5

UNIT - V

Web Searching

Syllabus

Introduction, Challenges, Web Characteristics, Search Engines : Centralized Architecture, Distributed Architecture, User Interfaces, Ranking, Crawling the web, Indices, Browsing, Meta-searchers, Searching using Hyperlinks, Trends and Research Issues, Introduction to Web Scraping : Python for web scraping, Request, HTML parsing, Beautiful Soup.

5.1 Introduction

- The World Wide Web which is abbreviated as WWW or W3 and commonly known as the Web.
- The World Wide Web is a system or service of interlinked hypertext documents accessed via the internet.
- With a browser one can access web pages. Web pages contain text, images, videos, and multimedia information's.
- Web pages also allow us to navigate between different web pages using hyperlinks.
- Tim Berners Lee created the World Wide Web. The first trials of the World Wide Web were at the CERN laboratories (one of Europe's largest research laboratories) in Switzerland in December 1990.
- By 1991 browser and web server software was available, and by 1992 a few preliminary sites existed in places like University of Illinois, where Mark Andreessen became involved. By the end of 1992, there were about 26 sites.
- As we know that with World Wide Web large amount of textual data and other media data are available so we can say that World Wide Web is very large, unstructured but ubiquitous database.
- As it is a database we need some efficient tools to manage, retrieve and filter information from such large database. Similarly this problem is very important in large intranets, where we need to extract information for data mining.

5.2 Searching the Web

University Questions

Q. Discuss different forms of searching the web. Explain with proper example. **SPPU : May 12, 8 Marks**

Q. List different forms of searching the web. **SPPU : May 14, 4 Marks**

- In today's modern word every person is using computer. And similarly if you want some information regarding anything you go for web search.
- So web searching has become a part of day-to-day life of any person seeking for any information.
- Searches vary from words, phrases, products, people, images, services, videos and many more forms.

- Web searching is searching for the information on the World Wide Web.
- For effective web searching there are few basic skills that you can learn. These skills help you to make search more successful and less frustrating.

1. If you are looking for a specific phrase then you can use quotation marks that will help search engine only bring back pages that includes search terms exactly how you typed them in-order.
2. You can use Google to search within a site as most of the search sites are not that much great like Google.
3. You can use the In url syntax as it allows searching words within the Uniform Resource Locator. It also helps us to search web and web sites that you will not find by just entering in a query word or phrase.
4. Use basic math to search more information on the web.
5. You can use Google cheat sheet shortcuts as it will help to locate whatever you are searching for.
6. If you want to limit your search to a specific domain like .gov, .in, .edu, .org and so on you can use the site: command.
7. You can use different sites like Google or yahoo to get information about weather forecast instead of waiting for the news reports.
8. We know that every engine returns different results. Similarly there are some engines that focus on games, blogs, forum, and book. So you can use different search engines for different searches.
9. You can use the web to find out definitions, synonyms or anonyms of different words.
10. We know that toolbars are the software applications that give searches the ability to perform searches and other functions very quickly. So you can use different toolbars while searching on the web to save lot of time.

- There are three different types for searching the web as shown in Fig. 5.2.1.

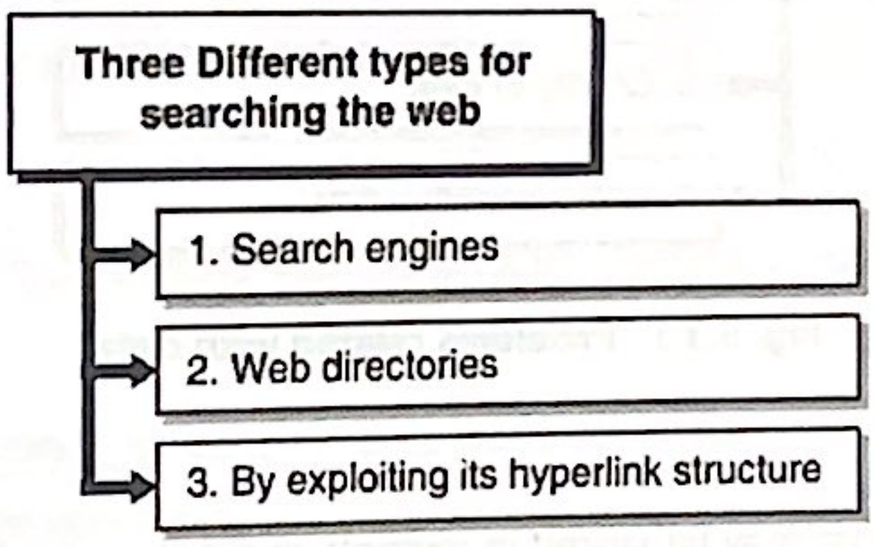

Fig. 5.2.1 : Types for searching the web

1. Search engines

A search engine also called as search service is a document retrieval system which is designed to help to find information which is stored on a computer system, such as on the World Wide Web, inside a corporate network, or in a personal computer.

2. Web directories

Web directory lists the web sites by category and subcategory. And this categorization and sub categorization is based on web site rather than web pages or set of keywords. Most web directory entries are also not found by web crawlers but by humans.

3. By exploiting its hyperlink structure

The first two are very well known and are frequently used.

5.3 Challenges

University Questions

Q. Write short note on Challenges in web search. **SPPU : May 13, 16 Marks**

Q. Explain Challenges of searching the web. **SPPU : May 14, 4 Marks**

Q. Discuss challenges involved in web searching. **SPPU : May 16, 8 Marks**

Q. What are the challenges while searching the web? **SPPU : Dec. 16, 12 Marks**

- Here we will see the main problems faced by web. These problems are divided into two categories :
  1. Problems related with data
  2. Problems regarding the user and his/her interaction with the information retrieval system.
- First we will see the problems related with data as shown in Fig. 5.3.1.

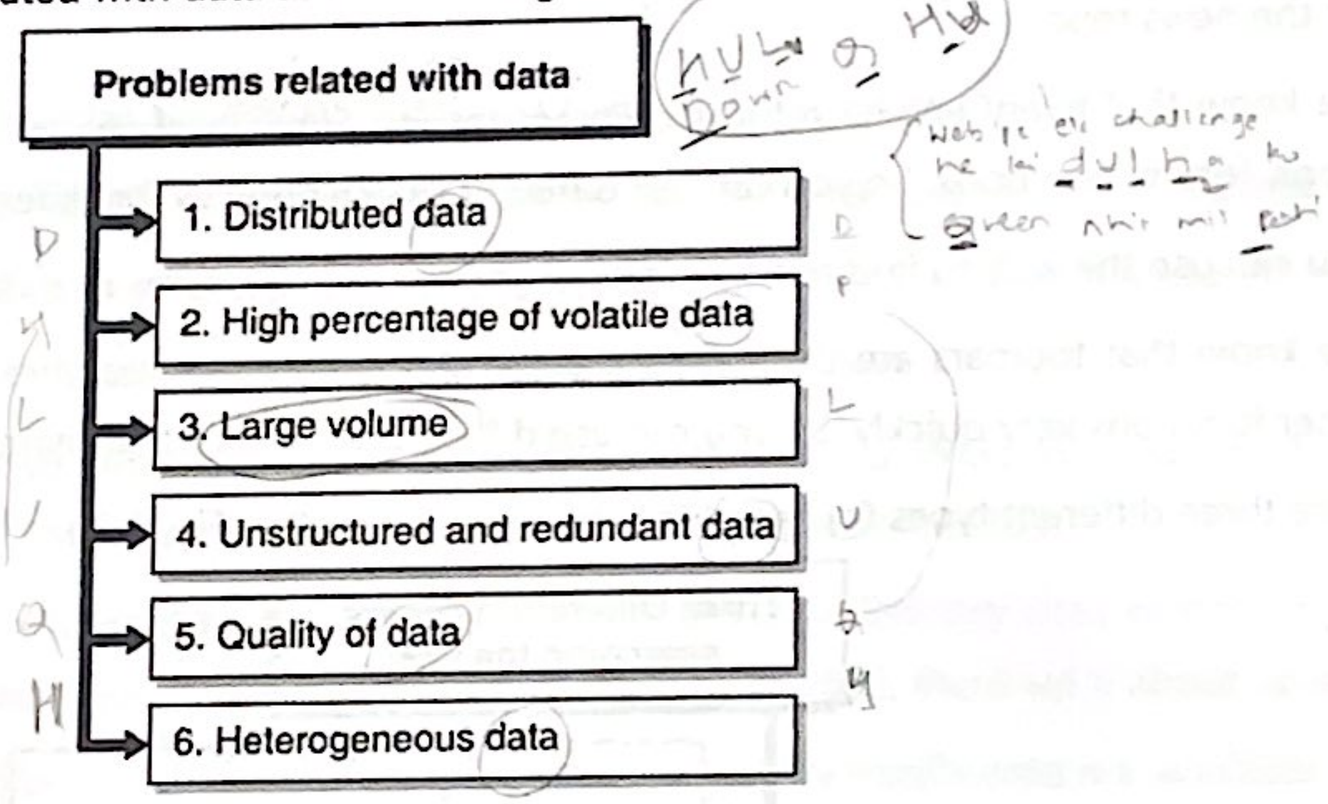

Fig. 5.3.1 : Problems related with data

1. Distributed data

- Distributed data is the data that may be stored in multiple computers or platforms.
- These computers may be located in the same physical location or these computers may be dispersed over a network of interconnected components. The interconnection of computer has no predefined topology.
- Similarly the available bandwidth and reliability on the network connection varies widely time to time.

