3.1 Introduction

Perhaps the most important development of the last decade has been the rise of the Internet. The development of protocols joined with the use of transmission devices permits large-scale interconnection between computers. The growth rapidly becomes worldwide; before 2003 the USA dominated the Web in both addresses and pages, but, other countries, particularly in Asia, are now increasing their presence. Online content has flooded onto the Web from publishers, libraries, museums, research scholars, and practically anyone else with keyboard or scanner. A uniform interface has encouraged the proliferation of online content. For instance, in mid 2003 the search engine giant Google claimed to be indexing more than 3 billion Web pages (www.google.com). Today the Web has more than 150 terabytes of text, more than all the very largest research libraries.

The Web was started by Tim Berners-Lee in 1990. The interface followed five years later and it took a while for enough materials to accumulate for an interface to become important. The years 1995 to 1997 saw an explosion of public interest in the Internet and of online materials and public access. Virtually, there were no Web pages in mid 1993 and by January 1996, 75,000 hosts with names beginning www,
and by the end of 2002 there were 100 million such hosts (Gulli and Signorini, 2005).

3.2 World Wide Web

The current Web can be traced back to a project at CERN, the European Particle Laboratory in Geneva, Switzerland, in 1989 when Tim Berners-Lee and Robert Cailliau built ENQUIRE (short for Enquire Within Upon Everything) (Berners-Lee, 1998). In that time, CERN was the largest Internet node in Europe, and Tim Berners-Lee, the creator of the World Wide Web (WWW), saw an opportunity to join hypertext with the Internet. He used similar ideas to those underlying the ENQUIRE system to create the WWW, for which he designed and built the first Web browser and editor (called WorldWideWeb and developed on NeXTSTEP) and the first Web server called httpd (short for HyperText Transfer Protocol daemon). The first Web site built was at http://info.cern.ch/ and was first put online on August 6, 1991. It provided an explanation about what the WWW was, how one could own a browser and how to set up a Web server. It was also the world's first Web directory, since Berners-Lee maintained a list of other Web sites apart from his own.

The WWW is a hypertext system that operates over the Internet. Hypertext is browsed using a program called a Web browser, which retrieves pieces of information (called "documents" or "web pages") from Web servers (or "web sites") and displays them on the screen. We can then follow hyperlinks on each page to other documents or even send information back to the server to interact with it. Every file or page on the Internet whether it is a Web page, telnet site, etc. has an Internet address. Web addresses are called Uniform Resource Locators (URLs) and
they specify a file’s unique location on a computer connected to the Web. The act of following hyperlinks is often called "surfing" the web. Instead of question answering, information retrieval of whole texts becomes common with the boom of the Web. Thousand of people and institutions have posted valuable information on the Web, and the volume is increasing day by day with an enormous rate.

What we’ve seen is people finding interesting new ways to use the Web to showcase their information and their expertise; particularly in niches in all kinds of subjects where it's really just opened the door to new uses of the Web. Whether it is sharing photographs on Flickr.com, showing off an amateur video on YouTube, or looking for a mate on Match.com, Web sites have also become a way to bond and belong. In both the business world and the social scene, a Web site is now an identifier almost as common as a phone number or an e-mail address.

3.2.1 World Wide Web: Growth and Change

The growth of the World Wide Web is an unprecedented phenomenon. Fours years after the Web’s birth in 1990, a million or more copies of the first well-known Web browser, Mosaic, were in use (Abbate, 1999). This growth was a result of the exponential increase of Web servers and the value and number of Web pages made accessible by these servers. In 1999 the number of Web servers was estimated at about 3 million and the number of Web pages at about 800 million (Lawrence & Giles, 1999), and three years later, in June 2002, the search engine AlltheWeb (www.alltheweb.com) announced that its index contained information about 2.1 billion Web pages. And the most popular search engine Google has indexed more than 4 billion pages as on 2004 (www.google.com). Measuring the Web is a
difficult task. However, different Web experts have attempted to calculate the size and growth of the web. Two different types of measurement may be notified as:

i) the number of websites, and

ii) the number of Web pages

An OCLC project that measure the growth of the Web by counting the number of Web sites reported that there were 9,040,000 websites in the year 2002 (Online Computer Library Center, 2004). According to Hois.net, a domain based research services, 47,501,249 domains have been registered till December 2004 (http://www.hois.net). The ISC Internet Domain Survey has identified 433,193,199 numbers of hosts in January 2007 and again 600,000,000 numbers of hosts in July 2007 (Internet Systems Consortium, 2007). However, Netcraft presented a totally different number in February 2007, saying that there was at the time around 109,000,000 Web sites with domain names, out of which 47 to 48 million were active sites. Recently CNN has reported that there are 100 million operating websites in the Internet. The Web's growth has been accelerating, there were just 18,000 Web sites when Netcraft, based in Bath, England, began keeping track in August of 1995. It took until May of 2004 to reach the 50 million milestone; then only 30 more months to hit 100 million, late in the month of October 2006. The February 2007 figure from Netcraft is 109,000,000 (Pandia Search Engines News, 2007). The bottom line here is that the more unwieldy the Internet becomes, the more central search becomes as the main navigational tool. All these studies leads us to estimate that at present the number of Web pages must be somewhere between 15 and 30 billion — and probably closer to the latter than the former, given the increase in the number of Web users, bloggers and companies moving online since 2000. It is
also worth mentioning here that there are over 1 billion Internet users in 2007, as counted by the Internet World Statistics (Miniwatts Marketing Group, 2007). The following table shows a range of approaches to sizing the Web as of 2004.

**Table – 3.1**

Methods for Sizing the World Wide Web

<table>
<thead>
<tr>
<th>Company</th>
<th>Domains</th>
<th>Method</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.whois.net">www.whois.net</a></td>
<td>47,501,249</td>
<td>Domain name registration, this number changes continuously</td>
<td>Continuously</td>
</tr>
<tr>
<td><a href="http://www.netcraft.com">www.netcraft.com</a></td>
<td>56,923,737</td>
<td>Web servers responding to HTTP request. Each domain name is counted as a server</td>
<td>Monthly</td>
</tr>
<tr>
<td><a href="http://wcp.oclc.org/">http://wcp.oclc.org/</a></td>
<td>9,040,000</td>
<td>IP addresses responding to HTTP request</td>
<td>Yearly</td>
</tr>
</tbody>
</table>

*Source: (www.sims.berkeley.edu)*

The number of pages on the Web can be estimate by looking at the number of pages indexed by different Search Engines. As of March 2003, Google is reported to be the largest Search Engine, having indexed over 3.8 billion Web pages (www.searchenginewatch.com). However, the total number of pages on the Web is much more than the highest number of Web pages indexed by any particular search engine. Netcraft says the United States, Germany, China, South Korea and Japan show the greatest Web site growing spurts. Today there are seemingly endless Web sites for shopping, social networking, and, of course, sleaze. When the Web was started, it was started as a mechanism for sharing high energy particle physics data. The creator of that Web site, Tim Berners-Lee, wanted experts to be able to share data on particle smashing, even if they weren't at CERN in Switzerland where he
was doing research. Soon, a Web "explosion" took place when businesses realized they could use the Internet to make money. Web sites begin to be incorporated into advertising. And by the mid-'90s the cost of personal computers had fallen enough so that the Internet began entering peoples homes and schools as well as their workplaces. The cost and the complexity of creating Web sites have both diminished since the beginning of the 21st century. Computer users no longer have to be experts in HTML, or hypertext markup language, to be masters of their own Web sites.

With the increasing of the Web resources, the number of Internet penetration is also flying high in worldwide. In May 2006, the ComScore Networks, a leader in measuring the digital age, announced the launch of ComScore World Metrix, the first true estimate of global online audience size and behaviour based on activity from the world's largest online behavioral research panel. The company announced that 694 million people, age 15+ used the internet world wide from all location in March 2006, representing 14 percent of world's total population (http://www.comscore.com). Notably, ComScore survey includes measurement of major Asian countries, China, Japan, India, and Korea which represent nearly 25 percent of the total world wide online population. In March 2007, the number of world wide online user increase to 747 million people. Among the top fifteen countries, ranked by penetration, Internet audience in India, Russian Federation and China increased the most in 2006. China now represents the second largest internet population in the world, with 86.8 million users, after the U.S. which rose two percent year-over-year to 153.4 million users in January 2007. The following chart shows the top fifteen countries by Internet penetration in between 2006-2007.
### Table 3.2

