain text, a Web page has more structure. Web pages are also regarded as semi-structured data, which are represented as a DOM structure. The DOM structure of a Web page is a tree structure, in which every HTML tag in the page represents a node in the DOM tree. The Web page can be segmented by some predefined structural tags. Useful tags include <P> (paragraph), <TABLE> (table), (list), <H1>-<H6> (heading), etc. Thus the DOM structure can be used to facilitate information extraction. From the DOM structure it will extract the document or web page URL, title and description. The extracted details are the stored to the web database.

3.4 Redundant URL Eliminator

Since the metacrawler combines the results from different search engines, there is a chance for the link suggested by one search engine will also suggested by another search engines. Here the redundant URL eliminator component will check whether an extracted link is a duplicate one by looking if it is already in the web database URL list. If the URL is already present in the web database it is duplicate, so it will not add into the result page.

As soon as results are available from one search engine, they are displayed on the screen to view. The results are therefore that much faster, and while you are down through the list of hits, other results are being added to the page even as you view. The new results found will be added to result page only after checking if it is duplicate or not. To check this for each new link suggested, the system will find its domain name. If this domain name matches with the domain name of another link in the same result set then the previous one will kept and the later one will avoid. So here duplicate link elimination is not only on the basis of URL similarities but also on the basis of domain a name matches. It will help to optimize the results.

3.5 Modified Ranking

Ranking is used for ranking or ordering the web pages based on their relevance. This assigns a score to each link in the result set. The system is based on two phase ranking. First phase is based on the domain name of each web page. Second phase is based on outgoing and incoming link counts of each web page.

Phase 1: First the system will find the domain name of each link in the result set. Then it will calculate the count of each domain name in the result set. Then arrange the links in decreasing order of their domain name counts in the result set.

Phase 2: In this phase a web page is assigned a rank based on its domain name count and its popularity. A web page that has more incoming and outgoing links is considered as more popular and relevant. So that phase 1 follows a modified ranking which calculate a rank for each web page based on this.

Algorithm: Modified Ranking

Input: Result Set R which contains links from phase 1

Output: Ordered list of Result Set R

1. Start
2. Initialise the score of all links to their domain name count in the result set R
3. **for each** link L_i **in** result set R
4. Crawl web page corresponding to link L_i
5. Find all its outgoing links and store them into list OL
6. **for each** link L_j **in** OL
7. **if** L_j is in R **then**
8. Increment score of L_i by 1
9. Increment score of L_j by 1
10. **end if**
11. **end for**
12. **end for**
13. Order links in R in the decreasing order of their scores
14. **return** R to result page
15. End

The above algorithm calculates the score of the web pages based on their outgoing links. The incoming links of a web page are included in the scoring when outgoing links of another web page is calculated. For example if link to the web page A is included in web page B, then A is the outgoing link of B and B is the incoming link of A.

3.6 Results Generation

Alternatively the ordering process is performed and the ranked list is displayed as a single list with most relevant links on the top of the result window. The result will contain the extracted title of the web page, URL, and description. The relevancy of the pages is calculated based on input query and the web pages are ranked using the modified ranking algorithm. The pages after ranking are displayed to the user. The top links are the links with highest rank. When moving from top to bottom the rank of web pages is decreasing. Fig.6 shows the ranked results for the search query 'Sachin'. It gives the details of web pages suggested by each of the four search engines Google, Yahoo, Bing and Ask. The results shows the search results merged from the first result page given by each of the search engines. The Meta Search Engine also supports searches other than web search. Fig.7 shows the image search results, Fig.8 shows the video search results and Fig.9 shows the latest news search results.

Home Web Search Image Search Video Search Latest News Update Profile Log Out

>>>>>META SEARCH ENGINE<<<<<<

Sachin

Google Yahoo Ask Bing

Sachin Tendulkar - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Sachin_Tendulkar

Sachin Retires, ODI Weeps - The Hindu
<http://www.thehindu.com/opinion/blogs/blog-hawk-eye/article4234484.ace>
After the retirement of Sachin Tendulkar, ODI cricket faces a bigger challenge to sustain itself.

Sachin Tendulkar leads Twitter tributes for retiring Mahela Jayawardene and Kumar Sangakkara | The Indian Express
<http://indianexpress.com/article/sports/cricket-world-cup/2320059/sachin-tendulkar-leads-twitter-tributes-for-retiring-mahela-jayawardene-and-kumar-sangakkara/>
Kumar Sangakkara and Mahela Jayawardene ended their careers and as the dust settled the adulation poured in.

