Notes

18

339en5B18

SEARCH TECHNIQUES: WEB BASED SEARCH

18.1 INTRODUCTION

Internet has become the biggest repository of information in the world. It can be considered as a global library where variety of information in different languages and formats is stored in digital form. The volume of information on web is enormous and it has become near to impossible to estimate its size. Because of its size and storing mechanism, finding relevant and precise information has become a difficult task. For searching information from this vast repository, we use search engines. There are thousands of search engines available on internet. For example, if you visit http://www.thesearchenginelist.com/, you will find a classified list of search engines. This list is category-wise and includes all-purpose search engines in various fields like accounting, blogs, books, legal, medical, etc.

In Lesson 17, you studied basic concepts of search techniques. Here, you will learn various aspects of searching information on web.

18.2 OBJECTIVES

After studying this lesson, you will be able to:

- explain purpose of simple and advanced search techniques;
- develop search string using Boolean logic on a given topic;
- illustrate search string with the help of a diagram;
- give examples of simple search and advanced search on internet;
- identify various Search Engines, viz. Google, Yahoo, Google Scholar;

Notes

- identify Search Engines on internet in different vernacular languages;
- illustrate search in specific categories, viz. maps, images; and
- modify search strings to get precise results.

18.3 PURPOSE OF THE SEARCH

People search information on web for various purposes. The best way to begin a search for information is to define information needs of the user. Information need is an individual or group's desire to locate and obtain information to satisfy a conscious or unconscious need.User may need an overview, a comprehensive search, a quick reference or fact, or an in-depth treatment of a topic. Once it is decided what is needed, a source can be selected which is likely to have the desired information. A search strategy is then planned that includes various sources. There is a vast range of sources available to help locate desired information.

18.3.1 Search Engines

A search engine is a software program that searches for websites based on terms referred to as search terms. Internet search engines are thus special sites on the Web that are designed to help locate information stored on other sites. There are differences in the way various search engines work, but they all perform three basic tasks:

- Search the Internet, or select parts of the Internet based on important words,
- Keep an index of the words they find, and where they find them, and
- Allow users to look for words or combination of words found in that index.

18.3.2 Kinds of Web Search

Information need is defined as 'an individual or group's desire to locate and obtain information to satisfy a conscious or unconscious need.' It is this perceived need for information that leads users to use an information retrieval system in the first place. The perceived need for web search can be of three types:

(i) Navigational Search

(ii) Informational Search

(iii) Transactional Search

(i) Navigational Search

A navigational search is a keyword search in which the searcher wishes to go to a specific website, or a web page on a specific site. In other words, here the searcher uses a web search engine to navigate (go to) a website. For example, if you wish to go to the website of the'President of India'.To do so, just type the query 'President of India'in a search engine (say Google) and search the web. The list provided by Google contains a link to the President of India website along with other links.Just by clicking the link, you will reach the website. The result of this search is given in Fig. 18.1.

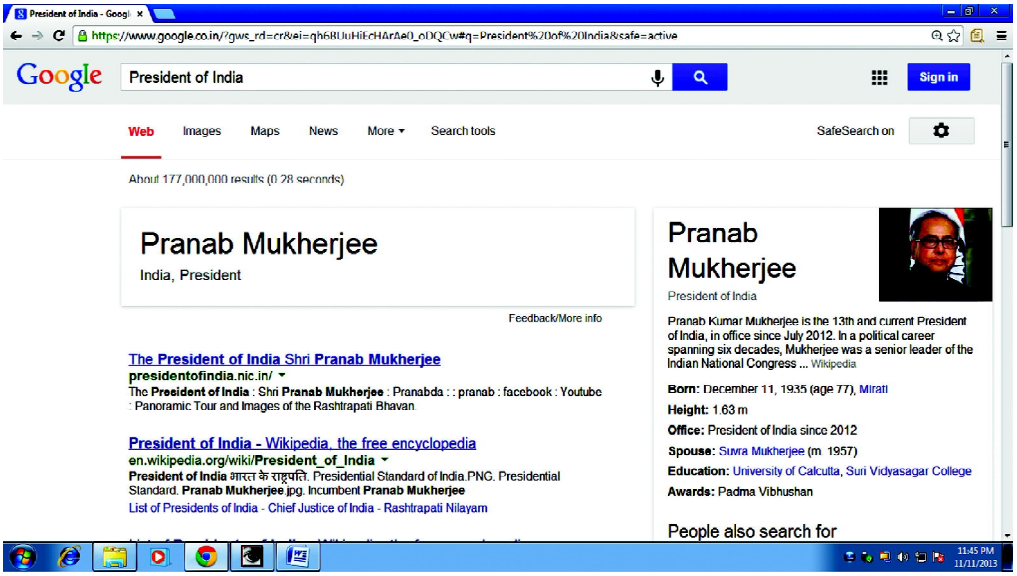


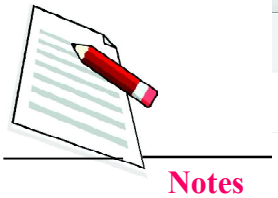
Fig. 18.1: The search result of a Google search on the topic 'President of India'

(ii) Informational Search

The intent of the informational search is to acquire some information, assuming it is available on the internet. This kind of search is conducted for study, research or any other purpose where scholarly information is required. For example, a person wants to find information on a topic 'Career in library and information science'. When the query is put to Google search engine on the web, it provides a list as search results, which contains references of 42,400 hits from across the web. The result is shown in Fig. 18.2.

After analysing the results, we find that the list contains references of different websites. The websites are of academic, research, commercial and many other institutions as well as organisations.

Search of specialized information systems, such as, LCOC, PubMed, OPAC of a particular library or likewise also fall in this category.



