ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

A Survey on Web Search Engines

D.Keerthika[1] G.Sangeetha[2]

M.E Student, Dept of CSE, Valliammai Engineering College, Kanchipuram, India[1]

Assistant Professor, Dept of CSE , Valliammai Engineering College Kanchipuram, India[2]

ABSTRACT: Finding a information in today's world is easy and quick .This work is made easier by the use of system called Search Engine. Search Engine makes the user's task simpler and all information can be found in a more effective manner. Previous days the search will be made only through typing the keywords in search bar but now a days we can search a content using voice etc .This paper is a comparative study of different search engines and their methodology's used.

KEYWORDS: Search Engine, Information, Effective.

I. INTRODUCTION

Web mining is one of the techniques of data mining. And it is used to recognize the patterns from World Wide Web. It is further classified into three types. They are Web content Mining, Web Structure Mining, and Web Usage Mining. Web Content Mining is Mining, Abstraction, and Synthesis of fruitful data, information and knowledge from web page content. It is also called as Text Mining. In this mining whole web page is scanned to determine the relevance of content to keyword. Web Structure Mining is used to find the relationship between web pages related by information. This mining is used to generate structural outline about websites and web pages. Web usage mining is used to fetch the usage patterns from World Wide Web. It consists of three phases preprocessing, pattern discovery and pattern analysis.

Long days ago there were many search engines in market. But only some search engines will get a long reputation. A search engine is analyzed based on some of the characteristics like Quality, Scalability, and Efficiency. Quality is the quality of search results. Scalability is increasing the performance of search by scaling the infrastructure. Efficiency is how quickly the results are displayed to the user. Time complexity is one of the major factor. It shows how fast the results are displayed.

II. TERMS USED IN SEARCH ENGINE

Index: The search engine facts are stored in a document called "index". It is the place where the facts are collected and stored for search purpose. The plural form of index is called "indexes". It is also called as "catalog"

Keyword search: A search for a query containing one or more words combinations that are specified by a search engine user

Phrase search: A search for documents containing a absolute sentence or expression specified by a search engine user.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