Sachin Tendulkar | India Cricket | Cricket Players and Officials | ESPN Cricinfo
<http://www.espncricinfo.com/india/content/player/35320.html>
Sachin Tendulkar's Cricinfo profile

Amazon.in: Sachin Tendulkar
<http://www.amazon.in/Sachin-Tendulkar/s?ie=UTF8&page=1&rn=i%3Aaps%2Ck%3ASachin%20Tendulkar>
Amazon.in: Sachin Tendulkar

Sachin Tendulkar Latest News, Photos, Biography, Stats, Batting averages, bowling averages, test & one day records, videos and wallpapers at CricketCountry.com
<http://www.cricketcountry.com/players/sachin-tendulkar>
Sachin Tendulkar Biography - Get Sachin Tendulkar full profile with all the records, quotes and latest news. Also have a look on Sachin Tendulkar career statistics and performance analysis with batting, fielding & bowling averages at CricketCountry.com


Sachin Tendulkar | India Cricket | About Sachin Tendulkar Stats, Records | NDTVSports.com

Fig.6: Web Search Results


Home Web Search Image Search Video Search Latest News Update Profile Log Out

Sachin


Google



Yahoo



Bing



Ask




Fig.7: Image Search Results

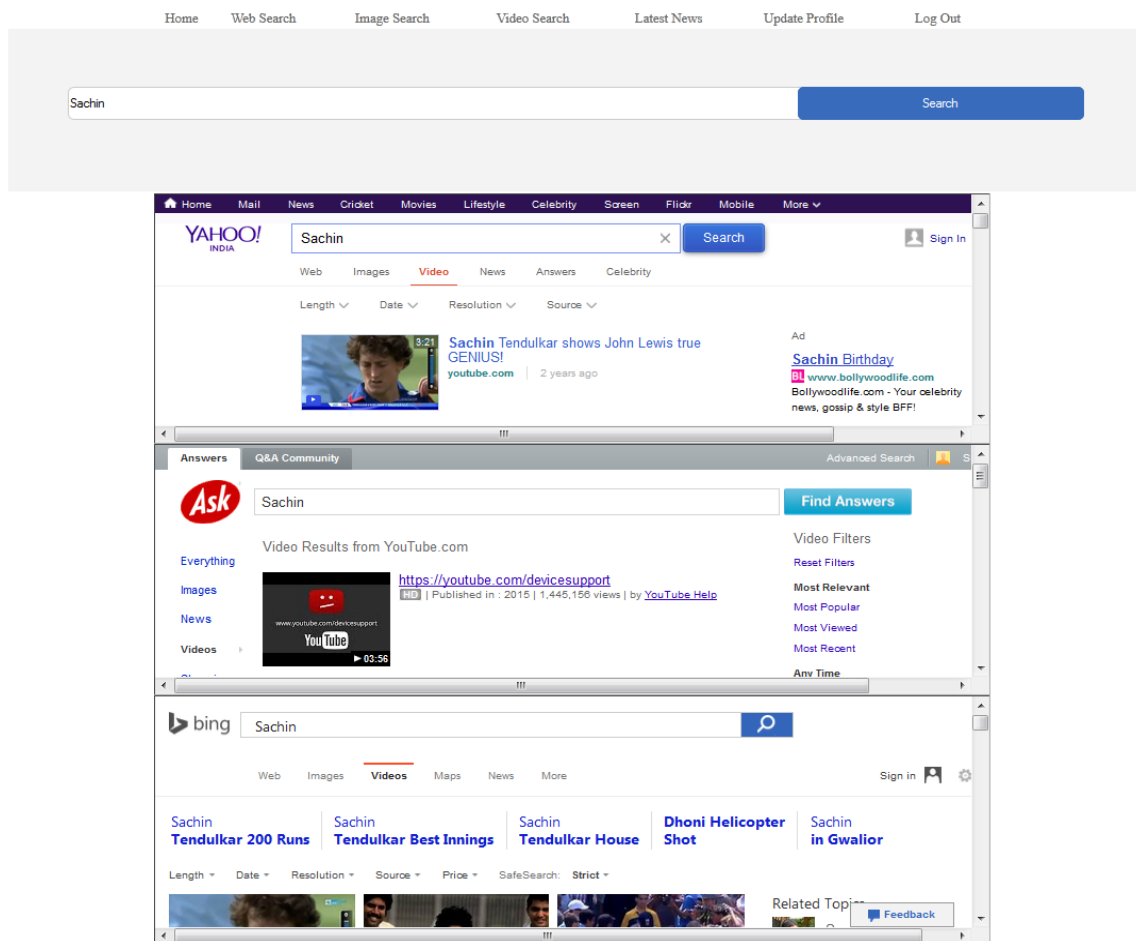


Fig.8: Video Search Results

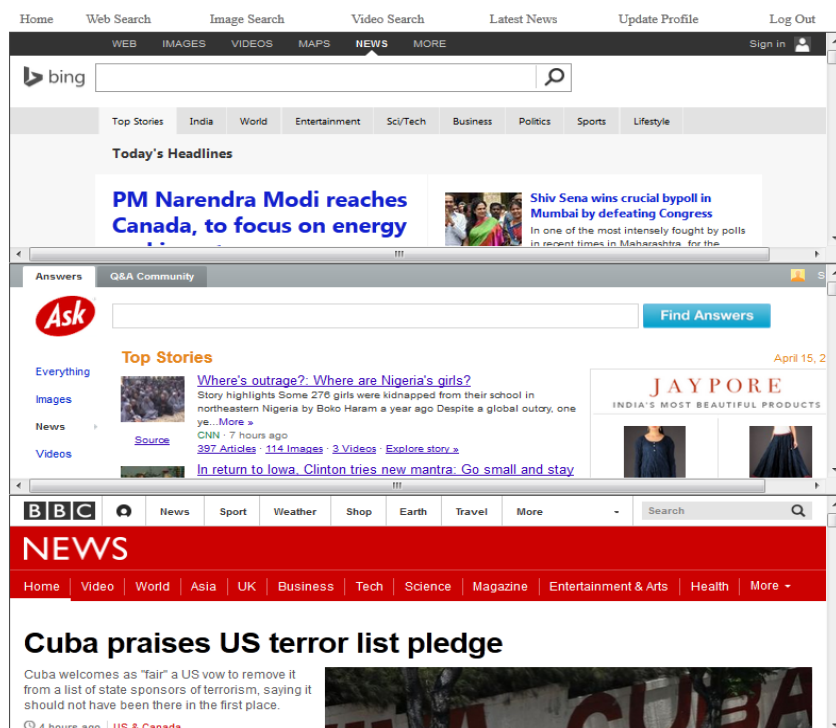


Fig.9: Latest News Search Results

IV. Results

The two basic measures for checking the effectiveness of the search engines in document retrieval are precision and recall.