Notes

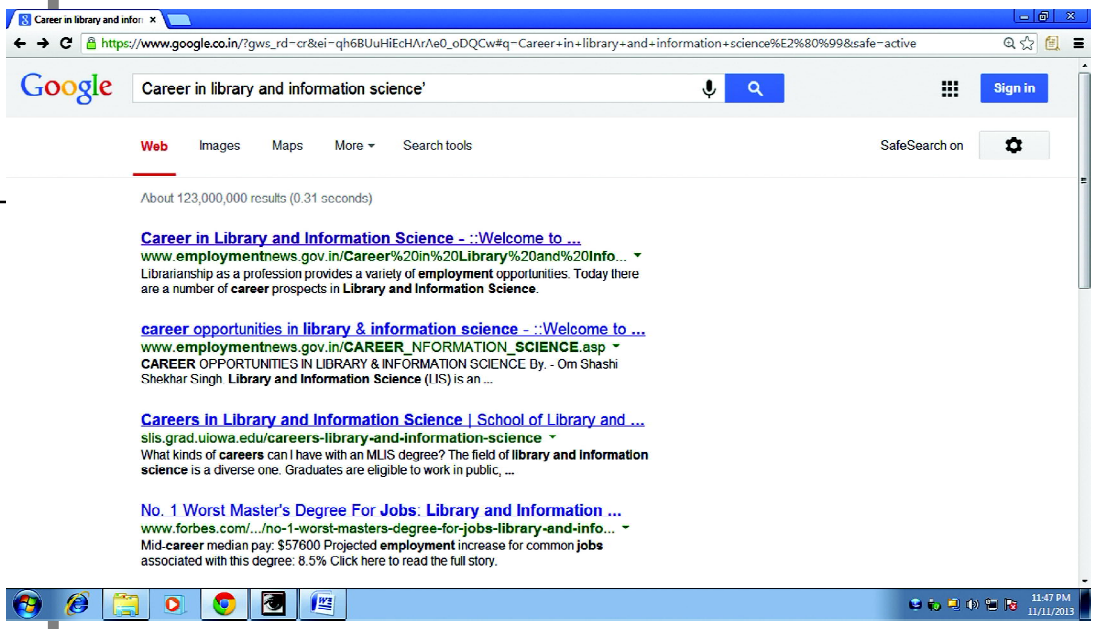


Fig. 18.2: The search result of a Google search on the topic'Career in library and information science'

(iii) Transactional Search

Transactional search intends to reach a website for further interaction or some other activities. The purpose of such queries could be shopping, downloading various types of files, as image, song, movies, etc. and various web mediated services like gaming, etc. For example, a person can go to websites where one can buy online tickets for airplane, train, bus, movie, etc.

Searching information on web for navigational and transactional search is general in nature. But searching for informational purpose needs certain level of knowledge and skills.

INTEXT QUESTIONS 18.1

1. Define a search engine.
2. Describe the purpose of informational search on internet?

18.4 SEARCHING INFORMATION ON INTERNET

Searching information on internet is both an art and a science. One can get expertise in finding relevant information on internet. Information in the form of simple websites, databases, books, maps, journal articles, audio-visual materials, multimedia, or any desired topic can be found. As information in several formats is available on internet, finding through a particular type of

search technique or searching with the help of one search engine is not possible. In the Lesson 17, you have studied the search process, different aspects of search techniques, designing search string or query, etc. In this lesson, we will apply these aspects for searching information on internet. We can search information with the help of **simple** or **advanced** search techniques as explained below:

18.4.1 Simple Search Technique

Simple search technique is a mechanism of finding information on internet for beginners. Here, we select keywords and make a simple string or query. This query when submitted to any search engine provides a list of references of all those resources which have these keywords in their content. These keywords may appear in any part of the document, that is, title, body text or anywhere else. As the volume of information on internet is very large, huge number of references are provided by the search engines in simple search mechanism. For example, a search was conducted on www.google.com with a string 'Career selection after senior secondary' and result showed about 4,460,000 items. Fig. 18.3 shows the result of this search.

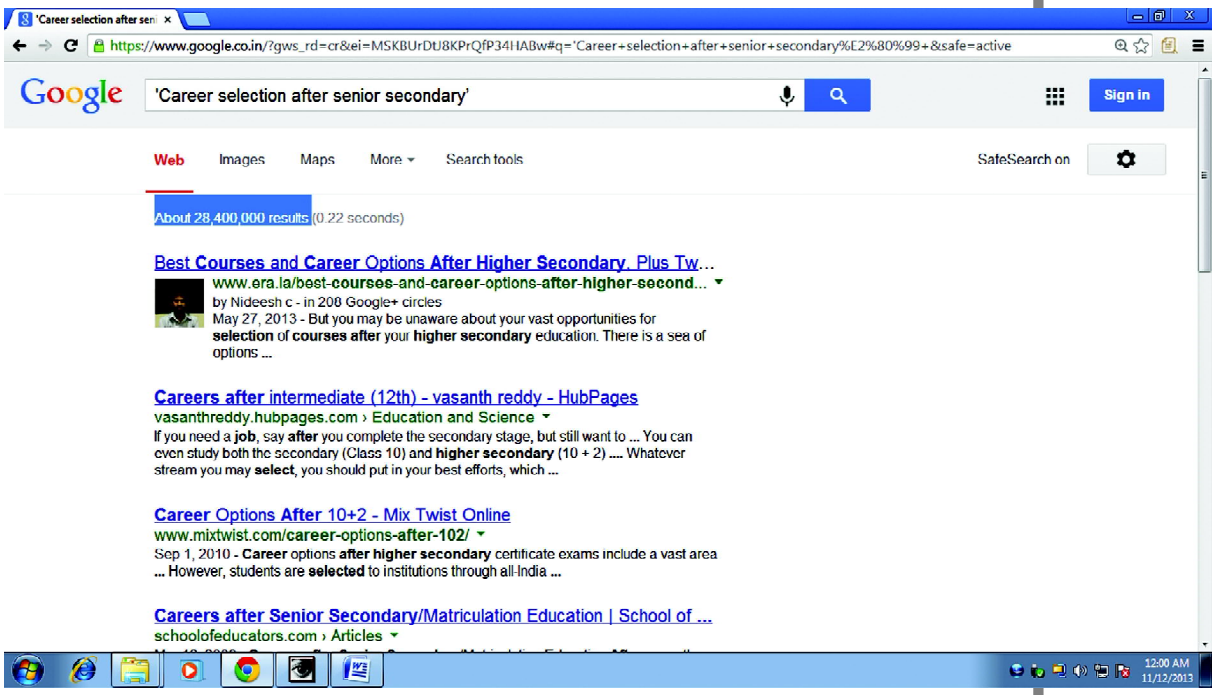


Fig. 18.3: Search Result of search 'Career selection after senior secondary' on Google

Further, the same concept was redesigned as 'Career selection after 12th' and searched through the same search engine, Google. The search showed about 8,580,000 results. Fig. 18.4 shows the result of this query.