**Top 15 Countries by Internet Penetration (January 2007 vs January 2006)**

<table>
<thead>
<tr>
<th>Country</th>
<th>Jan-06 (000)</th>
<th>Jan-07 (000)</th>
<th>Percentage Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worldwide</td>
<td>676,878</td>
<td>746,834</td>
<td>10%</td>
</tr>
<tr>
<td>United States</td>
<td>150,897</td>
<td>153,447</td>
<td>2%</td>
</tr>
<tr>
<td>China</td>
<td>72,408</td>
<td>86,757</td>
<td>20%</td>
</tr>
<tr>
<td>Japan</td>
<td>51,450</td>
<td>53,879</td>
<td>4%</td>
</tr>
<tr>
<td>Germany</td>
<td>31,209</td>
<td>32,192</td>
<td>3%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>29,773</td>
<td>30,072</td>
<td>1%</td>
</tr>
<tr>
<td>South Korea</td>
<td>24,297</td>
<td>26,350</td>
<td>8%</td>
</tr>
<tr>
<td>France</td>
<td>23,712</td>
<td>24,560</td>
<td>4%</td>
</tr>
<tr>
<td>India</td>
<td>15,867</td>
<td>21,107</td>
<td>33%</td>
</tr>
<tr>
<td>Canada</td>
<td>18,332</td>
<td>20,392</td>
<td>11%</td>
</tr>
<tr>
<td>Italy</td>
<td>15,987</td>
<td>18,106</td>
<td>13%</td>
</tr>
<tr>
<td>Brazil</td>
<td>12,845</td>
<td>14,964</td>
<td>16%</td>
</tr>
<tr>
<td>Spain</td>
<td>12,206</td>
<td>12,710</td>
<td>4%</td>
</tr>
<tr>
<td>Russian Federation</td>
<td>10,471</td>
<td>12,707</td>
<td>21%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>10,772</td>
<td>11,077</td>
<td>3%</td>
</tr>
<tr>
<td>Mexico</td>
<td>8,624</td>
<td>10,149</td>
<td>18%</td>
</tr>
</tbody>
</table>

*Source: ComScore World Metrix*

#### 3.2.2 The Invisible Web

The part of the Web that is hidden and cannot be easily access by the so-called Search Engines is called the “Deep Web”. Deep Web sources store their content in searchable databases, which can only be produced dynamically in response to a direct request. The deep Web contains nearly 7500 terabytes of information compared with 19 terabytes of information on the surface web. Public information on the deep Web is currently 400 to 550 times larger than the commonly defined World Wide Web (www.brightplanet.com). The visible Web is what can be searched and therefore be seen by using search engines and directories. However, the contents of databases rarely show up in search engine results. Search engine spiders cannot or will not go inside database tables and extract the data. Database content is therefore ‘invisible’ to them. Pages that cannot found in this fashion are part of the so called ‘invisible web’. Because of the fact that many answers may be found through a simple search with one of the main search engines, we tend to
forget that there is also this something variously called the invisible web, the hidden web, the deep web. There are countless hidden treasures out there that cannot be found with conventional search engines.

The terms used in a search query are sent into that specialized database, and are returned in another Web page that is dynamically generated for the answer. It is not retained anywhere after the search. Search engines cannot access such dynamically generated pages because the computer robots or spiders that build them cannot type the searches needed to generate the pages. Spiders find pages by visiting all links in the pages they ‘know about’. Unless there are links somewhere that the spiders can use to regenerate specialized database searches, the contents of the database cannot be reached by the spiders. Pages requiring passwords to access them are also closed to search engines, because spiders cannot type in the required password needed to access the page.