2. High percentage of volatile data

- Volatile data is the kind of data that changes frequently. We know that internet is very dynamic in nature i.e. new data or information may be added or removed easily.
- Similarly new computers can be added or removed easily. If the domain or file names changes in the internet then we can face the problem of dangling link and relocation.

3. Large volume

We know that web is growing exponentially time to time. And due to this exponential growth scaling issue arises in the web. So it is very difficult to handle this scaling problem.

4. Unstructured and redundant data

- Unstructured Data or unstructured information is the kind of data or information that either does not have a pre-defined data models or does not fit well into relational tables.
- Some people say that web is a distributed hypertext. But which is not exactly true.
- Because in case of hypertext conceptual model is used and this conceptual model adds consistency to the data and the hyperlink. And we know that this is not always true in web. And even this is not true for individual documents. Similarly each HTML page is not well structured or semi-structured.
- Redundant data is nothing but data duplication. That is same data of same organization is replicated on multiple site. And we know that web data is replicated.

5. Quality of data ACID

- Data quality is an assessment of data such that how it will fit to serve the required purpose in the given context.
- There are some aspects of data quality that includes accuracy, completeness, update status, relevance, and consistency across data sources, reliability, and appropriate presentation accessibility and so on.
- In the organization data quality is compulsory to operational and transactional process.
- Maintaining data quality requires scanning the data periodically, updating standardizing and de-duplicating.
- As web is used for publishing purpose, data can be false, invalid, poorly written with many errors from different sources. And errors can be typo, grammatical mistakes.

6. Heterogeneous data

- Heterogeneous Data is the type of data which is from any number of sources. These sources can be unknown and unlimited.
- Similarly they can be in various formats. We can also have different languages, different alphabets. These languages and alphabets are very large.
- So most of the problems discussed above are not solvable by just improving the software implementation.
- Similarly many of them will not change as these problems are very inherent to human nature.
- The second type of problems regarding the user and his/her interaction with the information retrieval system.
- Actually these are the problems which are faced by user while interacting with the information retrieval system.
- There are basically two kinds of problem related with it :
  1. The different ways used by the user while specifying the query.
  2. The different ways used by the user while interpreting the answers provided by the information retrieval system.

- If the query is very simple then we can specify it very precisely but if it is not simple then it is very difficult to specify the query.
- Even though the user specifies the query very precisely then he will get the answer with thousand of web pages.
- So again there is a problem that how to handle such large answer with multiple pages.
- Similarly how to rank the document of answers? How to select the document which is actually required for the user or to select the document in which user is interested? And how to browse efficiently and quickly in large document?
- So in addition to the above discussed problems faced by web, main challenge is how to submit the good query to the search system and how to obtain the manageable and relevant answers from the multiple documents

5.4 Characterizing the Web

University Questions

Q. How do you characterize the web ? **SPPU : Dec. 12, 8 Marks**

Q. Write short note on characterizing the web. **SPPU : Dec. 13, May 16, 8 Marks**

- The characterization of web is a new task of the web consortium.
- We will see the following characteristic related with the web as shown in Fig. 5.4.1.

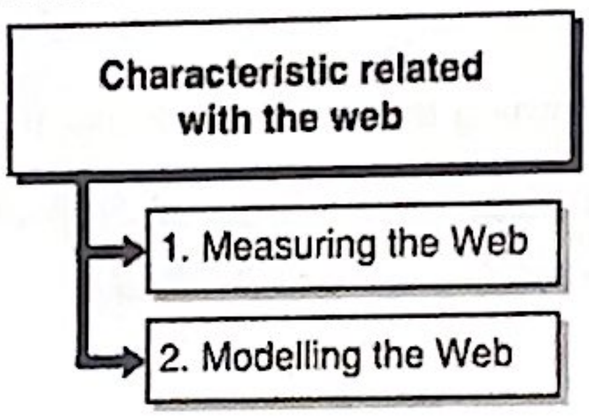

Fig. 5.4.1 : Characteristics related with web

5.4.1 Measuring the Web

- We know that internet and in particular web is very dynamic in nature.
- Due to this dynamic nature it is very difficult to measure it. There are more than 40 million computers in more than 200 countries which are connected to the internet. And many of them are hosting the web servers.
- The estimated number of web servers ranges from 2.4 million according to Netsizer and to over three million according to Netcraft web survey.
- There are many web sites which share the same web servers using virtual hosts, but all of them are not accessible.
- For the second half of 1993, the Web had a doubling period of less than 3 months, and even today the doubling period is still nearly 5 months.
- Additionally, the percentage of Web sites that are commercial has increased dramatically.
- The numbers below show the percentage growth in the com domain, which excludes foreign commercial sites such as the .co.uk domain, etc. This is given by Matthew Gray of MIT.

Month/Year	Web Sites	% of .com sites	Hosts Per Web Server
6/93	130	1.5	13,000
12/93	623	4.6	3,475
6/94	2,738	13.5	1,095
12/94	10,022	18.3	451
6/95	23,500	31.3	270
12/95	100,000	50.0	94

- Other estimation were made by sampling 0.1% of all internet numeric addresses obtaining about 2 million unique web sites or by counting domain names starting with www which in July 1998 were 780,000 according to the internet domain survey.
- However not all web servers have this prefix so real number is even larger. Considering that in July 1998 the number of internet hosts was estimated 36.7 million there is about one web server per every ten computers connected to the internet.
- Bray and Woodruff et al. studied different statistical measures of the web. The first study uses 11 million pages while second study uses 2.6 million pages.
- Estimates at the beginning of 1998 ranges from 200 to 320 million with 350 million as the best current estimate. The later study used 20,000 random queries based on a lexicon of 400,000 words extracted from yahoo.
- Those queries were submitted to four search engines and the union of all the answers covered about 70% of the web. Between 1997 and 1998, the size of the web doubled in nine month and is currently growing at a rate of 20 million pages per month.
- It is estimated that the 30,000 largest web sites account for approximately 50% of all web pages.
- HTML is the most popular document used for web with GIF and JPEG images, ASCII text, and Postscript. While Zip, GNU zip and compress are the most popular compression tools used for web document.
- Most of the HTML pages are not standard, not comply with all HTML specification. Even though HTML pages are instance of SGML.
- HTML document start with formal document type definition and similarly they are small usually contain few images.
- An average page has between 5 and 15 hyperlinks and most of them point to their own web server hierarchy.
- No external server points to any given pages. In 1995 around 80% of these home pages had less than ten external links pointing to each of them.
- There have been three studies regarding the language used in web pages. The first study is done by Funredes from 1996 to 1998. This uses Alta Vista search engine and is based on searching different words in different languages.
- But this technique is not statistically. The second study was done by Alis Technology and is based on automatic software that can detect the language used.

- The results obtained with this are consistent. The third study was done by OCLC in June of 1998 by sampling Internet numeric address and using the SILC language identification software.
- The variations for Japanese might be due to an inability to detect pages written in Kanji. Some languages are growing fast and the total number of languages exceeds 100.

5.4.2 Modelling the Web

- Modelling the web is addressing the specific issues related to design and development of large-scale web applications.
- It focuses on the design notations and visual languages that can be used for the realization of robust, well-structured, usable and maintainable Web applications.
- Designing a data-intensive Web site amounts to specifying its characteristics in terms of various orthogonal abstractions.
- Table 5.4.1 shows the languages of the web.

Table 5.4.1

Language	Funredes (1998, %)	Alis Tech. (June 1997, %)	OCLC (June 1998, %)	Spoken by (millions)
English	76.4	82.3	71	450
Japanese	4.8	1.6	4	126
German	4.4	4.0	7	118
French	2.9	1.5	3	122
Spanish	2.6	1.1	3	266
Italian	1.5	0.8	1	63
Portuguese	0.8	0.7	2	175

- There are no experiments on large web collections to measure the parameters like β and θ.
- There is a model which is related to the distribution of document sizes. According to this model document sizes are self-similar i.e. they have larger variance.
- This can be modelled by two different distributions. The main body of the distribution follows a logarithmic distribution such that the probability of finding a document size x bytes is given by -

$$P(x) = (1/x\sigma sqrt(2\pi)) \exp - (\ln x - \mu)^2 / 2\sigma^2$$

Where the average μ and standard deviation σ are 9.357 and 1.318 respectively.

- The majority of the documents are small, but there is a non-trivial number of large documents. The good fit is obtained with Pareto distribution

$$P(x) = \alpha k^{\alpha} / x^{1+\alpha}$$

Where x is measured in bytes and k and α are parameters of the distribution.

- For text files α is about 1.36 and it is smaller for images and binary formats.