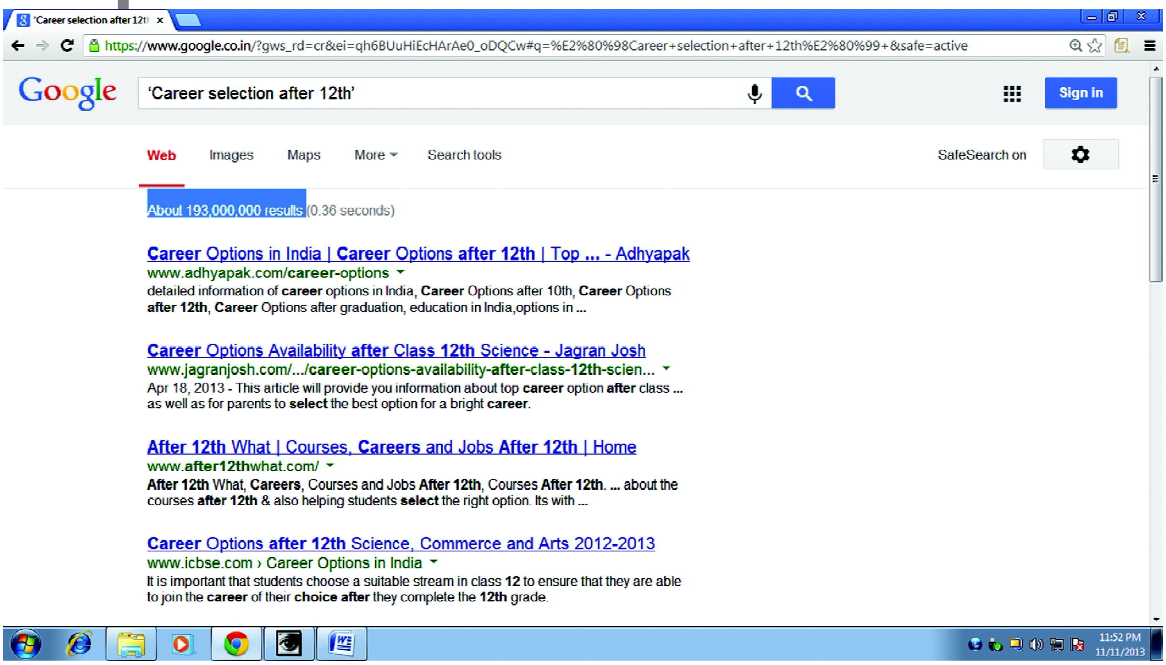


Fig. 18.4: Search Result of search 'Career selection after 12th' on Google

After analysing the results of these two searches, you can understand that even if the concepts are same, a change in keywords yields different search results. Hence, the advanced search techniques are required to be used to get precise search results.

18.4.2 Advanced Search Techniques

Advanced search techniques are a searching mechanism, which uses different parameters for getting precise search results. In previous lesson, you have studied different operators and parameters used in designing a string to get precise results. Different search engines provide different set of parameters and operators to get relevant and precise results. It is recommended that a searcher should follow the guidelines provided by a particular search engine and accordingly design a query. Further, for reducing the number of references from the set of search results, add more parameters. For example, we selected the Google Advance at http://www.google.ca/advanced_search, for our search query 'Career selection after senior secondary'. The Google Advance provides parameters for limiting the search for getting more relevant results. The search conducted on the basis of these parameters showed 379, 000 references. The query has been shown in Fig. 18.5 and the result in Fig. 18.6.

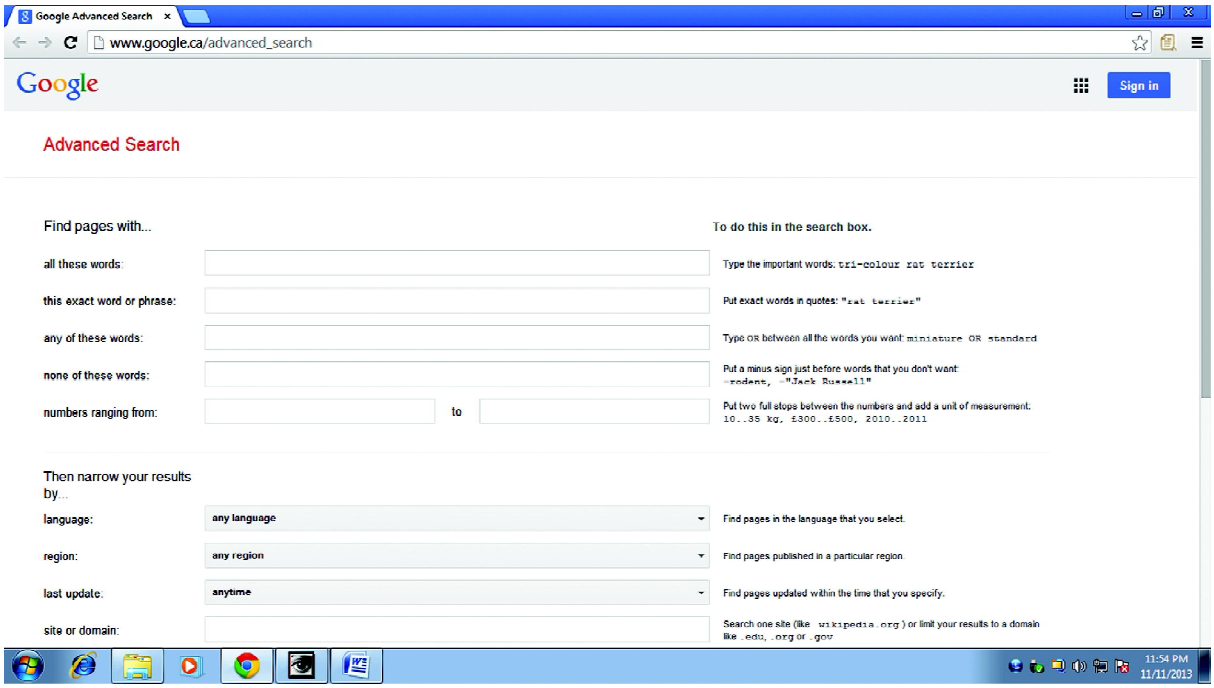


Fig. 18.5: Query 'Career selection after senior secondary'submitted to Google Scholar