Precision: This is the percentage of retrieved documents that are in fact relevant to the query. It is defined as:

$$\text{Precision} = \frac{|{\text{Relevant}} \cap {\text{Retrieved}}|}{|{\text{Retrieved}}|}$$

Recall: This is the percentage of documents that are relevant to the given query and were, in fact, retrieved. It is defined as:

$$\text{Recall} = \frac{|{\text{Relevant}} \cap {\text{Retrieved}}|}{|{\text{Relevant}}|}$$

The total number of relevant documents retrieved by executing ten different queries on the traditional search engines and the new Meta Search Engine, with the average precision value is given in Table 1. It represents the performance of the proposed Meta Search Engine in terms of precision.

Table 1: Relevant Documents Retrieved

Query Number	Search Engines				
	Google	Yahoo	Ask	Bing	Meta Search Engine
1	7	5	5	6	13
2	9	7	6	7	15
3	8	7	5	6	16
4	5	5	5	4	12
5	7	6	5	5	15
6	8	6	6	6	11
7	7	6	5	6	14
8	7	7	6	6	16
9	6	6	4	5	14
10	6	5	5	6	12
Average Precision	0.93	0.88	0.83	0.87	0.96

The Fig.10 shows the comparison between the precision of different search engines. From the results we can see that the new Meta Search Engine has a higher precision value when compared to the search engines Google, Yahoo, Ask and Bing. So we conclude that the Meta Search Engine is more efficient in relevant document retrieval and its performance is better than the performance of individual search engines.

