DOI: 10.15415/jtmge.2016.72002

Getting Bulk Data Through Google: An empirical study

Shama Rani
ME Scholar, Chitkara University, India.

Jaiteg Singh
Professor, CURIN Chitkara University, India.

Abstract

To store the information in a database is one of the major tasks. The efficient storage of data is important for future use. Information retrieval is a method of gathering information related to input queries from the various sources or stored databases. To retrieve the information, a search engine plays an important role. A web search engine creates an index to match queries. The quality of information is improved with the help of search engine. For retrieving the information, a search engine comprises some modules such as query processor, a searching and matching function, document processor and page rank capability. This paper focuses on the retrieval of web documents against input queries and stores them in to database. A Google search API can be used to fetch the results. It analyses the data by processing through these modules and downloads the content available in different formats.

Keywords: *Web crawling, indexing, page ranking, retrieve pdf documents, query processing, search engine implementation, web search.*

INTRODUCTION TO INFORMATION RETRIEVAL

Information retrieval is process of accessing desire information form storage system, The data can both be structured or unstructured information retrieval process not only works for large data set but also for small set of data. It helps to reduce time required to extract the information. A search engine is a source of retrieving information for the specified keywords and lists the result. This paper focused on the text data retrieved from google and the comparision of that documents. It has its own importance in different fields such as it is used for business perspective, education purpose, marketing and in the filed of research. It surveys the sites and create a database by retreiving the information from a query[1][3].

LITERATURE REVIEW

Nowadays search engines are very helpful to extract information from world wide web. It contains billions of web pages that are classified, indexed and ranked. The journey of search engines begin since 90's. Nowadays search engines are more advanced. The number of search engines originates time to time with different functionalities. But some of them are active and rest are inactive. Some of them are described here:

Journal of Technology Management for Growing Economies
Vol. 7, No. 2
October 2016
pp. 39-48

1. Archie: it was the first search engine that was based on the FTP sites that stores the listings of the pages are capability of being downloaded, but the index didn't link to user because of limited space. It was found in 1990[3].
2. Veronica and Jughead: As the archie was not able to connect with the user, veronica and jughead were two programs for search that works like GOPHER system. Gopher System, Veronica and Jughead both were launched 1991. Gopher system allows user to find, distributes the data over the web.
3. WWW Excite, Wanderer, Aliweb, Basic Web search are the search engines launched in 1993 based on bots. But now they are inactive.
4. Altavista, web Crawler, Yahoo(1994): it was the first search engine which provids the access to the user to insert and delete the queries with in some given time. It processes the natural language queries with unlimited bandwidth.

   Another search engine named WEB CRAWLER is also launched din 1994 which index the entire pages. YAHOO is also come into exixtence in same year which was cabaple to increase the size of directory of pages searched. These all search engines are active now a days[1].
5. Google, MSN(1996-1998) : The Google search came into existance by the founders working on a Backhub project. This project was related to extract the pages that has most relevance results to the input query. Then 1998 MSN launched by Microsoft is normally used by users.
6. Bing(2006) : Bing is also one of known search engines. Later many more search engines DuckDuckGo, SIRI, CORTANA, BLEKKO are launched recently for various platforms e.g Apple, Windows 8.1[10].

INFORMATION RETRIEVAL MODELS

Information retrieval is a method of gathering information related to input queries from the various sources or stored databases. It is basically a process to recover the information available on various sources into a database. The process starts with the query entered by user. These queries are not uniquely defined because several results match to those queries during the search. These queries are matched with the database information which is already stored on the computer. The generated results are ranked according to page ranking algorithm depends on the relevancy of the information. The top ranked pages are shown to users. The various models are used to retrieve the information on different basis. Boolean model, Vector Space Model, Statistical Model, Probilistic model are the main models used for retreiving information by search engines[11].

1. Boolean Model: Boolean model is based on the boolean algebra. It generally consist three modules
   - Representation for input queries
   - Representation for text Documents
   - A context for represent input queries, document and relationship among them.

The documents are the set of keywords and terms that are presented by Boolean model and the queries input by user are the boolean expressions for keywords. The queries consists of AND, OR and NOT operators. A Document is predicted relevent to input query expression if it satisfies the following condition:

((input text OR information) AND retrieval AND NOT theory)

Where OR is the union of two sets , AND is the intersection of two sets and NOT inverse the sets. Boolean model has some disadvantages :

OR:one matching word is as good as many

AND: one missing word as bad as all

Queries are difficult to express because a keyword has several meanings.