- Taking all web document into consideration we can have $\alpha = 1.1$ and $k = 9.3$ Kb. That is 9.3 Kb is the cut point between both distribution and 93% of all the files have a size below this value.
- For less than 50 Kb images are the typical files from 50 to 300Kb we have number of audio files and over several megabytes video files are more frequent. Recent data collection in 1998 shows that the size distributions have the same form but parameters are changing

5.5 Search Engines

- The World Wide Web has revolutionized the way that people access information, and has opened up new possibilities in various areas such as digital libraries, scientific and general information retrieval, education, commerce, entertainment, government and health care.
- We know that web search services are now a major source of information for a growing number of people. So use such services we need some mechanism to access information.
- And web search engines are designed to search for information's on the World Wide Web and FTP servers.
- A search engine also called as search service is a document retrieval system which is designed to help to find information which is stored on a computer system, such as on the World Wide Web, inside a corporate network, or in a personal computer.
- Search engines are very different from subject directories. Directories are organizes by human while search engines depends on computer programs called spiders or robots to crawl the Web and log the words on each page.
- Using search engine, keywords related to a topic are typed into a search box then the search engine scans its database and returns a file with links to websites containing the word or words specified.
- As the databases are very large, search engines often return thousands of results. So it is not easy to find needed information without search strategies or techniques.
- Search engines use regularly updated indexes to operate quickly and efficiently. The very first tool used for searching on the Internet was Archie.
- The name stands for "archive" without the "v". It was created in 1990 by Alan Emtage, a student at McGill University in Montreal.
- Here we will discuss the different architectures of information retrieval system that model the web as full text database. There are differences between standard information retrieval system and the web.
- One of that is in the web all queries should be answered without accessing the text. In case of information retrieval system answering queries require either storing them locally or accessing remote pages through the network at query time.
- And this approach is very slow as time is spent in storing and all.
- Similarly this affect on indexing and searching algorithms as well as on query languages.

5.5.1 Challenges in Web Search Engine

University Question

Q. Discuss the challenges involve in web search engine. **SPPU : May 19, 8 Marks**

Following are the challenges in designing the web search engine.

1. **Spam :** It is observed that user of web search engine are interested mostly to see on the first of the search results. Research showed that for 85% of the queries, only first page of the result screen is requested. Mostly 10 results are shown on the first page as the result of such queries. In these cases, the page rank of the first pages on these 10 results are much more compared to other ranked results and other pages. Hence, the commercially oriented sites are interested to get listed in the first 10 results of the search.
2. **Content Quality :** It is the ethical guideline that the web pages should contain the true and accurate information. All documents should contain high quality data. However, in realistic situation, quality is the documents is unpredictable and there is question regarding the trust of data displayed on the page.
3. **Quality Evaluation :** The search result quality depends on the ranking algorithm used. Commercial search engines take care about the quality of the results produced by the ranking algorithm but this does not happen with every search engine. Use feedback plays important role in improving the quality of the results but most of the times users are not interested to give the feedback.
4. **Duplicate Hosts :** Web search engines try to avoid duplicate pages or near-duplicate pages as the part of query result. But, the problem becomes more trackable if we limit it to finding duplicate hosts, that is, two hostnames that has the same content.
5. **Vaguely-Structured Data :** The degree of structure present in data has had a strong influence on techniques used for search and retrieval. At one extreme, the database community has focused on highly-structured, relational data, while at the other the information retrieval community has been more concerned with essentially unstructured text documents. Of late, there has been some movement toward the middle with the database literature considering the imposition of structure over almost-structured data. Document management systems use accumulated meta-information to introduce more structure. The emergence of XML has led to a flurry of research involving extraction, imposition, or maintenance of partially-structured data.

5.5.2 Centralized Architecture

University Questions

Q. Explain the Crawler-indexer architecture. **SPPU : May 12, May 14, May 16, 8 Marks**

Q. Explain centralized search engine architectures. **SPPU : May 13, May 17, 4/5 Marks**

- Crawlers are the software agents that traverse through the web and sends new or updated pages to main server where from where they are indexed.
- Crawlers are also called as robots, spiders, wanderers, walkers and knowbots. Crawlers does not move to remote machine or it does not run on remote machine instead of that it runs on local machine and it sends requests to remote web servers. It uses the index in centralized fashion to answer queries submitted from different places in the web.
- Most of the search engines use centralized crawler-indexer architecture. Fig. 5.5.1 shows the crawler-indexer architecture which is based on the AltaVista architecture. In 1998 AltaVista system was running on 20 multi-processor machines.
- All of these machines are having 130 GB of RAM and 500 GB of hard disk space. And out of that query engine uses more than 75% of these resources.

- It has two parts one that deals with user and second part consists of the crawler and indexer modules.
- The first part that deals with user consists of user interfaces and query engine.
- The crawler module extracts URLs appearing in the retrieved pages, and gives this information to the crawler control module. This module determines what links to visit next, and feeds the links to visit back to the crawlers.
- The crawlers also pass the retrieved pages into a page repository. Crawlers continue visiting the Web, until local resources, such as storage, are exhausted.
- The indexer module extracts all the words from each page, and records the URL where each word occurred. The result is a generally very large.
- The query engine module is responsible for receiving and filling search requests from users.
- The engine relies heavily on the indexes, and sometimes on the page repository.
- Because of the Web's size, and the fact that users typically only enter one or two keywords, result sets are usually very large.
- The ranking module therefore has the task of sorting the results such that results near the top are the most likely ones to be what the user is looking for.

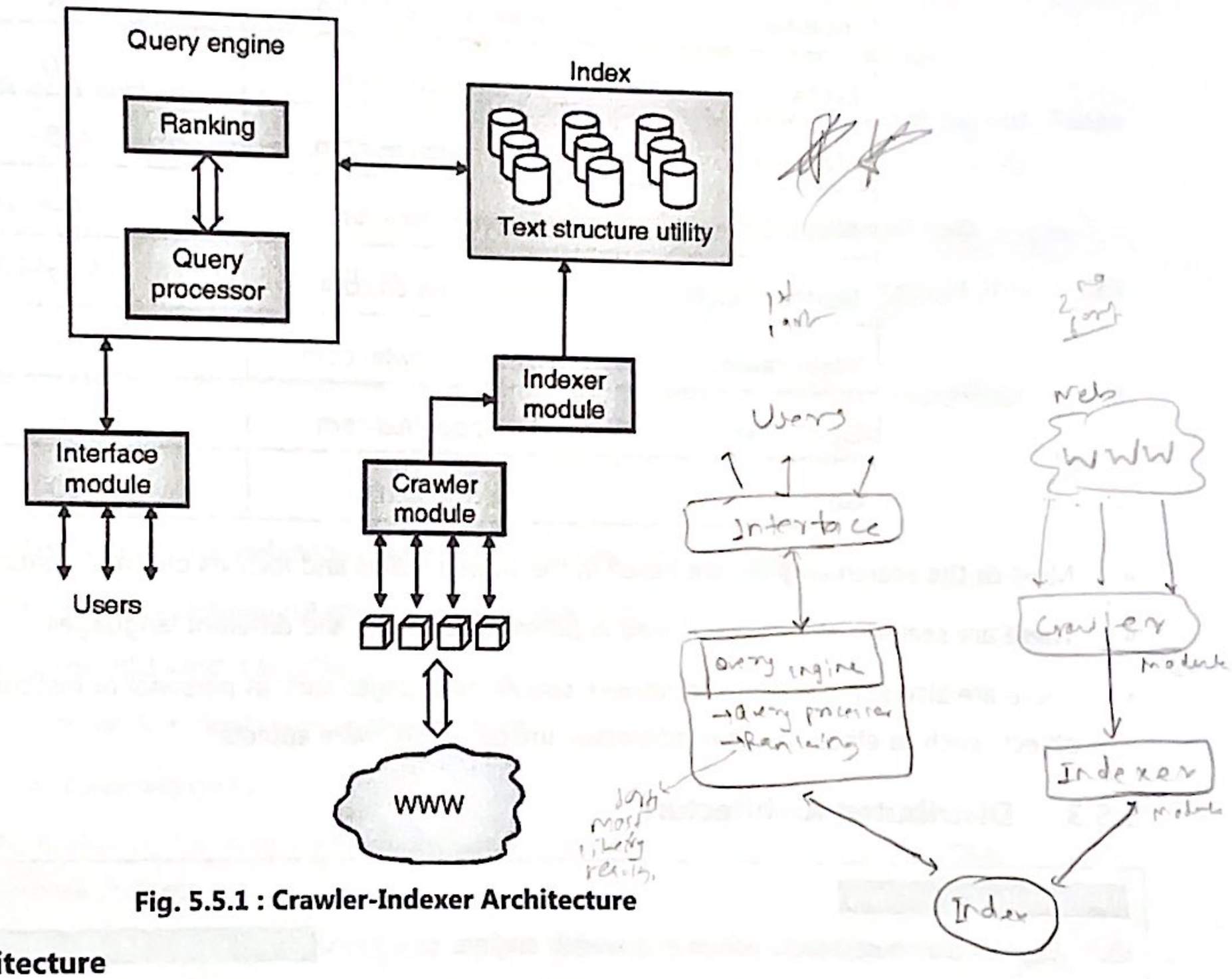

Fig. 5.5.1 : Crawler-Indexer Architecture

Disadvantage of this Architecture

1. Difficult to gather data due to dynamic nature of web.
2. Saturated communication links
3. High load on the servers.
4. Volume of the data. That means the crawler-indexer architecture is unable to deal with web growth in near future
5. Not that much good balancing between the different activities of search engine internally and externally.

- By considering web coverage in June 1998 the AltaVista, HotBot, Northern Light and Exite are the largest search engines; these engines cover 28-55% or 14-34% of all web pages.
- Table 5.5.1 shows the most of the search engines and their estimated size and corresponding URL.

Table 5.5.1

Search Engine	URL	Web pages indexed
AltaVista	www.altavista.com	140
AOL Netfind	www.aol.com/netfind	–
Excite	www.excite.com	55
Google	google.standford.edu	25
GoTo	goto.com	–
HotBot	www.hotbot.com	110
Infoseek	www.infoseek.com	30
Lycos	www.lycos.com	30
Magellan	www.mckinley.com	55
Microsoft	search.msn.com	–
Northern Light	www.nlsearch.com	67
WebCrawler	www.webcrawler.com	2
Open Text	www.opentext.com	–
Go	www.go.com	30

- Most of the search engines are based in the United States and focuses on documents in English.
- There are search engines specialized in different countries and different languages.
- There are also some engines to retrieve specific web pages such as personal or institutional home pages or specific objects such as electronic mail addresses, images or software applets.