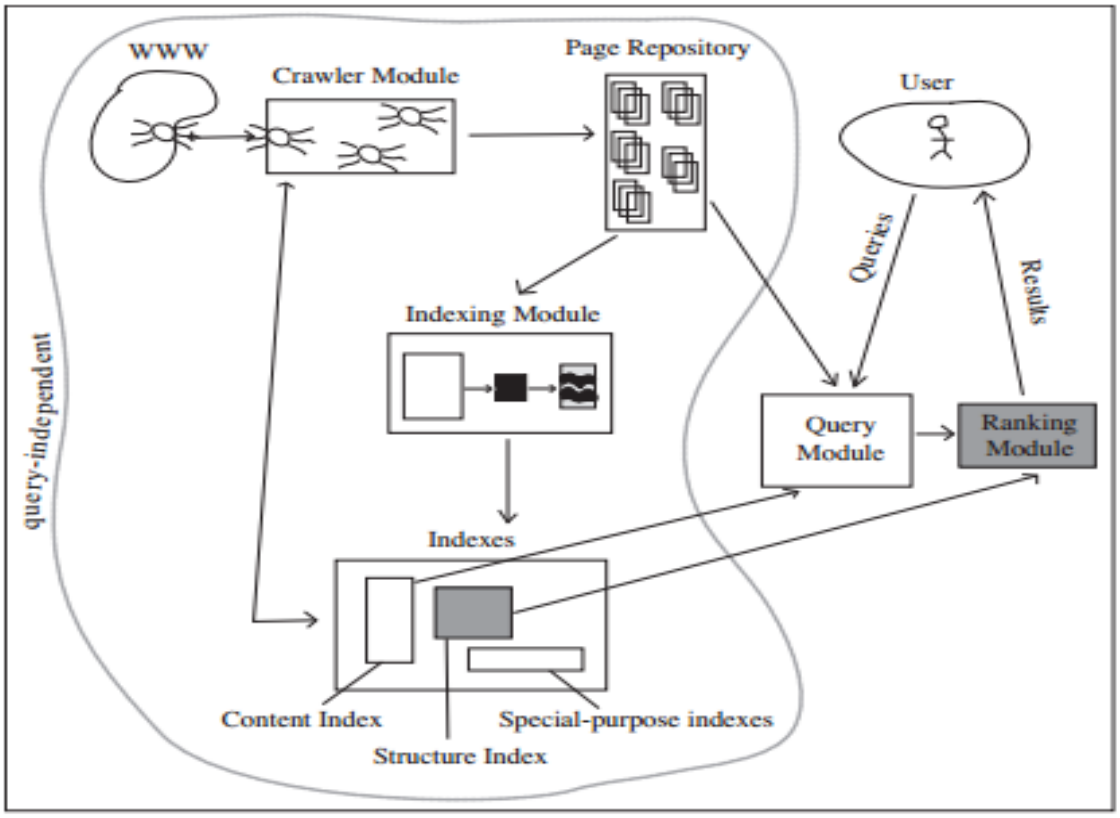


Fig.1 Basic Working of Search Engine

Proximity search: A search where search engine users to specify that documents returned should have the words close to each other.

Query-By-Example: A search where a search engine user instructs a search engine to find more documents that are akin to a accurate document. Also called "find similar".

Recall: It is the degree in which an engine returns all the identical documents in a acquisition. There may be 100 identical documents, but an engine may only find 80 of them. It will then list these 80 and have a recall of 80%.

Relevancy: How well a search result provides the information a search engine user is looking for, as measured by the user. It shows the degree of the results.

*Spider***:** It is software that scans documents and adds them to an index by following links. Spider is often used as a synonym for search engine. It is also called as bots.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Precision: The degree in which a search engine lists documents matching a search engine user query. If the matching percentage is higher it has higher precision. For ex. in a search results if the search query lists 85 results but only 30 of them satisfies the criteria then it has 20%precision.

*Stemming***:** The ability for a search to include the words at the start or end of the words. For example, stemming allows a search engine user to enter "reading" and get back results also for the stem word "read."

Stop words: Connectivity words, prepositions and articles and other words such as and, from and a that appear often in documents yet alone may contain little meaning.

Thesaurus: It is a list of equivalent phrase a search engine can use to find matches for particular words if the words themselves don't appear in documents.

Boolean search: A search allowing the inclusion or exclusion of documents containing certain words through the use of phrase such as AND, NOT and OR.

Concept search: A search for documents related conceptually to a phrase, rather than specifically containing the phrase itself.

Full-text index: An index containing every phrase of every document cataloged, including stop phrases (defined below).

Fuzzy search: A search that will find matches even when phrase are only partially spelled or misspelled[9].

III. ANALYSIS OF SEARCH ENGINES

A. *Archie:*

It is the first internet search engine .It was implemented in the year 1990 by Alan Emtage. It was a simple search engine that would keep a key of the file lists of all public FTP servers it could find. In this way, users would be able to find federally available files and download them. This provided a much better way to find files, as previously people could only know much about files. It is still used in University of Warsaw. It has different features such as customizing your search For example, besides being able to choose between “Anonymous FTP” and “Polish Web Index”, you can also choose whether your search coming in should be treated as:

- A sub string (as long as a part of the filename includes what you searched)
- An exact search (everything that doesn’t match the query exactly is rejected),and
- A regular expression