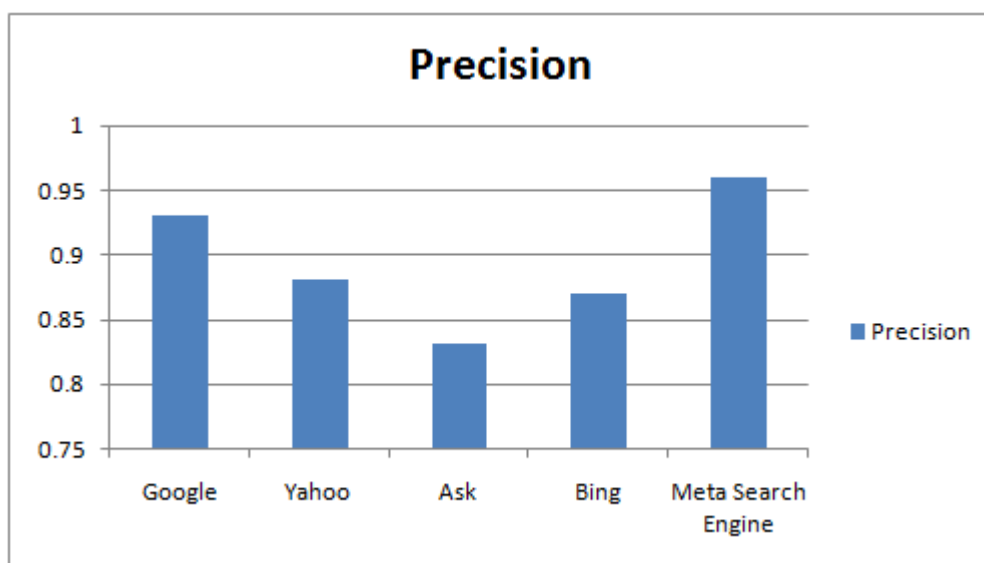


Fig.10: Precision

V. Advantages

The proposed Meta Search Engine has many advantages compared to other search engines. It extends the search coverage of the topic and allows more web pages to be found. It can utilize some of the specific functionality of the search engines that they employ. It searches all the major search engines at once so it reduces the work of users from having to separately type in different search engines to look for resources. It is easy to use and similar format to other search engines. It supports and utilizes the searching syntax of each search engine, including booleans, phrase searching, pluses and minuses, wildcards, field searching, etc. It can consolidate and remove duplicates from the results. It is customizable for user's preferences. It use the indexes created by other search engines, combining and post processing results in unique way. Using this more results can be retrieved with the same effort. Overall it provides improved recall and precision.

VI. Conclusion

WWW search engines have a number of deficiencies including: periods of downtime, low coverage of the WWW, inconsistent and inefficient user interfaces, out of date databases, poor relevancy ranking and precision, and difficulties with spamming techniques. In this paper a new Meta Search Engine has been introduced which address some of these and other difficulties in searching the WWW. It is understood that Meta Search Engine retrieves efficient results than any Search Engine and its performance also improves factors like recall and precision. It will combine and rank the results in an efficient way using the modified page rank algorithm. It is powerful because one search can highlight strengths of a number of top search engines such as Google, Yahoo, Bing and Ask.

References

- [1]. What Is A Search Engine? Digitallearn.Org a PLA Initiative, <http://digitallearn.org/sites/default/files/class/Basic%20Search/supplemental-materials/basicsearchhandout.pdf>
- [2]. Subarna Kumar Das, "Role Of Meta Search Engines In Web- Based Information System: Fundamentals And Challenges", 4th Convention Planner -2006, Mizoram Univ, Aizawl, 09-10 November, 2006
- [3]. D.Minnie, S.Srinivasan, "Meta Search Engine With An Intelligent Interface For Information Retrieval On Multiple Domains", International Journal of Computer Science, Engineering and Information Technology (IJCEIT), Vol.1, No.4, October 2011
- [4]. Minky Jindal , Nisha Kharb, "K-Means Clustering Technique On Search Engine Dataset Using Data Mining Tool", International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 6 (2013), pp. 505-510
- [5]. Sonali Kushwah, Avdhesh Singh, Indu Gupta, "Analysis on Meta Search Engine "
- [6]. Trupti V. Udupure, Ravindra D. Kale, Rajesh C. Dharmik, "Study Of Web Crawler And Its Different Types", IOSR Journal of Computer Engineering (IOSR-JCE) e-Volume 16, Issue 1, Ver. VI (Feb. 2014), PP 01-05
- [7]. Google Search Engine Optimization Starter Guide, <http://static.googleusercontent.com/media/www.google.com/en/webmasters/docs/search-engine-optimization-starter-guide.pdf>
- [8]. Kimber ly McCoy , "Search Engines, Subject Directories, And Meta-Search Engines", Winter 2000 The OLRC News Volume 5, No: 1
- [9]. Bassma S.Alsulami, Maysoun F. Abulhair, Fathy E. Eassa, "Near Duplicate Document Detection Survey", International Journal of Computer Science & Communication Networks, Vol 2(2), 147-151
- [10]. K R Remesh Babu, AP Arya, "Design of a Metacrawler for Web Document Retrieval", 12th International Conference on Intelligent Systems Design and Applications (ISDA), 27-29 Nov. 2012, pp.478-48