2. Vector Space Model: This is an algebric model that represent the text documents as index terms. This model is most suitable to filter the pages, checking relevance pages to the search term or keywords or to index and rank those pages. . This approach gives the vector representation of the documents. If cosine similarity value is 0 or the angle between objects is 90 degree then the documents do not share any attributes or words. The representation of queries and documents are done as following:

   query =($word_1$,q ,$word_2$,q, …….,$word_n$,q)

   $document_j$=($word_1$,j,$word_2$,j,…….,$word_t$,j)

- A text document contains a list of key terms with their weights.
- D = ($term_1$;$weight_i$…… $term_n$;$weight_n$)
- Term = input query or search keyword enterd by user.
- Weight = it is the measure of importance of term expressed for information available in the specific Tex Document.
- Term frequency (tf): the count of repeated words relevant to content..
- Inverse document frequency (idf): uncommon term is more important
- Normalize the length of document
  - ✓ Large documents contain many distinct words.

✓ Large documents contain same word many times.

- Weight = term-frequency * inverse document frequency /normalization
- Measure vocabulary overlap between user query and documents. $T_1 \ldots T_n$

$$Q = q_1 \ldots q_n$$

$$D = d_1 \ldots d_n$$

$$Sim(Q,D) = \frac{\vec{Q}.\vec{D}}{\vec{Q}.\vec{D}} = \sum_i q_i \sum_i q_i * d_i d_i$$

Term -> $(Document_{i1}, word_{i1}) \ldots\ldots\ldots\ldots (D_{ik}, w_{ik})$

- Steps to calculate Cosine Similarity:

1 Take the dot product of vectors A and B.
2 Calculate the magnitude of Vector A.
3 Calculate the magnitude of Vector B.
4 Multiply the magnitudes of A and B.

- Divide the dot product of A and B by product of the magnitudes of A and B
- The similarity score can be calculated as:

$$(\cos\theta) = \frac{A.B}{||A||\,||B||} (\cos\theta) = \frac{A.B}{||A||\,||B||} = \frac{\sum_{i=1}^{u} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} X \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

$$\frac{\sum_{i=1}^{u} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} X \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

SEARCH ENGINE

A search engine is a system that is developed to retreive the inform ation stored on computer system. It is the largest source to use the data available on web. The search engines are of various types such as crawler-based search engines, web-based search engines, hybrid Search engines, Custom search, meta search engines, specialty search engine etc. The web search engine is very popular because they used to retrieve the information world wide web where other search engines works on personal search engines and intranet[7].

1. Crawler-based Search engines: Crawler-based search engines are use d automatic software programs to classify web pages. Search engine use some programs called spiders or crawler to access web pages. A crawler finds a web page or portable document format and downloads

it to analyze the content available on that page. Then the web pages are added to the databases. When user hits the query, search engine checks the database to retrive the list of links for the query or keywords. These search engines continually searching new web pages and update its database. Some crawler-based search engines are Google, Yahoo, Bing and Ask jeevs. The crawler used in case of Google is GoogleBot.

2. Web-based Search engines: Web Search engine is used to retrieve the information available world wide web. The results generated searching by queries are represented in line of results ang called search engine result pages (SERP). The information can be images, web pages, text documents and other files. The directories are used to categorize the web pages. The human directories are used to check information on web pages according to pre-defined rules and rank it[2][3].
3. Hybrid Search engines: These are the search engines which uses the different types of data to generate results based on web crawling. Earlier only text data was searched by these search engines. Hybrid search engine is a combination of the data searched by web search engine and Directories.
4. Meta Search engines: These are the search engines which uses the data searched by another searche engines and combine the results into one large list. The user inputs the query and queries are tranferred to third party search engines to obtain results and represented to users. MetaCrawler and Dogpile are the examples of Meta search engines.
5. Specialty Search engines: This search engine search the data from specially created database related to particular subject area. SearchDotNet, Yahooliqans, LawCrawler, MedHunt, CleanSearch Askjeeves are some popular search engines related to different subject areas. E.g LawCrawler is developed for legal profeesionals. MedHunt is related to medical field[3] [6].

Working of Search Engine

A search engine is designed to retrieve the information according to query or keyword hit by user. The information retrieved from queries are stored on WWW. The search engine retrieves a huge list of results that matches to the input terms. Search engine updates the information available on index servers so that the latest information can be retrieved in efficient way. The list of results generated by web servers are of two different types either the results are natural search or the paid search that is pay per click. When the user search search for any term and put the query into search box, large numbers of results are generated. The most relevant results matches to those queries are filtered and represented to users. The searching, indexing and ranking are the major functions of search engines to produce the results for input[1][4].

1. Crawling: It is a method of finding latest, updated and new pages to add into the google index. The detection and fetching of pages is done by crawler or spider. When a large amount of data is searched the work of crawling is done by Googlebot. It is a process that defines the how many pages are to be fetched form how many sites.

   The crawling process begins with list of URLs of web pages obtained from the different crawling processes. As the Googlebot crawler explores the websites it identifies the various links on each page and add them to the already crawled pages. The working of crawling process is shown in fig 1.2.