5.5.3 Distributed Architecture

University Questions

Q. Explain distributed architecture of a search engine. **SPPU : Dec. 14, May 15, Dec. 15, May 17, 5/8 Marks**

Q. Comment on drawbacks of harvest distributed architecture. **SPPU : May 13, 4 Marks**

- There are several variations of the crawler-indexer architecture. One of them is Harvest architecture. Harvest architecture is more efficient than the crawler architecture.
- This architecture is based on distributed architecture for gathering and distributing data.
- Fig. 5.5.2 shows the one of the example of Harvest architecture.

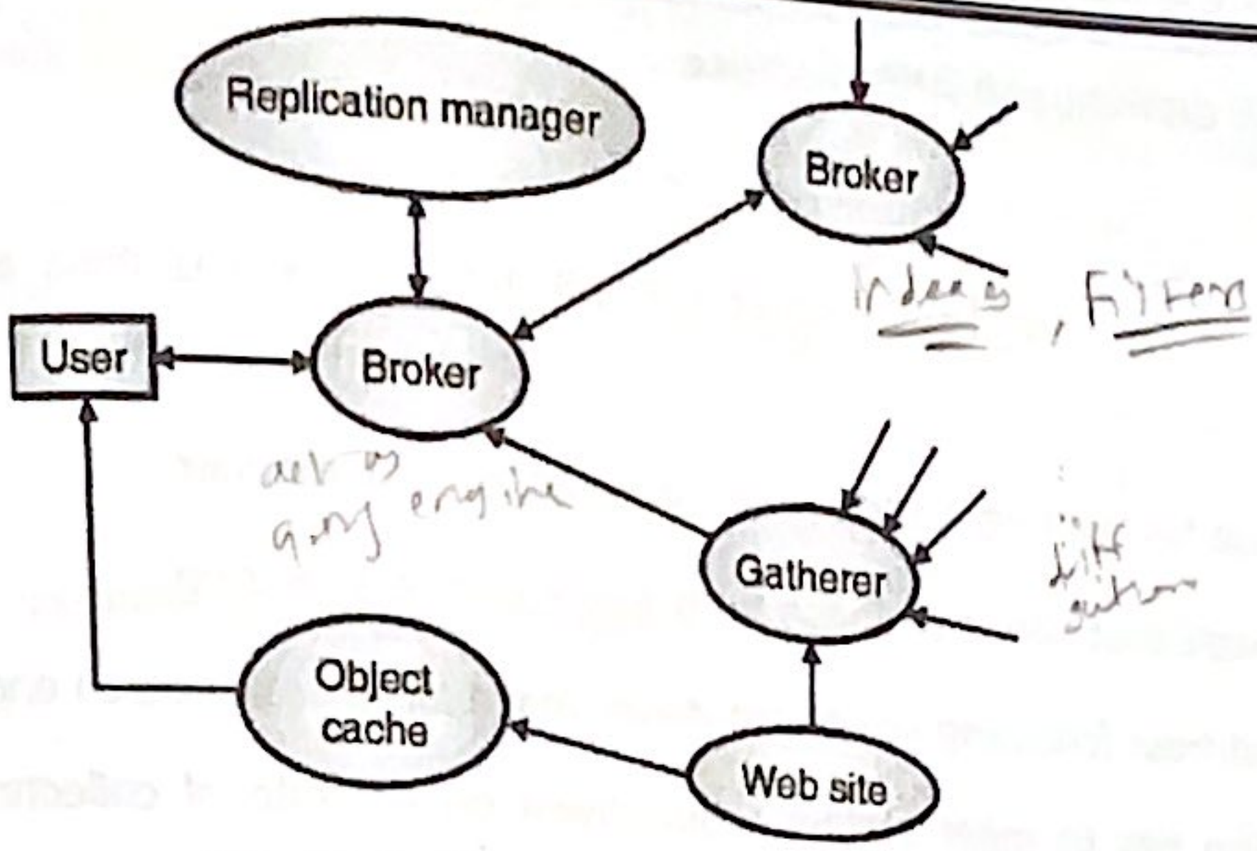

Fig. 5.5.2 : Harvest Architecture

- Harvest architecture solves following problems of the crawler-indexer architecture.
  1. Increase in web server load due to receiving requests from web crawlers.
  2. As crawler retrieve entire objects and discards most of the content but it increases web traffic.
  3. Without coordination of all search engines information is gathered by each crawler independently.
- To solve above mentioned problem Harvest introduces two main elements i.e. gatherers and brokers.
- Gatherer performs the task of collecting and extracting indexing information from one or more web servers. Times required for gathering are defined by the system and they are periodic.
- A broker performs the task of providing gathering mechanism and query interface to the gathered data.
- Broker retrieves information from one or more gatherer and also from other broker and updates their index incrementally.
- Different improvement on web server load and network traffic can be achieved by considering the configuration of gatherers and brokers.
- This achievement can be of :
  1. Without generating external traffic for web server gatherer can run on that web server.
  2. Gatherer can send information to several brokers by avoiding work repetition.
  3. Broker can filter information and send it to other brokers.
  4. Sharing of information and work in flexible and generic manner.
- Harvest architecture provides following goals :
  1. To build topic-specific broker by focusing on the index content and by avoiding many of the vocabulary and scaling problems of generic indices.
  2. Includes distinguished broker to allow other brokers to register information about gatherers and brokers. And this is helpful to search for appropriate broker or gatherer when you are building a new system.
  3. Provides replicators and object caches.
- Replicators are used to replicate servers to improve user-base scalability. It is also used to divide gathering process between many web servers.
- Object cache reduces network and server load. It also response latency when accessing web pages.
- There are hundreds of Harvest applications available on web like CIA and NASA.

Drawbacks of Harvest distributed Architecture

- Harvest requires coordination of several web servers.
- Harvest is lacking of flexibility in the sense that all gatherers and brokers are described in an administration registry.
- Harvest does not use his gatherers and brokers in a cooperative manner.
- Conceptual framework that allows all these tools to cooperate is missing.
- Harvest does not address following important issues that a large-scale search engine must face.
  1. A search engine has to meet certain requirement on the rate of collecting Web pages. Harvest does not consider this aspect.
  2. The Gatherer of Harvest will have good effect if it runs on the machines of Providers. However, it is impossible to make every Provider to do so.
  3. A Gatherer will discard URLs that could not be visited by itself. However, other Gatherers may be able to visit these URLs. So Harvest system does not resolve how to use cross URLs.
  4. Harvest has less effective controls on its Gatherers when it is used to collect information within a particular scope.

Comparison of centralized and distributed Architecture of search engine

University Question

Q. Compare centralized and distributed architecture of search engine. **SPPU : Dec. 13, 8 Marks**

Table 5.5.2

Sr. No.	Centralized Architecture	Distributed Architecture
1.	Crawlers are the software agents that traverse through the web and sends new or updated pages to main server where from where they are indexed.	This architecture is based on distributed architecture for gathering and distributing data.
2.	Increase in web server load due to receiving requests from web crawlers.	Gatherer performs the task of collecting and extracting indexing information from one or more web servers. So there will not be high load on server.
3.	As crawler retrieve entire objects and discards most of the content but it increases web traffic.	Without generating external traffic for web server gatherer can run on that web server.
4.	Without coordination of all search engines information is gathered by each crawler independently.	Gatherer can send information to several brokers by avoiding work repetition. Sharing of information and work in flexible and generic manner.
5.	Does not provides replicators and object caches.	Provides replicators and object caches.

Sr. No.	Centralized Architecture	Distributed Architecture
6.	Centralized architecture is less efficient than the distributed architecture.	Distributed architecture is more efficient than the centralized architecture.

5.5.4 User Interfaces

- There are two aspects of the user interfaces of search engines-
  1. Query interface
  2. Answer interface
- The basic query interface is a box where user can type one or more words.
- Even though the user expects given sequence of words in the same query in all search engine he or she will not get it.
- For example in AltaVista a sequence of words is reference to the union of all the web pages having at least one of those words.
- While in HotBot it is reference to the intersection of all web pages having all the words.
- While in case of answer interface relevant pages appear on the top of the list. Each entry in the list includes title of the page, URL, brief summary, size, date and written language.
- All search engines also provides a query interface for complex queries as well as command language including Boolean logic, phrase search, title search, URL search, date range search, data type search and other features.
- Fig. 5.5.3 shows query interface for HotBot engine.
- These search engines provides several filtering functions. The result can be filtered by additional words that must be present or absent from the answer or in a particular fields such as the URL or title, language, geographic region or internet domain, data range or inclusion of specific data types such as images or audio.