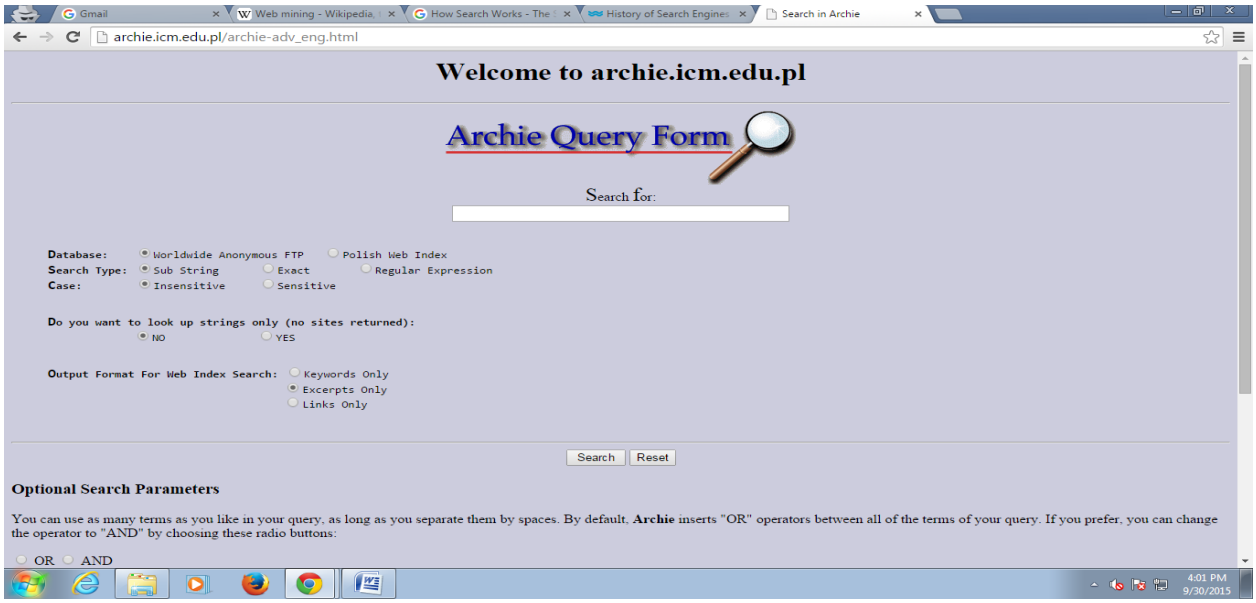
Other feature is the ability to search for strings rather than paths to files or websites. In other words, if this feature is enabled, it returns the filenames of what Archie finds, but not the actual place where the file was found so that you can download it. There are even three features for how the search results should be delivered, including keywords only, excerpts only, and connections only.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015



B.Gopher:

The **Gopher protocol** is a Transmission Control Protocol/Internet Protocol application layer protocol designed for distributing, searching, and get back booklets over the Internet. The Gopher protocol was strongly oriented towards a menu-document design and presented a replacement to the World Wide Web in its early stages, but ultimately HTTP became the dominant protocol. The Gopher ecology is often regarded as the forcible predecessor of the World Wide Web.

The protocol was invented by a group governed by Mark P. McCahill at the University of Minnesota. It offers some features not natively supported by the Web and levy a much stronger hierarchy on facts stored on it. Its text menu interface is well-suited to computing environments that depend heavily on remote text-oriented computer terminals, which were still common at the time of its creation in 1991, and the ease of its protocol facilitated a wide variety of client implementations. More recent Gopher revisions and graphical users added support for multimedia. Gopher was preferred by many network administrators for using fewer network supplies than Web services.

It's hierarchical structure provided a platform for the first full-scale electronic library connections. Gopher has been specified by some fans as faster and more efficient and far more classified than Web amenities. The Gopher protocol is still in use by fans, and although it has been almost entirely ejected by the Web, a small population of actively maintained servers remains [12].