2. Indexing: web crawler processes each web page to explore an enormous directory of words and their location on each page. All the attributes of pages are processed. The web crawler can explores content of many types but sometimes it is unable to process some dynamic pages and media files[8].

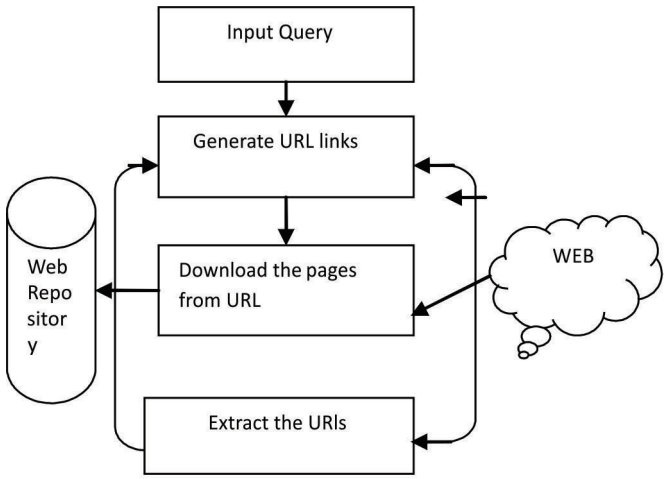

Fig 1.2: Process of crawling for web pages

Serving Results: When a keyword is searched in the form of query, the index of matching pages related to that query are found and most relevant results are returned to user.

3. Page Ranking: page ranking is process to provide weightage to the pages indexed by search engine. The relevancy of the web pages are depend on many factors, one of the important factors is page rank. Page ranking system is used by google search engine to check the relevancy of pages. To

determine the importance the web pages it counts the number and quality of link related to query.

The important websites are those websites which receives more links from other websites. The fig 1.3 is representing the process of search engine, how it search, index and rank the page[5][9].

IMPLEMENTATION AND RESULTS

To extract the results of Google search engine HttpUrlConnection, useragent or ApacheHttpClient are the different option to perform the task. A query parameter is a part of String Url where google search is an Http GET request. Jsoup is an open source HTML parser that is able to fetch the results in the form of urls as shown in fig 1.4.

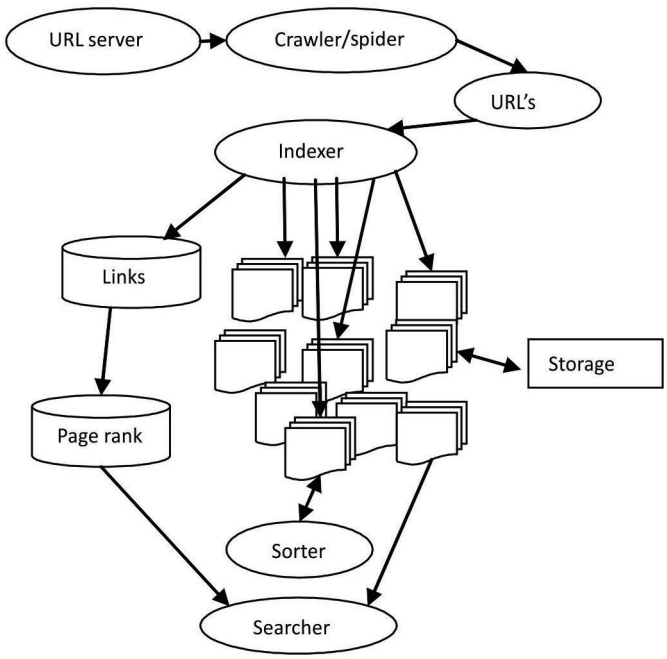

Fig 1.3 process of indexing and page ranking

User Agent : User Agent identifies the browser and the required details of computer system. These details are send to the servers of that pages that you are visiting. The details must be the version number of browser, operating system you are using. The Web server use this information to provide the required content related to particular browser.

The user can also enetred the number of results. This program can find the open source files available on the web and download that pages into database

shown in fig 1.4. Fig 1.5 represents the details of the files that are downlaoded from the urls extarcted. This can be achieved by the user agent Mozilla5.0.

Google search engine use different page algorithm then other search engines. A Google Search API is used to retrieve the results and HttpUtility Downloader is used to save that open source URL into Repository. Based upon that repository the comparison is being done between the text documents.