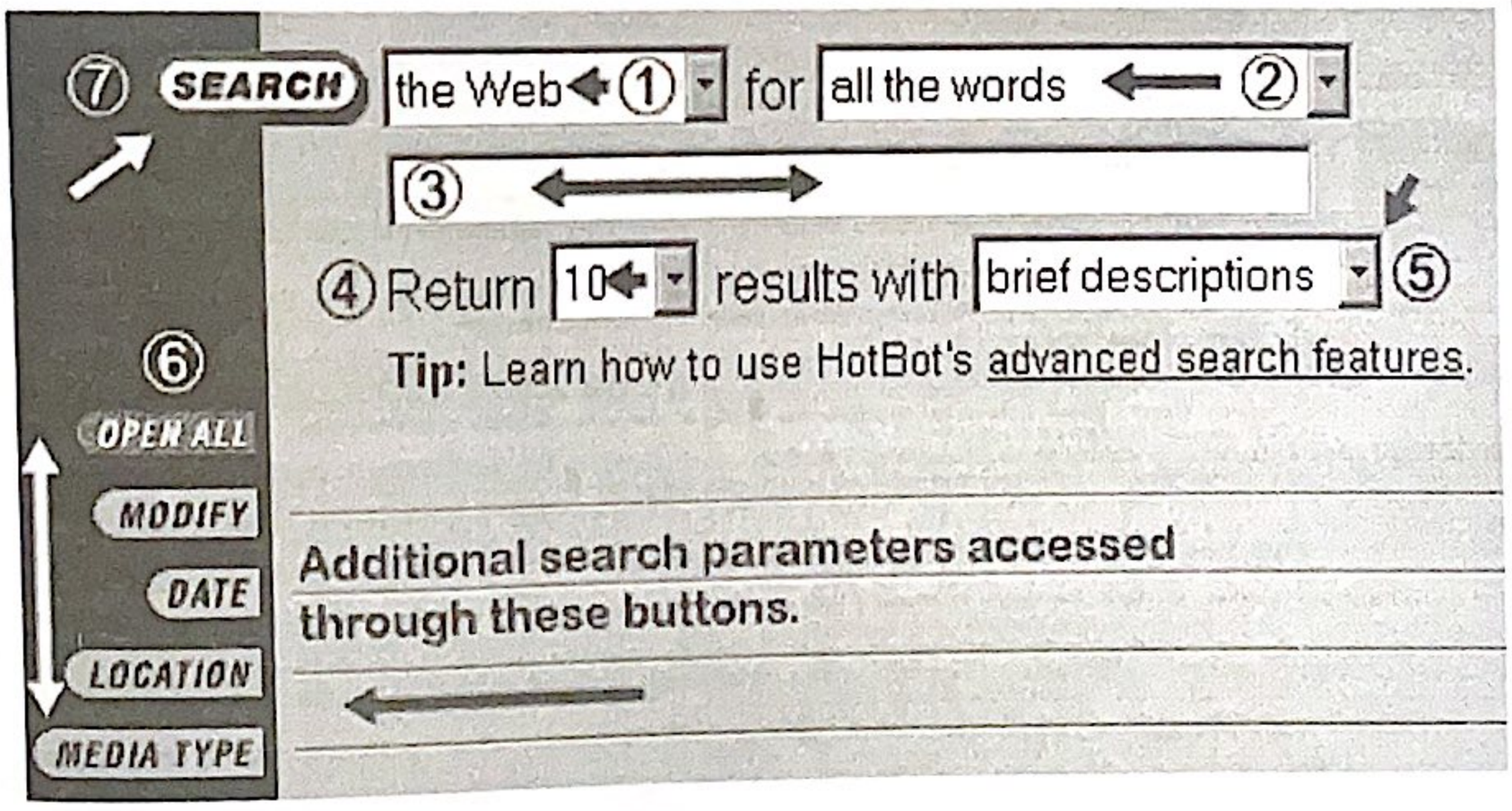

Fig. 5.5.3 : Query interface using HotBot engine

- Fig. 5.5.4 shows the query interface for Northern Light engine.

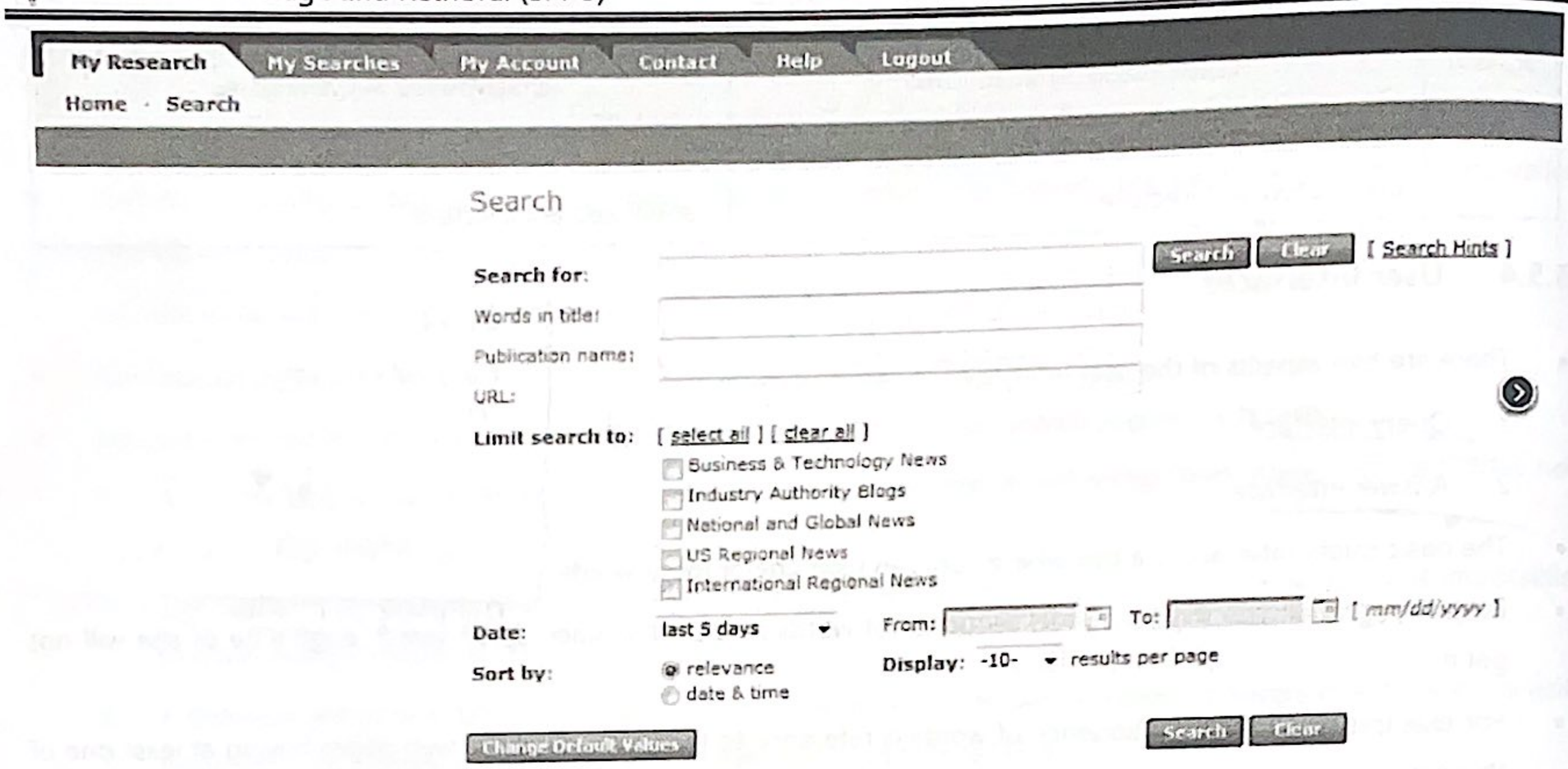

Fig. 5.5.4 : Query interface for Northern Light engine

- The search result consists of a list of the top ranked web pages.
- Fig. 5.5.5 shows the search result using Northern Light engine.