Gopher Protocol:

/Reference
1CIA World Factbook /Archieves/mirrors/textfiles.com/politics/CIA gopher.quux.org 70
0Jargon 4.2.0 /Reference/Jargon 4.2.0 gopher.quux.org 70 +
1Online Libraries /Reference/Online Libraries gopher.quux.org 70 +
1RFCs:Internet Standards /Computer/Standards and Specs/RFC gopher.quux.org 70

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

1U.S Gazetteer /Reference/U.S Gazetteer gopher.quux.org 70 +
IThis file contains information on United States fake (NULL) 0
Icities,counties and geographical areas.It has fake (NULL) 0
Ilatitude/longitude,population ,land and water area,fake (NULL) 0
Iland ZIP codes fake(NULL) 0
iTo search for a city,enter the city's name .To search fake (NULL) 0
ifor a country,use the name plus Country—for instance, fake (NULL)0
iDallas Country.fake (NULL) 0

C.Google:

Larry page and Sergey joined together and worked on a search engine called Backrub to analyze the back links pointing to a given website January 1996. A year later it got reputed among the students of Stanford University. It uses Citation for ranking .Citation is related to ranking the pages based on the number of view count. In 1998, google was launched with the page ranking technology[2]. Google contains 600 trillion individual web pages. It uses crawlers to navigate from one page to other page. Then the pages are sorted based on their content and other factors. Then algorithms will be applied to find the clues based on the search word. Using that clues relevant documents are retrieved based on the index. Then Google rank the results based on two hundred factors such as freshness, Quality, safe search, user context etc…Then the results are translated based on the end device. Google takes $1/8^{th}$ of second to do this. To keep the results relevant the Spam's are removed automatically and website owners will be reported about the spam[1]. The Google search engine architecture is shown in Fig.2.[8].

Page Rank :

Page Rank is a numerical value that represents the importance of a page present on the web. When one page links to another page, it is effectively molding a vote for the other page. Page Rank Notation- "PR"[10].

The original Page Rank algorithm which was specified by Larry and Sergey is given by PR(1) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) where,

PR(1) – Page Rank of page 1

PR(Ti) – Page Rank of pages Ti which link to page 1

C(Ti) - number of outbound links on page Ti

d - damping factor which can be set between 0 and 1[3].

Page rank Matlab code :

```
%Parameter M adjacency matrix where M_i,j represents the link from 'j' to 'I' such that for all 'j'
%sum(I,M_i,j)=1
%Parameter d damping factor
%Parameter v_quadratic_error quadratic error for v
%Return v,a vector of ranks such that v_i is the i-th rank from[0,1]
Function[v]=rank2(M,d,v_quadratic_error)
N=size(M,2);%N is equal to half size of M
```

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

```
V=rand(n,1);
V=v./norm(v,1);%This is now  L1,not L2
Last v=ones(N,1)*inf;
M_hat=(d*M)+(((1-d)/N).*ones(N,N));
While(norm(v-last_v,2)>v_quadratic_error)
Last_v=v;
V=M_hat*v;
%removed the L2 norm of the iterated PR
End
endfunction
```

Page Rank algorithm rank the pages with the value of 0 to 10.The first and top ranking page will get the rank as 10.The page which gets less rank will have the value of zero. Page rank increases the popularity of the web pages[1]