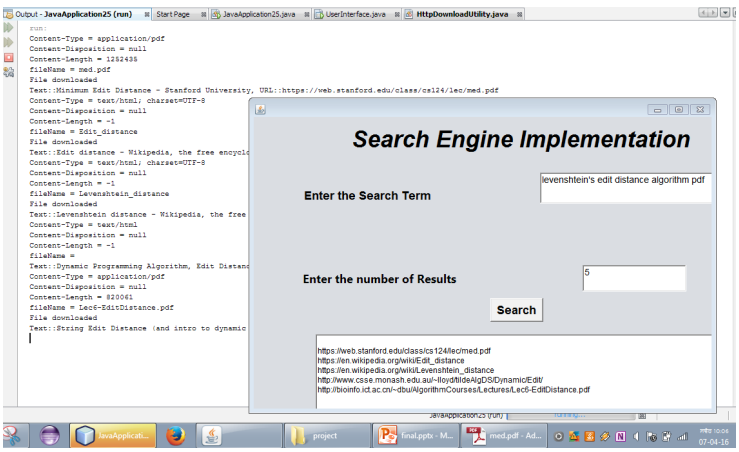

Fig 1.4 To fetch and download the files from open source URLs.

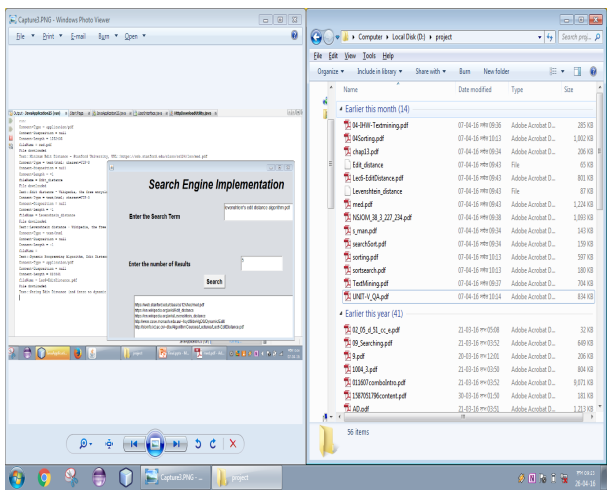

Fig 1.5 Downloaded files into database

COMPARISON WITH OTHER SEARCH ENGINES

Google provides more relevant information with comparison other search engines. As recent changes in Bing make it more Viable search option, it is

still lacking in terms of result quality in response to informative queries.

The most significant advantage on the Google internet search engine is maybe the sheer number of sites that indexes with comparison to yahoo search engine.

Giagblast was created to provide search engine functionality on the least amount of hardware possible at the current state of technology.

Fig 1.6 represents the comparisons of different search engines based on some defined factors.

Name of search Engine	**Pages Indexed**	**Daily direct queries**	**Dedicated servers**	**Personalized results**
Google	40 billon	320 million	yes	Default
Bing	13.5 billion	unknown	yes	Unknown
Yahoo	10 billion	unknown	partial	Unknown
Gigablast	>1billion	Unknown	---------	No

Fig 1.6 : comarison of various search engines.

CONCLUSION

There are many search engines are available but Google Search engine is very useful tool in present era of internet. The user agent helps to extract the most relevant results based on query. The Google search API is used to fetch the results generated by search engine implementation. It fetches the URL Links into database and then downloads the files linked to those URLs for the text comparison of that document.

REFERENCES

Beel, J., Gipp, B. and Wilde, . "*Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co.*". Accessible at: http:// www.beel.org/ files/papers/2010-ASEO--preprint.pdf (last accessed 24 August 2012); and Hoyt, Jason. Mendeley blog. 29 November 2010. Academic SEO – Market (and Publish) or Perish.

Madhu, G., Govardhan, A. and Rajinikanth, T. V. (2011) "*Intelligent Semantic Web Search Engines: A Brief Survey*" International journal of Web & Semantic Technology (IJWesT) Vol.2, No.1, January 2011.

Prakash, K. S. V. "*Concept of Search Engine Optimization in Web Search Engine*" International Journal of Advanced Engineering Research and Studies E-ISSN2249–8974.

Dirk Lewandowski "*New perspectives on Web search engine research*" Lewandowski, Dirk

(ed.): Web Search Engine Research. Bingley: Emerald Group Publishing, 2012.

Mike Grehan "*How Search Engines Work*" 2 publication of Search Engine Marketing: The Essential Best Practice Guide.

Bar-Ilan, J. (2004). *The use of web search engines in information science research*. In B. Cronin (Ed.), Annual review of information science and technology (Vol. 38, pp. 231- 288). Medford, NJ: Information Today, Inc

Mr.K. Tarakeswar and Ms. D. Kavitha "*Search Engines: A Study*" Journal of Computer Applications (JCA) ISSN: 0974-1925, Volume IV, Issue 1, 2011.

Mark Levene, "*An Introduction to Search Engines and Web Navigation*", John Wiley & Sons, Inc., 2010.

Sergey Brin and Lawrence Page "*The Anatomy of a Large-Scale Hyper textual Web Search Engine*".

Tom Seymour "*History of Search Engines"* International Journal of Management & Information Systems – Fourth Quarter 2011 Volume 15, Number 4.

Diana Inkpen "*Information Retrieval on the Internet*".