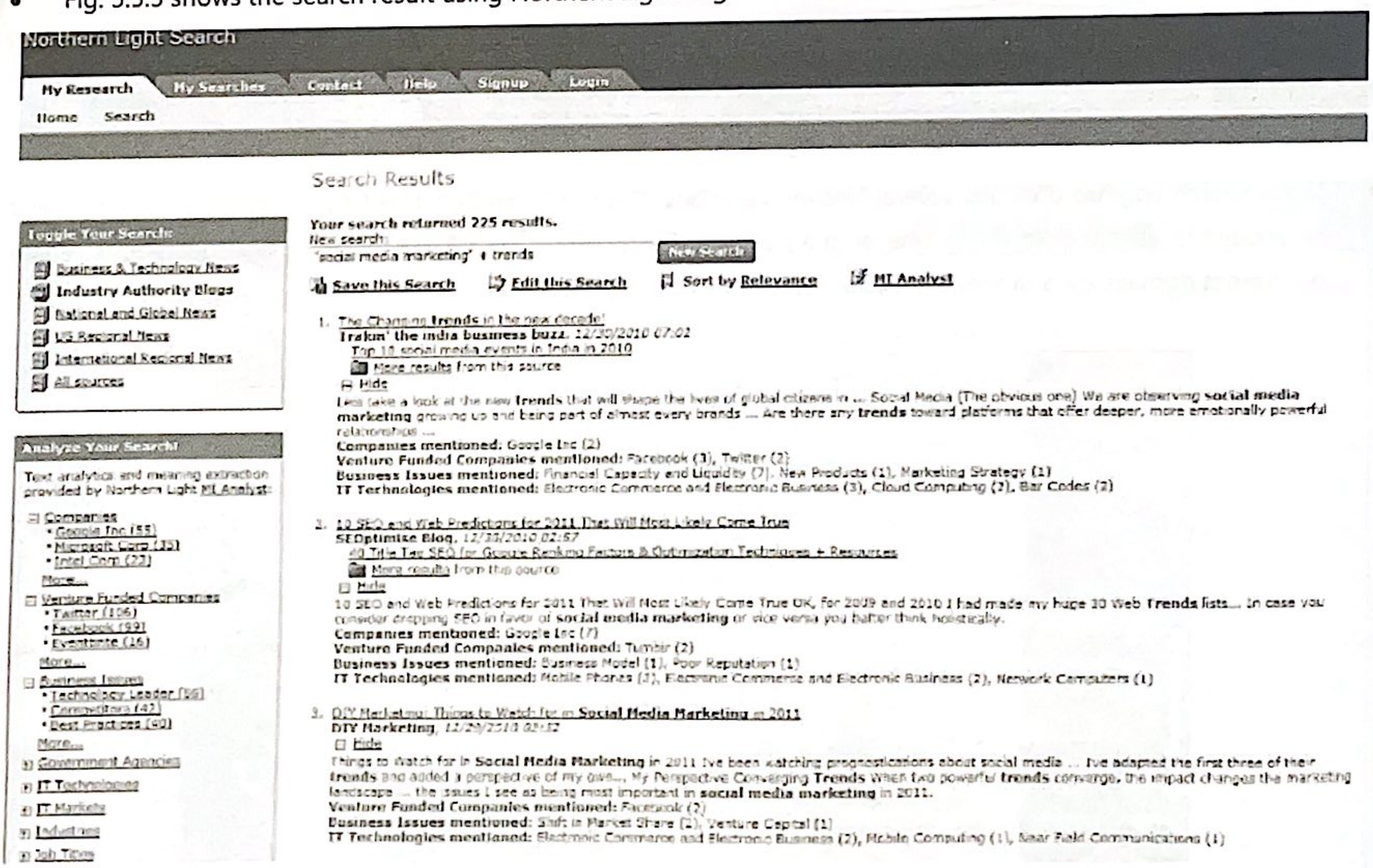

Fig. 5.5.5 : Search result using Northern Light engine

- Fig. 5.5.6 shows the search result using HotBot search engine.

- Each entry in the list includes some information about the document it represents. The information includes URL, size and the date when pages are indexed. Some search engine allows the user to change the number of pages returned in the list and the amount of information per page.
- The order of the list is by relevance. In addition most of the search engines have option for finding document similar to web page in the answer. The web pages retrieved by the search engine are ranked using statistics related to the terms used in the query and also sometimes terms included in metatags or the title.

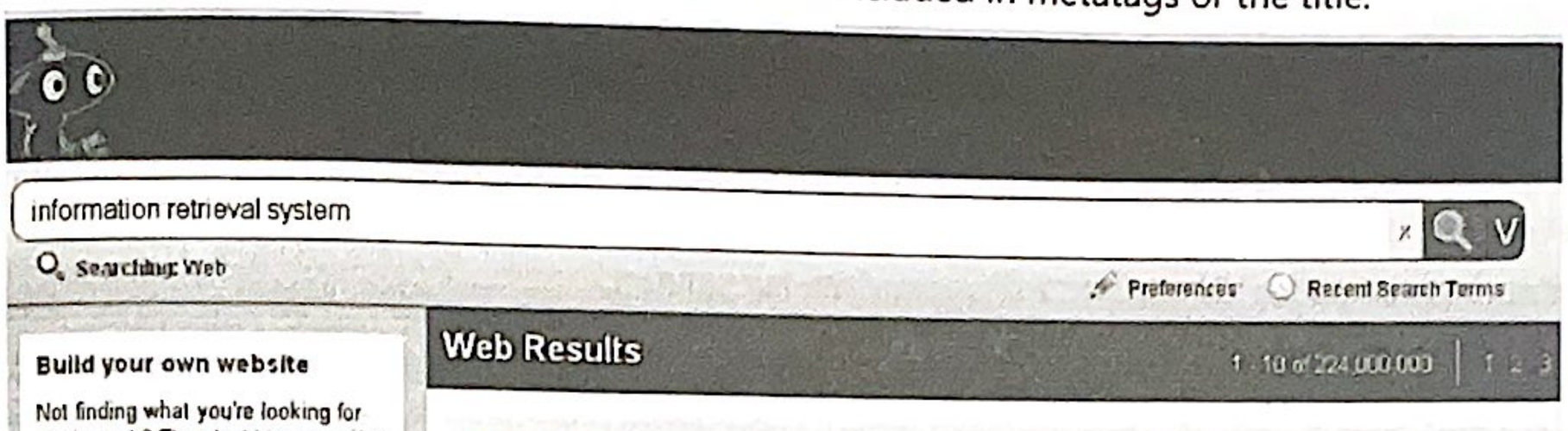

Fig. 5.5.6 : Search result using HotBot engine

5.5.5 Ranking

- Ranking refers to ranking document by relevance and in web searching Ranking has to be performed without accessing the text only by using the index.
- Most of the search engines use variations of the Boolean and vector model for the ranking.
- Yuwono and Lee proposed three ranking algorithms Boolean spread, vector spread and most-cited.
- First two algorithm i.e. Boolean and vector spread are extended from Boolean and vector model and they includes pages pointed to by a page in the answer.
- While most-cited is based on the terms included in pages having a link to the pages in the answer.
- Some of the algorithm uses hyperlink information. The first is WebQuery which allows visual browsing of web pages.
- It takes a set of web pages i.e. answer to a query and ranks them depending on the connection of web pages with each other. It extends the set by finding web pages that are highly connected to the original set.
- The second is developed by Kleinberg and used in HITS i.e. Hypertext Induced Topic Search.
- It depends on the query and considers the set of pages S that point to pages in the answer. And these pages that point to in S are called as authorities.

- The pages that have many outgoing link is called as hub. Better authority pages come from incoming edges from good hubs and better hub pages come from outgoing edges to good authorities.
- Let H (p) and A (p) be the hub and authority value of pages, then the value of H(p) and A(p) is given by the following equations -

$$H(p) = \sum_{u \in S \mid p \rightarrow u} A(u)$$

$$A(p) = \sum_{v \in S \mid v \rightarrow u} H(v)$$

- These values are calculated by using iterative algorithm. To avoid an explosion of the size of link matrix S a maximum number of pages pointing to the answer can be defined.
- One problem with this technique is it does not work with non-existent, repeated or automatically generated links. So solution to this problem is weight each link based on the surrounding content.
- Second problem is that the topic of the result can be diffused. Solution to this problem is analyzing the content of each page and assigns a score to it like traditional information retrieval ranking.
- One more example for ranking is Page Rank which is used by WebQuery and Google.
- It simulates a user navigating randomly in the web that jumps to a random page with probability q or follows a random hyperlink with probability 1-q.
- Here it is assumed that user will not go to previously visited pages following already traversed hyperlinks.
- Suppose that S (a) is the number of outgoing links of pages a and page a is pointed to by pages p1 to p2. Then page rank, PR (a) is given by the following equation :

$$PR(a) = q + (1 - q) \sum_{i=1}^{n}$$

Where q is set by the system.

- Page Rank can be calculated by using iterative algorithm. Google simulates users using the search engine to rank documents and Google uses citation graph i.e. 518 million links.
- So in order to help ranking algorithm page designers should provide information like titles, headings, Meta field, good links and so on.

5.5.6 Crawling the Web

University Question

Q. What is web crawling ? Explain techniques used by web crawlers to crawl the web.

SPPU : May 17, May 19, 8 Marks

- There are several techniques to crawl the web. Out that one is; start with the set of URLs and from that set of URL extract other URLs which are followed recursively in a breadth-first or depth-first fashion.
- Variation for above technique is start with set of popular URL i.e. URL containing frequently accessed information. But above two techniques works with one crawler.

- So it is difficult to coordinate more than one crawler to avoid visiting the same page more than once.
- There is one more technique which partitions the web according to country codes or internet names and then assigns one or more robots to each partition and then travel around each partition thoroughly.
- In the web crawling the order by which URLs are traversed are very important. So these can be traversed either using breadth-first or depth-first manner.
- In case of breadth-first technique looks at all the pages linked by the current page. In this technique web page coverage is wide but shallow and many requests will be sent to the web server.
- While in case of depth-first technique follow the first link of a page then do the same on that page until go deeper. It provides narrow and deep traversal.

5.5.7 Indices

- Search engine indexing collects, parses, and stores data to ease fast and accurate information retrieval.
- An inverted file is a list of sorted words or vocabulary and each list will have set of pointers pointing to the pages where it occurs. Most of the indices use variations of inverted file.
- Some of the search engines use elimination of stop words to reduce size of the index.
- The index is complemented with a short description of each web page i.e. date of web page creation, title, heading and this is used to give some idea about each document retrieval to the user.
- Initially user receives only the subset of the complete answer to each query but the search engine stores this whole answer set in memory to avoid recomputing if user asks for more documents.
- A user query is answered by doing binary search on the sorted list of words of the inverted file. If user is searching for multiple words then he has to combine the result to generate complete answer.
- Inverted files can also point to the actual occurrences of a word within a document. But this is very costly in space because each pointer has to specify a page and a position inside the page.
- But having the positions of words in a page we can answer phrase search by finding words that are near to each other in a page.
- Finding words which starts with a given prefix requires two binary searches in the sorted list of words.
- More complex queries like words with errors, arbitrary wild cards or regular expressions on words can be performed by doing a sequential scan over the vocabulary.
- This is slow but the best sequential algorithm for this type of query can search around 20 Mb of text stored in RAM in one second. Thus for several gigabytes we can answer those queries in a few seconds.
- There are some ideas but that cannot used for the web because sequential search is not affordable and it implies a network access.
- However in a distributed architecture as index is distributed logical blocks can be used.