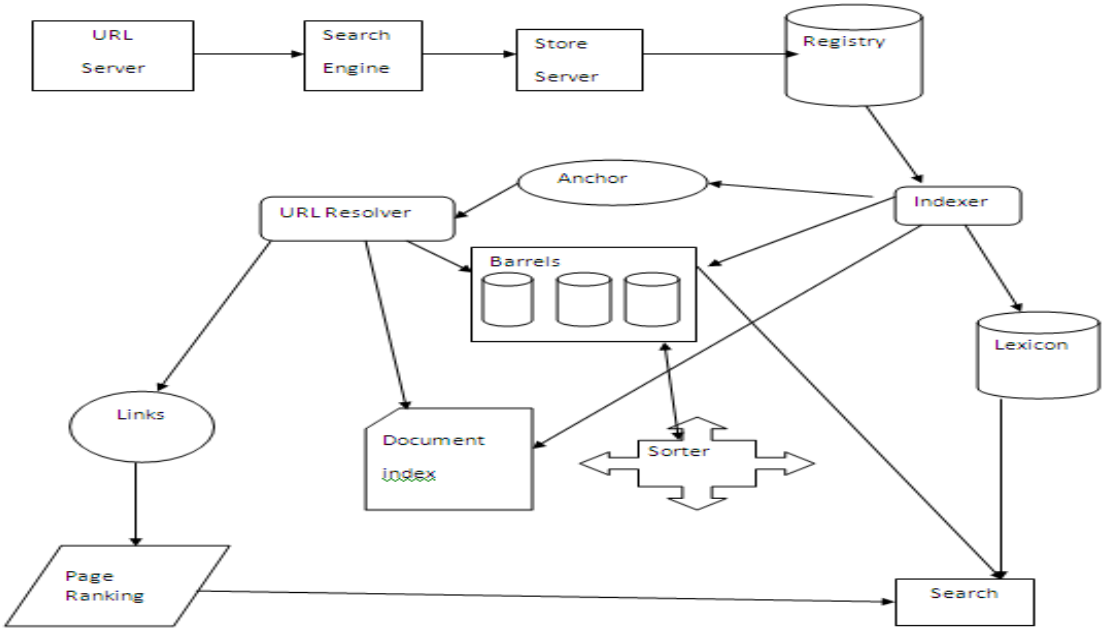


Fig.2 Google's search engine architecture

D.Bing:

Bing is a web search engine and it was developed by Microsoft. It is also called as live search, windows live search and MSN. It was announced by Microsoft CEO on May 28 2009. On July 29, 2009, Microsoft and Yahoo! announced a contract in which Bing would control Yahoo! Search. All Yahoo! Search global customers and partners made the transition by early 2012. In September 2013, a new-look Bing was released to tie in with Microsoft's "Metro" design language. As of February 2015, it is the 2nd largest search engine in the US with a query volume at 19.8%, while Yahoo Search, which Bing powers, has 12.8%. Its entrant Google is at 64.5%[4].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Bing Working:

Bing is working based on two factors .One is relevancy calculation and other is click distance. It takes each and every document in the web and parses it. The parsed documents are reduced to roots and useless words like articles will be removed .Each word produce a hash value and it will be stored in the frequency table. When a user enters a query, each word will be splitted and reduced to roots and useless words will be removed. Now each word produces hash value and it will be found in the frequency table. If it is found it is called essential page. Otherwise new documents will be found using selection Algorithms. The old and new document will be compared a. The old will be replaced .The essential pages will be ranked for next process [5].

The next step is calculating the click distance and URL depth will be measured.Bing uses click distance to compute how many clicks it takes to get from the homepage of the website to the page with the query on it. Pages closer to the homepage are treated more important, while pages even more from the homepage are less important. However, if the page has a Uniform Resource Locator depth of 4 (meaning the number of backslashes in the URL), the page is not likely to be terribly important, although the fact that it may have a link to itself on the homepage of a website. Thus, the URL depth can correct, reinforce, or refine the click distance. The click results and relevancy score is used to calculate the final score[6].

S.NO	GOOGLE	BING
1	It focuses on context of the words	It focuses on individual words
2	It uses Page Ranking algorithm	It focuses on number of back links.
3	Bots read through entire page.	Bots do not read through entire page.
4	It focuses on new content	It focuses on old content.

Table 1 –Comparison of Google and Bing[5]

E.ASK :

Ask.com (initially named as **Ask Jeeves**) is an interrogate answering-focused web search engine founded in 1995 by Garrett Gruener and David Warthen in Berkeley, California. The Ask toolbar is a web-browser add-on that can appear as an additional bar added to the browser's window and/or menu. It cannot be easily removed by using built-in uninstall features, therefore it is termed as "nonessential instructions". Whilom toolbar has been installed it captures the control of clients home page and relays the results to this search site. Another drawback of ask is delay of ten minutes for updation. It can be excluded by using Ask.Com Help Center. Norton antivirus has an ask engine for “Safe Search” toolbar. If the user did not install that tool then popup will be shown to get the tool. The old versions of Ask is a way of threat to PC and provide tools to search and expel them. In 2010, it discarded the search industry, because it could not with stand in the market against other search engines like Bing, Google etc…Now it has turned to be an answering engine and it outsourced its company to unnamed third party service provider[11].