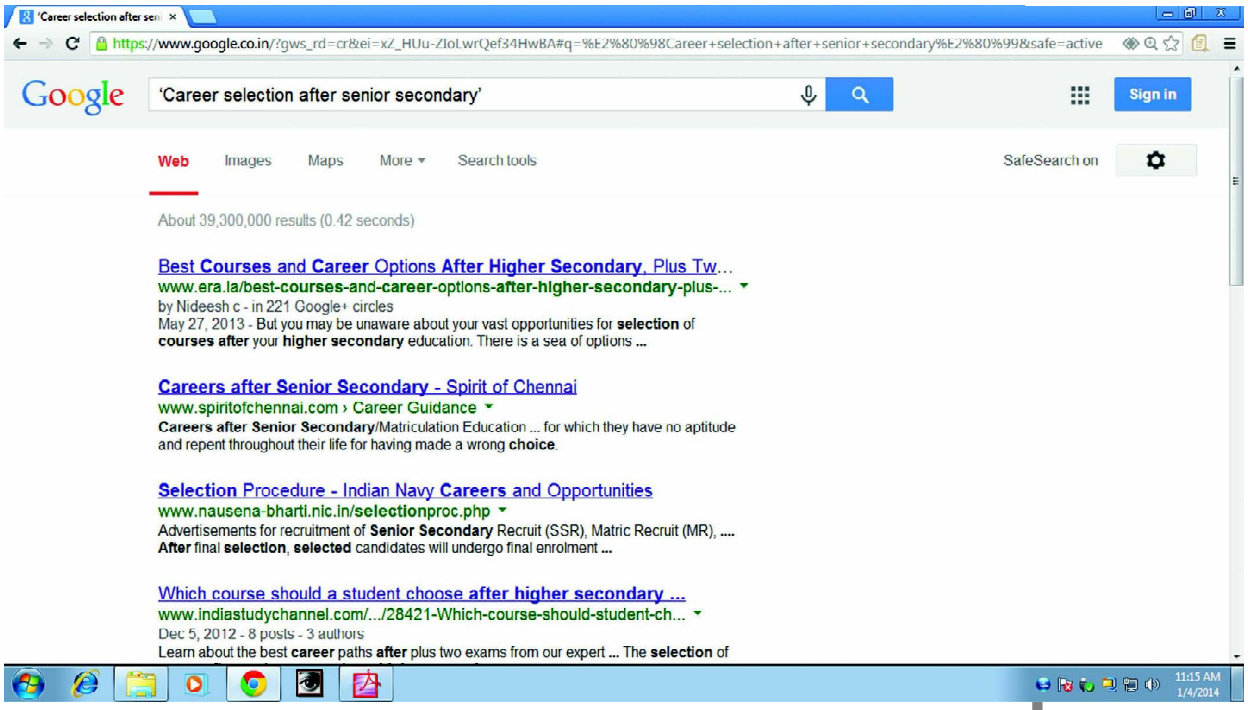


Fig. 18.6: Search result for query 'Career selection after senior secondary' by Google Scholar

Further, the same search was redesigned and more parameters were added. The input box containing keywords of string has been shown in Fig. 18.7.

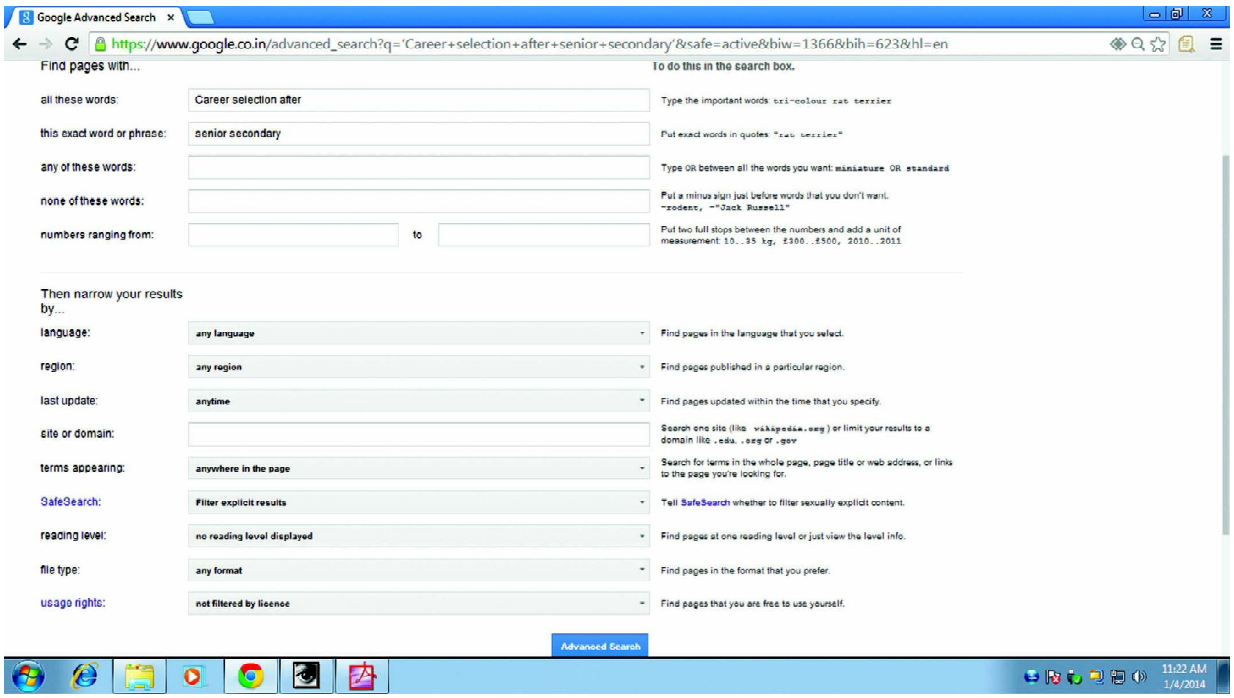


Fig. 18.7: Query 'Career selection after senior secondary' to Google Scholar using parameters

With these parameters, we got a list of 49,000 references. It clearly shows the impact of adding more parameters, provided by advance search engine.

INTEXT QUESTIONS 18.2

1. Define simple search technique.
2. Explain advanced search technique.

18.5 BOOLEAN LOGIC AND QUERY

Boolean Logic and its operators have already been discussed in previous lesson. Let us now design a few search strings and search these on internet to see the impact of the logic. The Google has given guidelines for using the Boolean Operators, "AND", "OR", "NOT"and other operators at http://support.google.com/websearch/bin/answer.py?hl=en&answer=136861. Before using Google, it is recommended to go through the guidelines to get better search results.

We designed the same search as 'Career selection after' and 'Senior secondary' OR '12th' with the language parameter 'English' and region 'India'. We got a list of only four references. The result is shown in Fig. 18.8.