5.6 Browsing

Browsing is nothing but to look for information on the internet. In this section we will discuss the tool i.e. web directories used for searching and browsing on the web.

5.6.1 Web Directories

- Web directories are also called as catalogs, yellow pages or subject directories.
- Table 5.6.1 shows the most important web directories including URL, number of web pages indexed and categories in the beginning of 1998.

Table 5.6.1

Web directory	URL	Number of Web sites	Categories
eBLAST	www.eblast.com	125	–
LookSmart	www.looksmart.com	300	24
Lycos Subjects	www.lycos.com	50	–
Magellan	magellan.excite.com	60	–
NewHoo	www.newhoo.com	100	23
Netscape search.	netscape.com	–	–
Search.com	www.search.com	–	–
Snap	www.snap.com	–	–
Yahoo!	www.yahoo.com	750	–

- Out of the entire web directory yahoo is the best and oldest example of web directory. And Yahoo!, eBLAST, LookSmart, Magellan, and Nacho are the most popular.
- Most of the above web directories allowed keyword searches. AltaVista Categories, AOL Netfind, Excite Channels, HotBot, Infoseek, Lycos Subjects, and WebCrawler Select allow category services.
- Directories are hierarchical taxonomies that classify human knowledge.
- Table 5.6.2 shows the first level of the taxonomies used by web directories. First level of taxonomies range from 12 to 26.

Table 5.6.2

Arts and Humanities	Local
Automotive	News
Business and Economy	Oddities
Computers and Internet	People
Education	Philosophy and Religion
Employment	Politics
Entertainment and Leisure	Recreation
Games	Reference

Government	Regional
Health and Fitness	Science and Technology
Hobbies and Interests	Shopping and Services
Home	Social Science
Investing	Society and Culture
Kids and Family	Sports
Life style	Travel and Tourism
Living	World

- In most of the web directories pages have to be submitted to the web directory where they are reviewed and if page is accepted it is classified in one or more categories of the hierarchy.
- Even though the taxonomy is looking like tree there are cross references so it is a directed acyclic graph.
- The advantage of the above technique is that we can find the useful answer in most of the cases.
- But the disadvantage is that the classification is not specialized and all web pages are not classified.
- Web directories also allow the user to perform classification based on taxonomy descriptors or in the web pages pointed by taxonomy.
- But as the numbers of classified web pages are small we can afford the cost of having a copy of all the pages but these pages must be updated frequently and it poses the problem of performance and temporal validity.
- In addition to that most of the web directories also send the query to a search engine and allow the whole web to be searched.

5.6.2 Combining Searching with Browsing and Some Helpful Tools

- The two paradigms of searching and browsing are currently almost always used separately.
- One can either look at the library card catalog, or browse the shelves; one can either search large WWW sites (or the whole web), or browse page by page.
- In web directories a search can be reduced to a sub tree of the taxonomy but sometimes some search miss the related pages which are not part of taxonomy.
- There are some search engines which finds similar pages using common words but this method is not effective.
- There are some tools which are helpful to combine browsing with searching. Out of that one is WebGlimpse.
- This tool allows the search to be limited to a *neighbourhood* of the current document. WebGlimpse automatically analyzes collections of web pages and computes only those neighbourhoods.
- With this WebGlimpse tools users can browse using the same pages; they can also jump from each page, through a search, to "close-by" pages related to their needs. This combined paradigm allows users to browse using hypertext links that are constructed on the fly through a neighbourhood search.
- The design of WebGlimpse concentrated on four goals: fast search, efficient indexing (both in terms of time and space), flexible facilities for defining neighbourhoods, and non-wasteful use of Internet resources.

- Alexa is a simple handy toolbar that will provide you all the "Inside Information" about the site, like contact details, traffic ranking, customer reviews, related sites and other sites that direct a link to the particular site you are surfing.
- It also provides you a quick and easy search without disturbing your browsing.
- We generally visit the search engine, for example www.google.com to fetch the desired sites typing suitable keywords. Alexa toolbar has this facility inbuilt, which means that there is no need to visit a search engine site.
- Just key in the suitable keyword in the Alexa toolbar and press search, you will get the matching results from the Google search engine as Alexa search engine is powered by Google.
- Alexa computes traffic ranking by recording the web usage of millions of Alexa toolbar users. Ranking is calculated based on the 3 months data of Alexa users & their number of visits to different web sites.
- There are also some commercial tools to visualize Web subsets i.e. Microsoft's Site Analyst, MAPA from dynamic Diagrams, IBM's Mapuccino, SurfSerf, Merzscope from Merzcom, CLEARweb, Astra Site Manager, Web Analyzer, from InContext, HistoryTree from SmartBrowser. Non commercial includes WebMap, Sitemap, Ptolomeaus.

5.7 Meta Searchers

University Questions

Q. What are Meta searchers ? Explain with suitable example. **SPPU : Dec. 12, Dec. 16, 6/8 Marks**

Q. Explain Meta searchers with examples. **SPPU : May 14, 8 Marks**

- Meta searchers also called as Meta search engines are powerful parallel search services or web servers that allow you to query multiple sites to several search engines, web directories and other databases simultaneously.
- It collects the answer from different sources and returns the uniform result to the user. It also to sort it by host, keyword, data, and popularity. Meta searchers can run on client machine as well the number of sources is adjustable.
- Fig. 5.7.1 shows the architecture of Meta searchers.
- The architecture of Meta searchers uses wrappers. Wrappers export a common data model view of each source's data. Wrappers also provide a common query interface.
- After receiving a query, a wrapper translates it into a source-specific query or command, hence giving interface transparency to the user.
- Then, the wrapper translates the query results from the underlying source into the common data model or format.

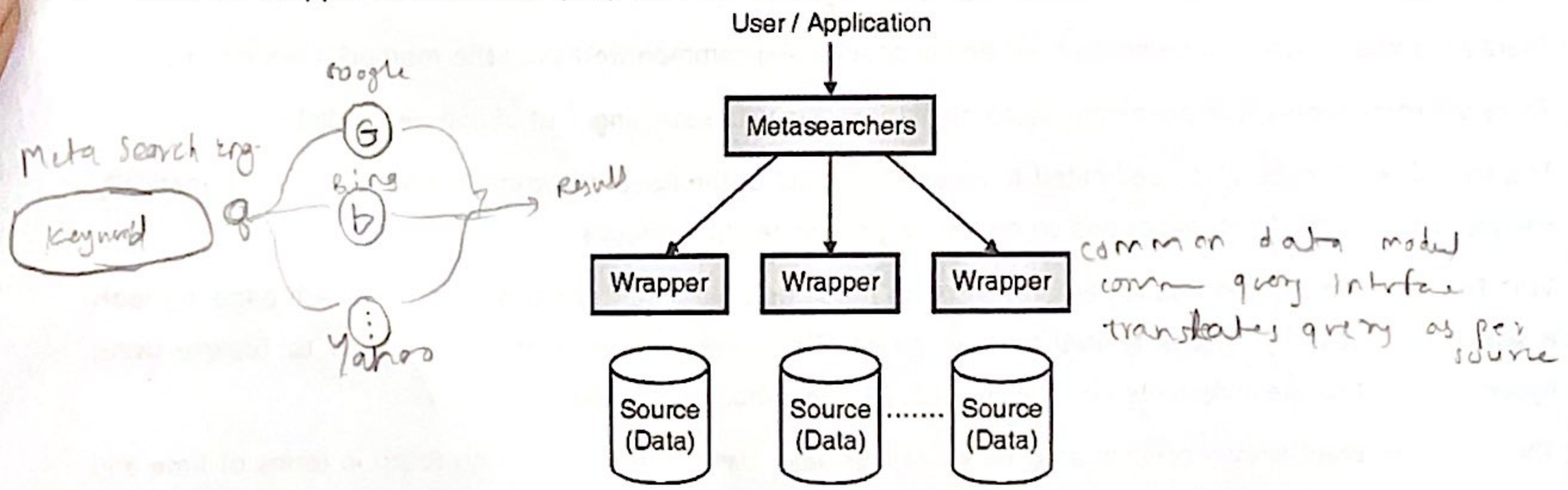

Fig. 5.7.1 : Architecture of Meta searchers

- Meta searchers provides the user with a virtual integrated view of the heterogeneous sources.
- User can access the view using a unified query interface. This query interface offers *location*, *model*, and *interface transparency*, i.e., users have the illusion of a single database and do not have to be aware of the location and interface of the sources.
- Although users and applications might access data directly through wrappers, Meta searchers offer an integrated view of the world, where information related to the same entity has been fused together, redundancies have been eliminated, and inconsistencies have been removed.
- Table 5.7.1 shows the examples of Meta searchers, their corresponding URL's, and number of sources that they use.

Table 5.7.1

Metasearcher	URL	Sources used
C4	www.c4.com	14
Dogpile	www.dogpile.com	25
Highway61	www.highway61.com	5
InFind	www.infind.com 6	6
Mamma	www.mamma.com	7
MetaCrawler	www.metacrawler.com	7
MetaMiner	www.miner.uol.com.br	13
Local Find	local.find.com	N/A

Advantages of Meta Searchers

1. It has ability to combine the results of many sources and the fact that the user can pose the same query to various sources through a single common interface.
2. The result can be sorted by different attributes such as host, keyword, date etc.
3. They can run on the client
4. Some of the Meta searchers search several sources and shows the different answers in separate windows.