F.Yahoo :

Yahoo is one of the oldest search engine and it was founded on 1994.It provides search results to over 350million peoples a month. It also provides other services like Yahoo News, Yahoo Mail. Yahoo ranging algorithm is similar to Google ranking algorithm but it concentrates more on the directory[7]. Yahoo’s Ranking Algorithm depends more on heading of the website. Its first ranking criteria are the heading of the website. The description of the webpage is the next criteria. Yahoo concentrates more on click through rate where Google does not concentrate on this criterion. When a website gets more rates it gets the first rank in the page. The characteristic of 1) Yahoo is more locuses on topic and region.2) It adds word at

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

the start or end of query in order to do automatic truncation.3) Yahoo is not case sensitive. The yahoo search engines architecture is shown in Fig.3[8].

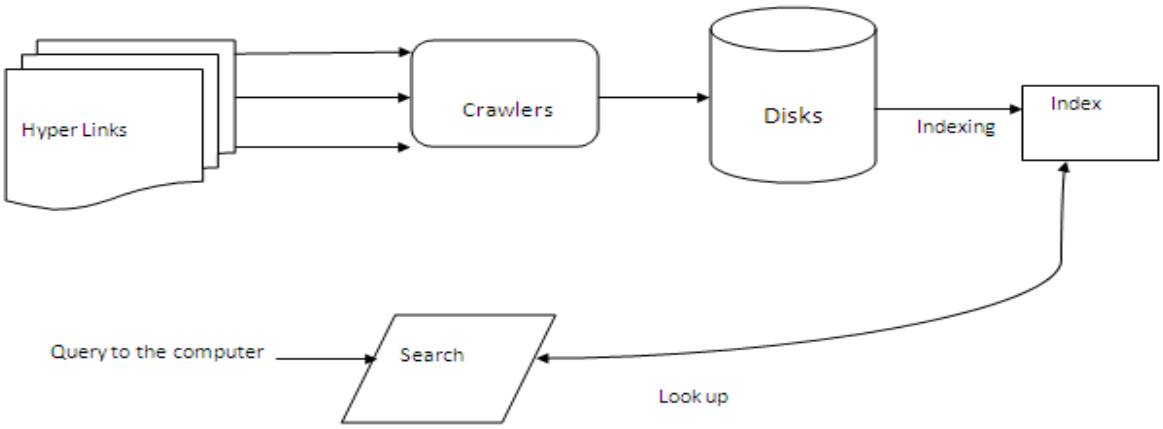


Fig.3 Yahoo Search Engine Architecture

IV. CONCLUSION

In this paper we studied the features, history and working of popular search engines like Google, Ask, Bing, and Yahoo. The basic function of all Search Engine is the same. Each search engine uses its own ranking techniques which differs one from other. But the search results and the efficiencies will differ depending upon the methodologies used. Further research can be extended to analyze the search engines.

REFERENCES

1. Google - http://www.google.co.in/insidesearch/howsearchworks/thestory/
2. http://www.searchenginehistory.com/#google
3. Google's Page Rank and Beyond by Amy N.Langville and Carl D.Meyer
4. https://en.wikipedia.org/wiki/Bing
5. http://scenic.princeton.edu/network20q/blog/?p=811
6. http://www.searchenginejournal.com/6-ways-bing-opposite-google/128935/
7. http://www.irkawebpromotions.com/search-engines/yahoo/
8. Amjad J. Khalil, Fadi K. Abu Alrub, "A COMPARISON OF SEARCH ENGINE's FEATUREs and MECHANIZMs", 2013.
9. Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hyper textual Web Search Engine".
10. M.Sangeetha, Dr.K.Suresh Joseph, "Page Ranking Algorithms used in Web Mining",2014
11. Brijendra Singh, Hemant Kumar Singh ,"WEB DATA MINING RESEARCH: A SURVEY ",2010.
12. https://en.wikipedia.org/wiki/Gopher_%28protocol%29