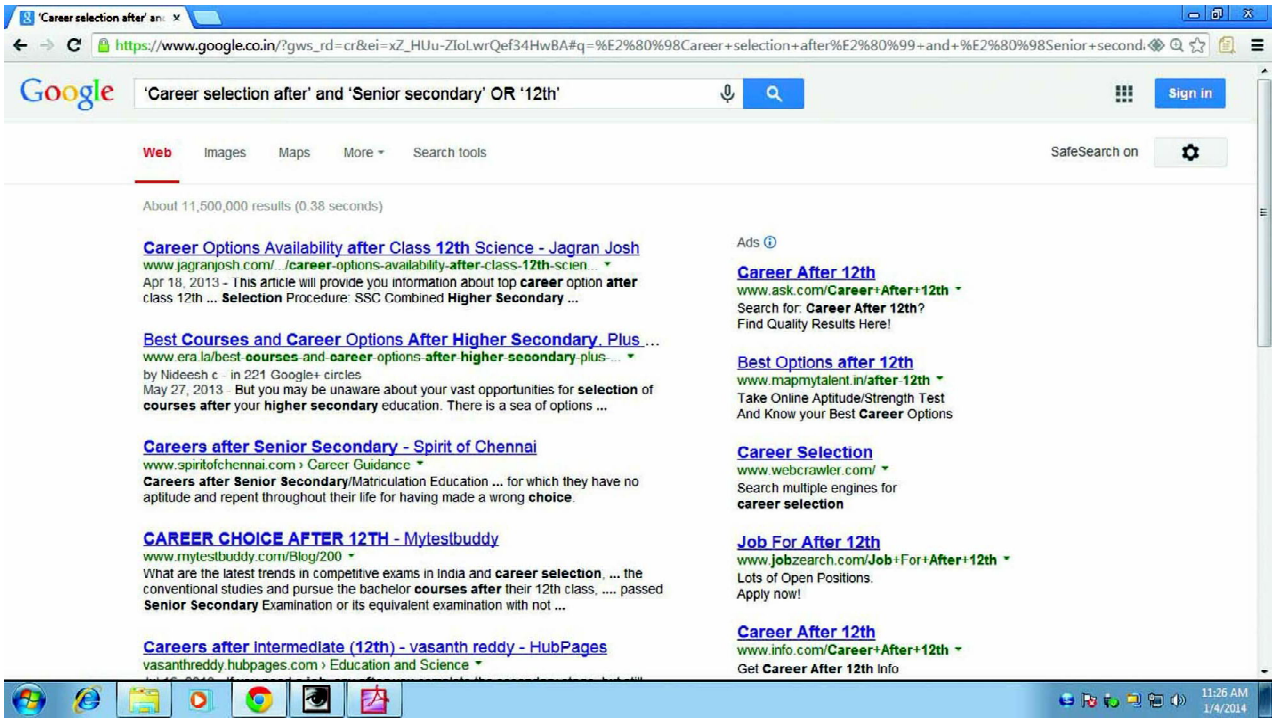


Fig. 18.8: Search result of query 'Career selection after' and 'Senior secondary' OR '12th' with the language parameter 'English' and region 'India'

INTEXT QUESTIONS 18.3

1. What is the purpose of "AND" in Boolean Logic?
2. Identify the purpose of "OR" in Boolean Logic.
3. Elaborate the purpose of "NOT" in Boolean Logic.

18.6 SEARCH ENGINES ON INTERNET

Hundreds of search engines are available on internet. There are a number of websites which provide studies and analyses about the search engines, those are active on internet. For further information, you can visit http://www.thesearchenginelist.com/. As it is not feasible to list all the search engines and their features here, a list of selected search engines with their brief introduction and categories is given below.

Search Engine	Description
Google	**Google**: The world's most popular search engine.
bing	**Bing** Search: Microsoft's entry into the burgeoning search engine market. Better late than never.
YAHOO!	**Yahoo!** Search: The 2nd largest search engine on the web (as defined by a September 2007 Nielsen Netratings report).
LexisNexis	**LexisNexis**: LexisNexis claims to be the "world's largest collection of public records, unpublished opinions, forms, legal, news, and business information". Searchable archive of newspapers, public records & more.
dieselpoint	**Dieselpoint**: Search & Navigation. Dieselpoint provides advanced full-text search with data navigation capability. It gives users highly relevant results not possible with either traditional search engines or SQL databases.
ORACLE SECURE ENTERPRISE SEARCH 10g	**Oracle Secure Enterprise Search 10g**: a standalone product from Oracle, enables a secure, high quality, easy-to-use search across all enterprise information assets.
SAP	**SAP** NetWeaver Search and Classification (TREX): finds information in both structured and unstructured data. TREX provides SAP applications with services for searching and classifying large collections of documents.
SAIC TeraText	**TeraText Suite:** Most data resides in semi-structured, primarily textual documents, not in structured, organizational repositories. Teratext is designed for text-rich data repositories.

Vivisimo Clustering Engine: developed by scientists based upon a mathematical algorithm and deep linguistic knowledge to find relationships between search terms and bring them to light. (Web search:Clusty)

Guruji.com: India - an Indian Internet search engine that is focused on providing better search results to Indian consumers, by leveraging proprietary algorithms and data in the Indian context.

Rediff: India - India's leading internet portal for news, mail, messenger, entertainment, business, mobile, ecommerce, shopping, auctions, search, sports and more.

Naukri.com (India): An India-focused job search engine.

WebMD: A source for health information, a symptom checklist, pharmacy information, and a place to store personal medical information. As the leading US Health portal, it scores over 40 million hits per month.

YouTube: Owned by Google, the web's largest media site. This search will search through the videos of YouTube only.

FindSounds: Search engine to find any kind of sound file: WAV, MP3, AIFF, AU - search by sample rate and quality... a great place to find those sound effects.

Ask Jeeves was designed to allow users to get answers to questions posed in everyday, natural language. Ask.com was the first such commercial question-answering search engine for the Web.

INTEXT QUESTIONS 18.4

1. What are the features of WebMD?
2. Explain the search engine Ask Jeeves.
3. Write down the features of Dieselpoint.

18.7 SEARCH ENGINES AND CATEGORIES OF RESOURCES

The search engines on internet have categorized the information sources on the basis of types or format. For example, the Google categorises the information sources as maps, image, news, scholar, scholarly papers, and many more. The details of the categories are available at http://www.google.co.in/intl/en/about/products/. The Yahoo has categorised information in web, video, news and local categories.

Searching information on web under these categories has become effective using category specific search. A search was conducted using this feature of the Google search engines, to find the effectiveness of the engine in search output. The query was 'NIOS, NOIDA, India'. The results are shown below:

(a) Under Map category, we got location of the NIOS, NOIDA, Uttar Pradesh, India.