Disadvantages of Meta Searchers

1. The number of results per search engine retrieved by the Meta searchers is limited.
2. All the result will not match with query.
3. Sometimes pages returned by more than one search engine are not more relevant.
4. Better ranking is required.

- NEC Research Institute Meta search engine, Inquirus will to better ranking but these are not available for general public. Inquirus downloads and analyze each page obtained and then displays each page, highlighting the places where the query terms were found.

- In this the results are displayed as soon as they are available in progressive manner otherwise the waiting time would be too long. This also allows non-existent pages or pages that have changed and do not contain the query any more to be discarded.

5.8 Web Crawlers

University Questions

Q. Explain the different components of web crawler. **SPPU : May 13, 8 Marks**

Q. Write short note on web crawlers. **SPPU : Dec. 13, 8 Marks**

Q. What is the role of crawler in web searching? Explain the strategies used by web crawler. **SPPU : Dec. 16, 6 Marks**

- A web crawler is a program or automated script, which browses the World Wide Web in a methodical, automated manner.
- A web crawler is a program that downloads the web pages associated with the URLs, extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks when given one or more URLs. The process is called Web crawling or spidering.
- Web crawler also defined as the software agents that traverse the Web sending new or updated pages to a main server where they are indexed.
- Also called as robots, web spiders, worms, wanders, walkers, ant, scrapers, knowbots. The first crawler, Wanderer was developed in 1993. Web crawlers run on local machine and send requests to remote Web servers. Fastest crawlers are able to traverse up to 10 million pages per day.
- Because of the huge amount of information on the web, it takes several weeks to crawl it totally.
- Even the largest search engines, like Google and Altavista, cover only limited parts of the web and much of their data are out of date several months of the year.
- Therefore, they usually do not cover some important information that changes hourly or daily like news. Most of the recent works done on crawling strategies attempt to minimize the number of pages that need to get downloaded, or maximize the benefit obtained per downloaded page.

5.8.1 How it works ?

- When a web crawler of search engine visits a web page, it reads the visible text, the hyperlinks, and the content of the various tags used in the site.
- These contents include keyword rich Meta tags. Then search engine will determine what the site is about and index the information using the information gathered by the crawler.
- The website is then included in the search engine's database and its page ranking process.
- Web crawlers may operate one time only i. e. for a particular one-time project. If it is used for long-term like search engines, then they may be programmed to explore through the Internet periodically to determine whether there has been any significant changes.
- If a site is experiencing heavy traffic or technical difficulties then the web crawler may be programmed to note that and revisit the site again i.e. after the technical issues have subsided.

5.8.2 Architecture of Web Crawler

- Any crawler must fulfil following two issues
  1. Good crawling strategy
  2. Should have a highly optimized system architecture that can download a large number of pages per seconds.
- Most of search engines use more than one crawler and manage them in a distributed method to get the benefits like increased resource utilization, effective distribution of crawling tasks with no bottle necks and configurability of the crawling tasks.
- Fig. 5.8.1 shows the high level architecture of standard web crawler and Fig. 5.8.2 shows the easy architecture of Web Crawler.

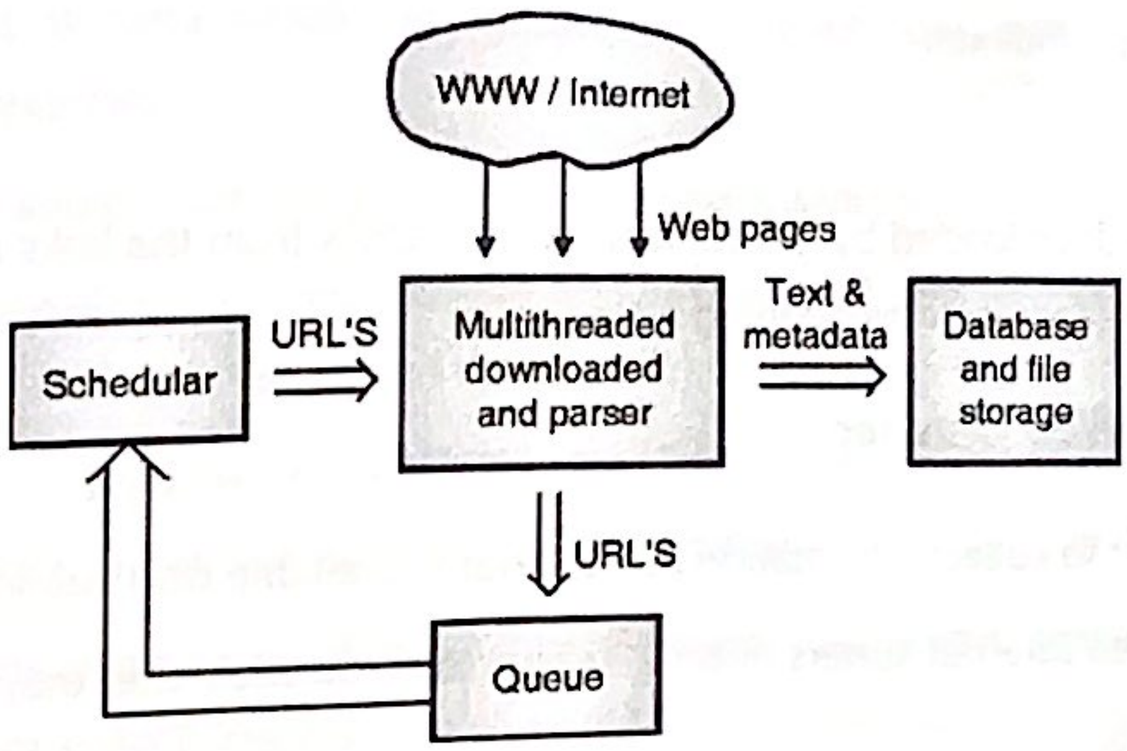

Fig. 5.8.1 : High-level architecture of standard web crawler

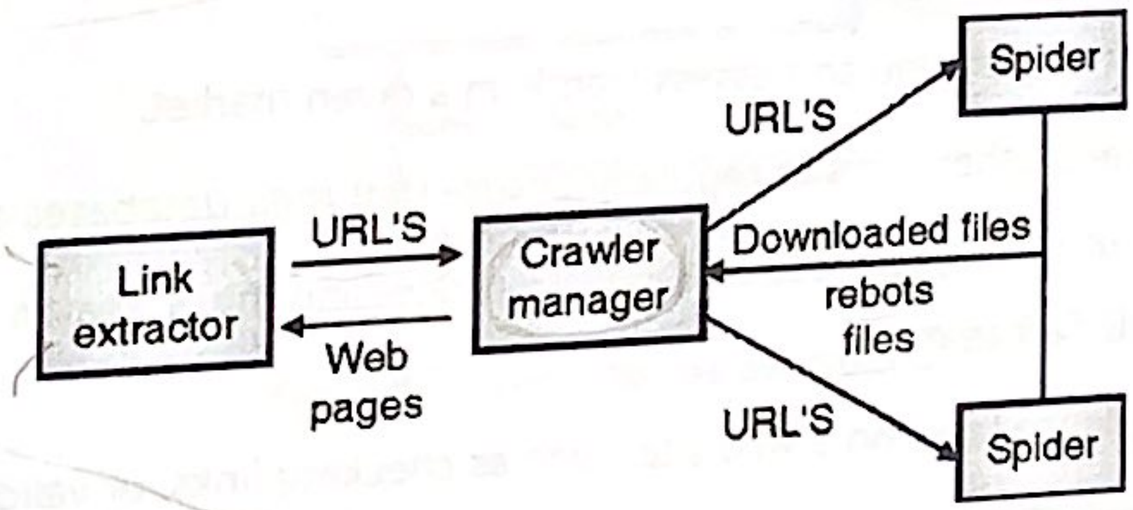

Fig. 5.8.2 : Easy architecture of web crawler

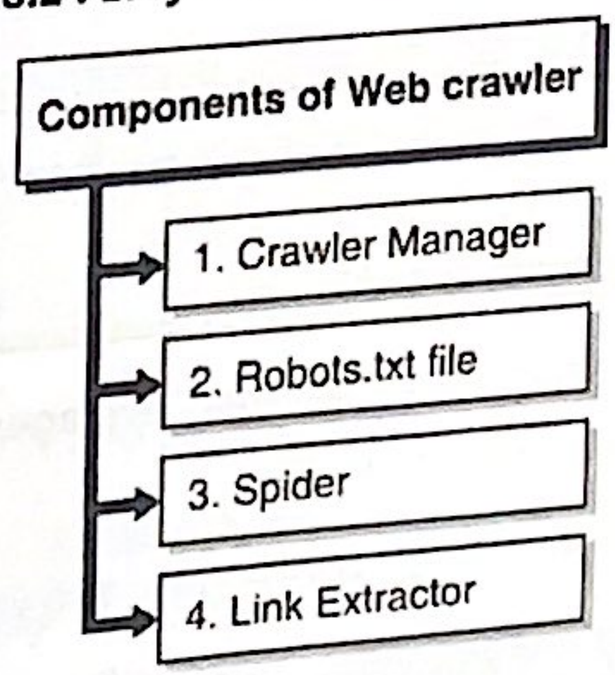

Fig. 5.8.3 : Components of web crawler