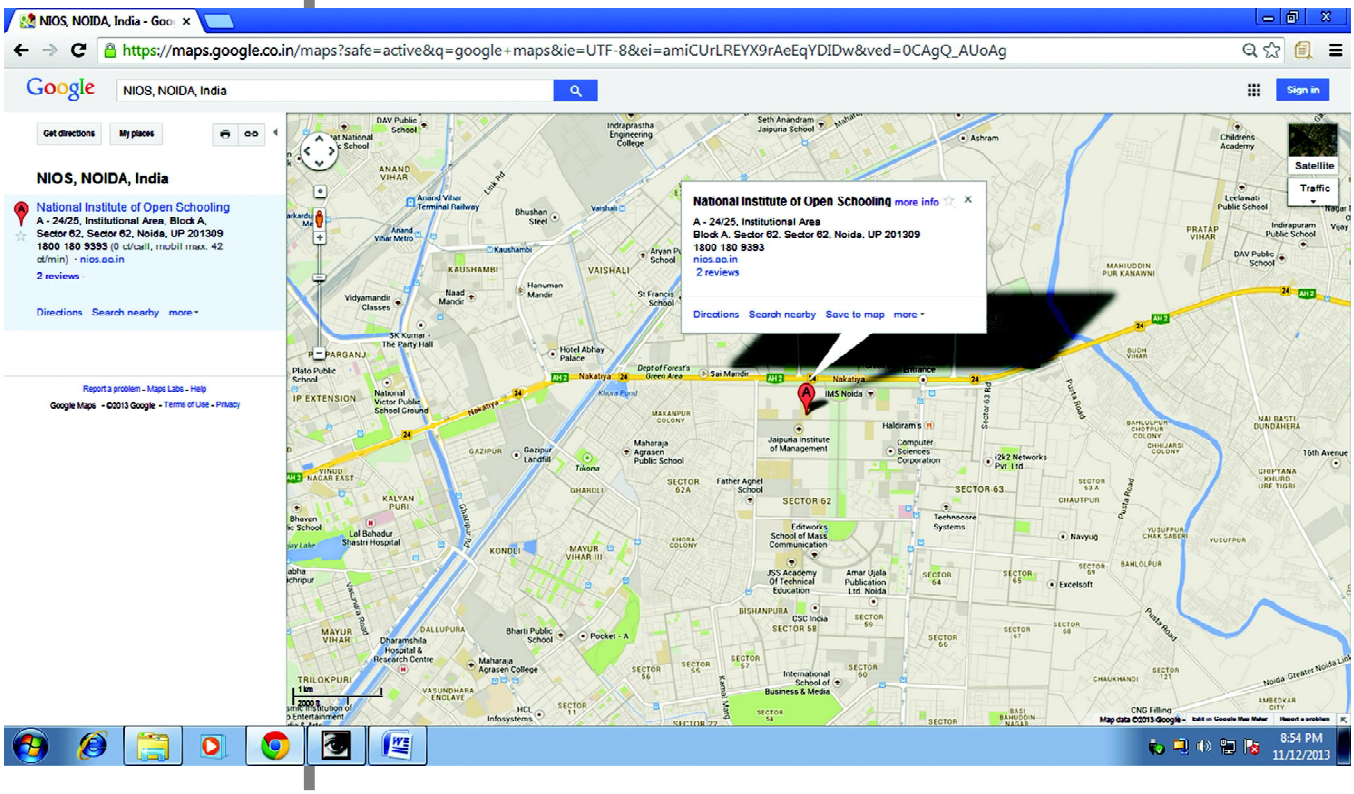


Fig. 18.9: Google's search output of 'NIOS, NOIDA, India' in map category

(b) Under image category, we got the search output of the sources which have images of the NIOS.

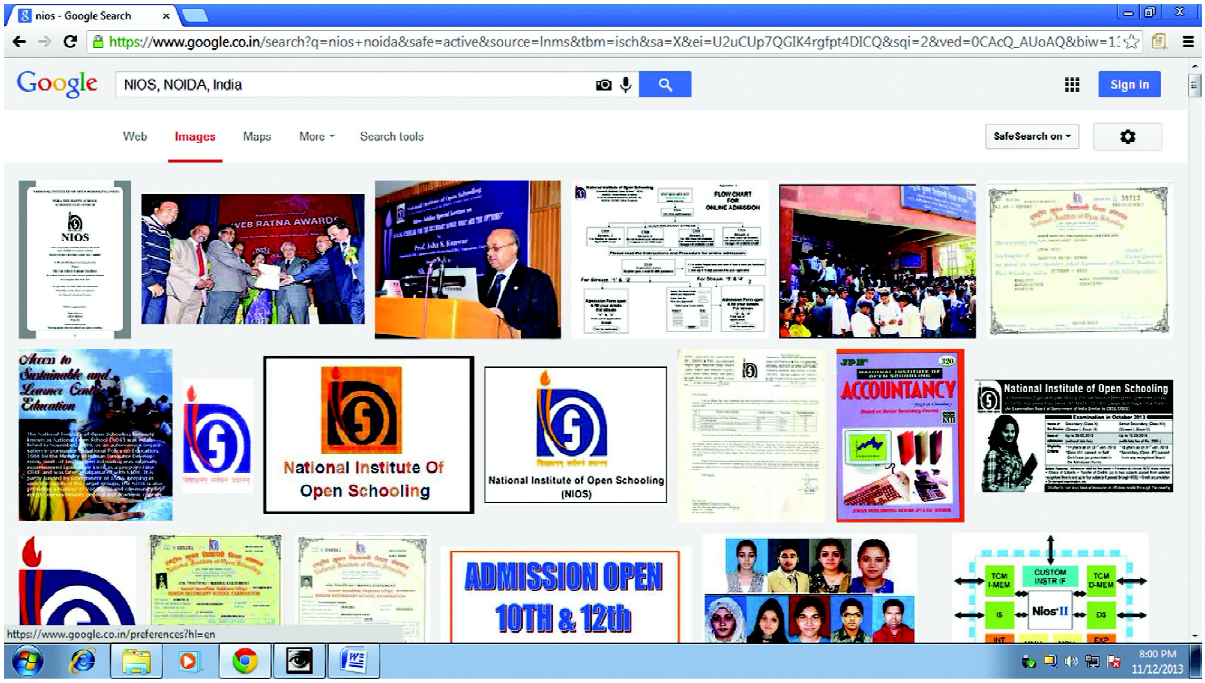


Fig. 18.10: Google's search output of 'NIOS, NOIDA, India'in image category

(c) Under scholar category, there were 34 hits 'NIOS, NOIDA, India'. After evaluating the articles, it was found that the articles contained this string in their text.

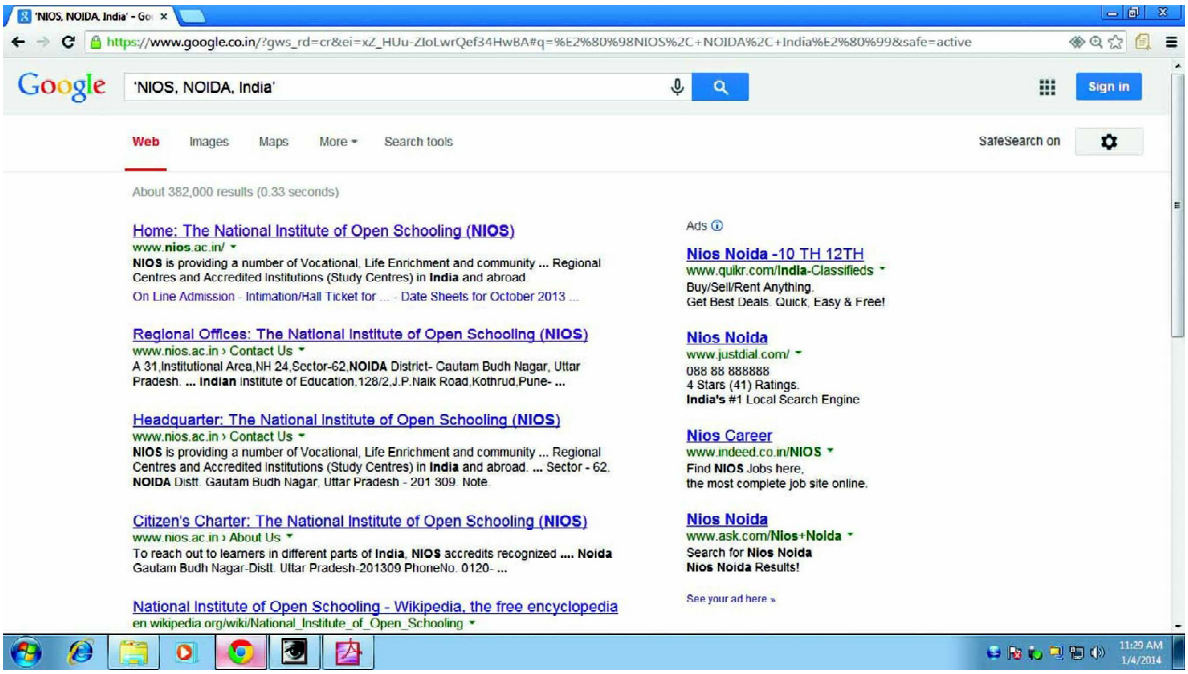


Fig. 18.11: Google's search output of 'NIOS, NOIDA, India' in scholar category

(d) Under web category, there were 293,000 hits for 'NIOS, NOIDA, India'. After evaluating a few of these references, we found that all the terms of the string appeared in the text of the web pages.

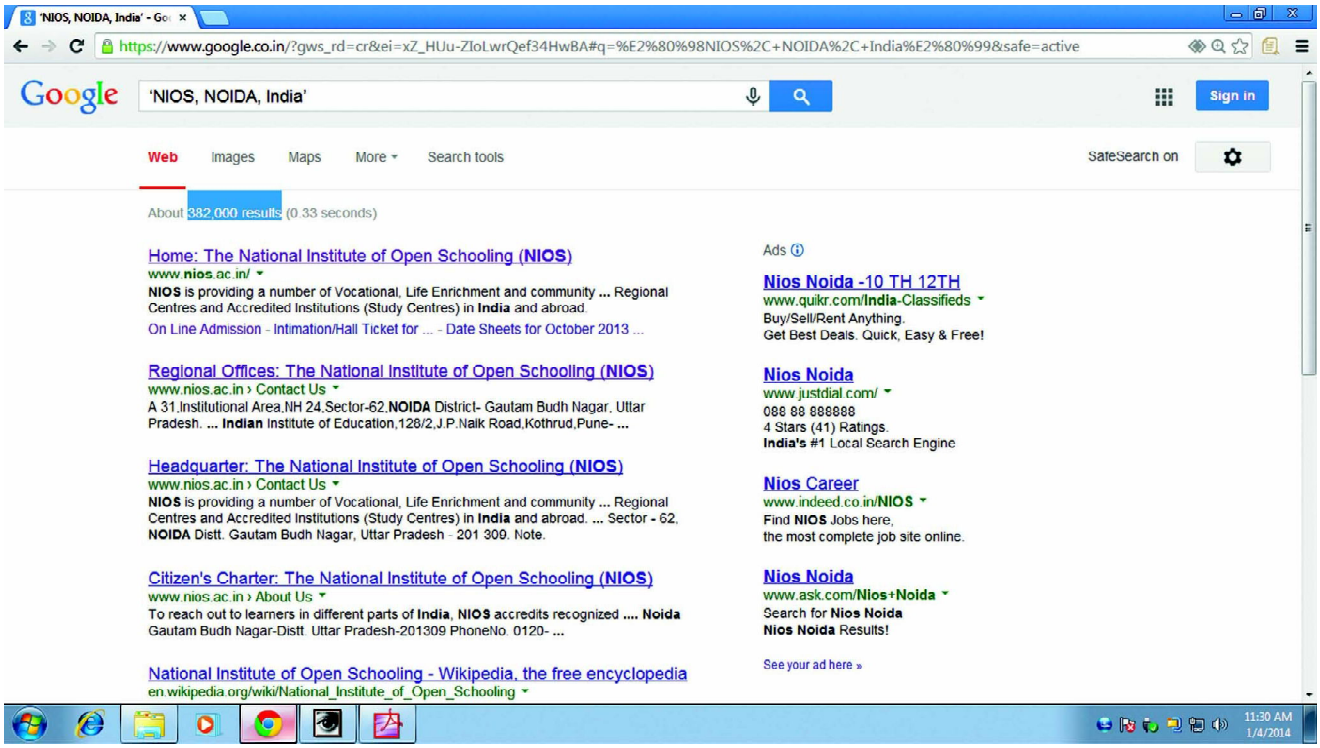


Fig. 18.12: Google's search output of 'NIOS, NOIDA, India' in web category

The analysis of these results shows that as per the need of the search, we should conduct category based search to get effective results. The category could be decided on the basis of the need of the search.

18.8 SEARCH ENGINES IN VERNACULAR LANGUAGES

Internet is the largest library in the world, if seen from the point of view of a library. There are hundreds and thousands of information sources available on internet in languages other than English. With the invention of the UNICODE, the web compatible documents are created in different languages and put on web. Therefore, now the search engines are providing facility to search information using vernacular languages. For example, the 'Google.co.in' which is the default page in India, provides search facilities in Hindi, Bengali, Telugu, Marathi, Tamil, Gujarati, Kannada, Malayalam and Punjabi other than English language. The effectiveness of the search engines can be evaluated by the people of concerned language. The official blog of Google says that, it provides search facility in forty languages. Hence, people of different languages can use the search engine for finding information on internet in their own language.

INTEXT QUESTIONS 18.5

1. What kind of information is covered by Google Scholar?
2. Name a few Indian languages in which Google can Search?

18.9 EFFECTIVE SEARCHING ON INTERNET

For getting effective results from the search engines, a searcher should have knowledge of behaviour and the features of the search engines. Almost all the search engines on internet use operators in one or another form. The operators, truncation signs and symbols or any others tools used by the search engines, should be first understood by the searcher before designing the search query and searching information. For this purpose, it is suggested that searchers should read the guidelines provided by the search engines on their websites. For example, the Google provides a dedicated website where, the guidelines for searchers have been given(http://support.google.com/websearch/?hl=en). 'Tips for search' and 'Explore Google search' are two important sections that help searchers to search information on internet effectively. Other search engines, like Yahoo, Ask, Bing, etc. also provide help to the searchers for effective search. The search process on internet can best be described with the help of a diagram.

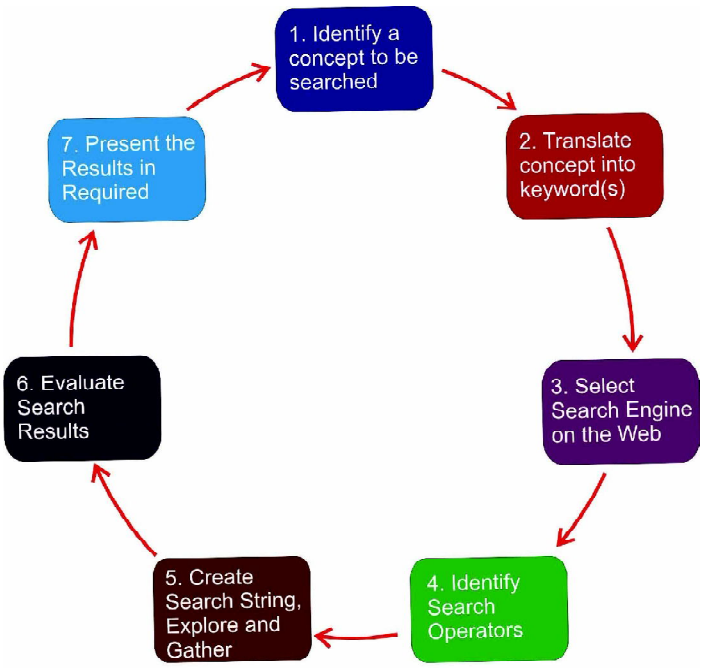


Figure 18.13: Diagram of search process on Internet

The Fig. 18.13 shows the process of searching information on internet in seven steps. The steps can be described as:

Step –1: Identify whole concept on which you want to conduct a search. Choose some keywords, subject descriptors, and/or sentences, and the types of information sources you are looking for.

Step –2: Translate the concept,terms, phrases, their alternatives and associated terms or subjects, etc. into keywords.

Step-3: Select information domain to be searched, like your library, OCLC, LISA, PubMed, etc., or suitable search engine on internet.

Step-4: Identify operators and symbols used by the search engine of the information domain.

Step-5: Create a search string using operators and/or connectors and conduct the search.

Step-6: Evaluate the result and if needed, modify the search by choosing alternate terms. Expand the scope of the subject or topics and put more parameters like date of publication, language, form, source, etc. provided in the search engines.

Step-7: Present the search results in a user friendly format.

These seven steps can make a searcher expert in finding information from Internet. The experiments with search query, permutation combination of the terms in the query, using set of symbols and analysis of the results for each and every query, as well as the impact of the change in the string, can make a searcher expert in searching the information on Internet. It is always recommended that, the searcher should use at least two search engines to get effective and relevant results.

WHAT YOU HAVE LEARNT

- Internet has become the largest repository of information in the world. It stores all kinds of information in digital form.
- There are three main purposes of searching information on internet, namely (a) Navigational Search, (b) Informational Search, and (c) Transactional Search.
- Searching information on internet is both an art and a science.
- Simple search technique is a mechanism of finding information on internet

for beginners in which one selects a search engine and enters a simple query and searches the information.

- Advanced search technique is a mechanism where one can use operators and other parameters provided by the search engines to get precise results.
- Boolean Logic uses 'AND', 'OR', 'NOT' as operators to express the concepts to be searched in terms of string to be understood by the search engines.
- There are a number of search engines on internet. One can select the search engine as per the requirements of the search.
- For getting better results, search engines have categorized information resources in different categories as Map, Images, Books, Articles, etc.
- Information resources are available on internet in languages other than the English language also. Therefore, the search engines are available in Indian languages also like Tamil, Malayalam, Bengali, etc.

TERMINAL QUESTIONS

1. Why a search engine is needed on internet?
2. Discuss the purposes of search on Internet.
3. Discuss features of advanced search techniques.
4. How do the Boolean operators work?
5. What do you understand by search engines in vernacular language?
6. Discuss various steps of effective search on internet.

ANSWERS TO INTEXT QUESTIONS

18.1

1. A search engine is a software program that searches for sites based on the words that we refer to as search terms. Internet search engines are thus special sites on the Web that are designed to help people find information stored on other sites. There are differences in the ways various search engines work.
2. With an ad hoc search, the searcher's goal is to find as many relevant documents as possible about a topic. An ad hoc search is informational in

nature, as searcher is looking for information about a subject from vast resources. The searcher may or may not have previous knowledge about the topic but, wants to read or learn more about it.The intent of the informational search is to acquire some information, assuming it is available on the internet.

18.2

1. Simple search technique is a mechanism for finding information on internet on the basis of a query submitted to search engine in term of simple keywords.
2. Advance search technique is a searching mechanism where different parameters are used for getting precise search results.

18.3

1. The purpose of the 'AND' operator is to find the information source where both the terms connected by this operator exist.
2. The purpose of the 'OR' operator is to find the information source where either of the two terms connected by this operator exist.
3. The purpose of the 'NOT' operator is to find the information source where first term exists but not the second term.

18.4

1. WebMD is a source for health information, a symptom checklist, pharmacy information, and a place to store personal medical information. As the leading US Health portal, it scores over 40 million hits per month.
2. Ask Jeeves is designed to allow users to get answers to questions posed in everyday natural language. Ask.com was the first such commercial question-answering search engine for the Web.
3. Dieselpoint provides advanced full-text search with data navigation capability. It gives users highly relevant results not possible with either traditional search engines or SQL databases.

18.5

1. The Google Scholar covers the full text of scholarly literature.
2. Indian languages in which Google can search are Hindi, Bengali, Gujarati, Marathi, Kannada, Punjabi, etc.

GLOSSARY

Browse: To browse through a web page is exploring what's there and seeing where the links take you. When you browse, you have to guess which words and links on the page pertain to your interests.

Browsers: Software programs that enable you to view web pages and other documents on the Internet. They "translate" HTML-encoded files into the text, images, sounds, and other features you see. The most commonly used browsers are Microsoft Internet Explorer, Firefox, Mozilla, Safari, Opera, and Chrome.

HTML: Hypertext Markup Language.

ISP: Internet Service Provider - a company that sells direct access to the Internet.

LCOC: Library of Congress Online Catalog

Meta-Search Engine: Search engines that automatically submit your keyword search to several other search tools, and retrieve results from all their databases.

OPAC: Online Public Access Catalogue

PubMed: It is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics.

Server/Web Server: A computer running that software, assigned an IP address, and connected to the Internet so that it can provide documents via the World Wide Web.

Site/Web-Site: This term is often used to mean "web page," but there is a difference. A web page is a single entity, one URL, one file that you might find on the Web. A "site," properly speaking, is a location or gathering or centre for a bunch of related pages linked to form that site.

UNICODE:A set of standard coding schemes intended to replace the multiple coding schemes currently used worldwide. The Unicode Consortium developed the original standard, Unicode Transformation Format-16 (UTF-16), in 1991 as a standard coding scheme to support multiple complex alphabets such as Devanagri (Hindi), Bengali, Chinese, Japanese, Korean, etc.

URL: Uniform Resource Locator. It is the unique address of any Web document.

Vernacular Language: A vernacular language is the native language or native dialect of a specific population, community or region.

WWW: The World Wide Web

Notes

WEBSITES

http://www.thesearchenginelist.com/

http://www.ncbi.nlm.nih.gov

www.swse.org

http://www.w3.org/standards/semanticweb/

http://www.google.com

http://www.yahoo.com