**Figure 3.1**
An illustration of different parts of the invisible web

*Source: Ford and Mansourian (2006)*

52
The content of search engines on the Web is itself stored in databases and available only through user query. The whole terminology doesn’t seem to be very logical. It is not strictly speaking invisible; it is just that a user query is necessary to find it. The term ‘deep web’ was first used by a company called BrightPlanet to describe the phenomenon of searchable databases on the web. This is much better since database content is visible with the appropriate search and retrieval technology. There are usually considered to be four parts to the invisible web:

i) **Almost invisible Web** - Conventional search engines index only parts of the whole web. From a website with roughly 200 pages, only about 50 are indexed, therefore 150 pages remain in the almost invisible web. This measure is sometimes called the ‘depth’ of crawling.

ii) **Gated Web** - The gated web consists of pages that require a registration, either with or without costs involved.

iii) **Professional Web** - Professional, multidisciplinary databases are usually subscription-based. They may be accessed through a browser interfaces. Some well-known professional databases include: ISI Web of Knowledge.

iv) **Vanishing Web** - The vanishing web describes web pages that are temporarily unavailable or simply do not exist any longer.

Another type of information that falls into the category of the deep web is information that is dynamically changing. These are like, News; Job Postings; Available airline flights and last-minute hotel deals; Stock exchange figures; etc.
3.2.3 Access to Information on the Web: The Tools

A user can get access to any websites by entering the URL (uniform resource locator; the address of a Web site) on the browser. A Web browser, like Netscape Navigator, Mozilla FireFox or Microsoft Internet Explorer, is a computer program, an essential tool for getting access to the web. A Web browser performs two major tasks:

i) It knows how to go to a Web server on the Internet and request a page so that the browser can pull the page through the network and into your machine.

ii) It knows how to interpret the set of HTML tags within the page in order to display the page on your screen as the page’s creator intended it to be viewed.

Although one could get access to any Web page by typing on the browser the URL of a sought website, and then moving into the site through the various links, there are several problems to his approach, especially when the user is interested to get some specific information on a given topic, or find answers to a given question. Problems arose because it is almost impossible for user to know which of the billions of Web pages contain the information they require and which of the millions of websites contains the required Web page. In order to solve these difficulties, several Web search tools have been developed that assist users in finding the information they need from the right Web page with relatively little effort.

There are basically two ways to find information on the web: by conducting a search using what is known as a search engine, or by following the links in a specially
designed list called a directory. Search engines allow user to enter search terms—Keywords and/or phrases—that are run against a database containing information on Web pages collected automatically by programs called spiders. The search engine retrieves from its database Web pages that match the search terms entered by the searcher. It is important to note that when a user conducts a search using a search engine, the search does not search for the information across the entire Web at the given instance. Instead it searches a fixed database, which is updated at regular intervals according to specific criteria employed by the search engine, located at the search engine’s website and containing information on selected Web pages.

3.3 Information Retrieval Systems

The term information retrieval was coined in 1952 and gained popularity in the research community from 1961 onwards. At that time information retrieval’s organizing function was seen as a major advance in libraries that were no longer just storehouses of books, but also places where the information was catalogued and indexed. Subsequently, with the introduction of the computers in information handling, there appeared a number of databases containing bibliographic details of documents, often coupled with abstracts, keyboards, etc., and consequently the concepts of information retrieval came to mean retrieval of bibliographic information from stored document databases. These information retrieval systems were document retrieval systems, since they were designed to retrieve information about the existence or non-existence of bibliographic documents relevant to a user query. Modern information retrieval systems can either retrieve bibliographic items, or the exact text that matches a user’s search criteria from a stored database of full texts of documents. Although information retrieval systems originally meant text
retrieval systems deal with multimedia information comprising text, audio, images and video, while many features of conventional retrieval systems are equally applicable to multimedia information retrieval.

The concept of information retrieval presupposes that there are some documents or records containing information that have been organized in an order suitable for easy retrieval. The documents or records we are concerned with contain bibliographic information, which is quite different from other kinds of information or data. The major objective of a bibliographic information retrieval system, however is to retrieve the information – either the actual information or the documents containing the information- that fully or partially match the user’s query. The database may contain abstracts or full texts of documents, like articles, handbooks, dictionaries, encyclopedias, legal documents, statistics, etc., as well as audio, images and video information. Whatever may be the nature of the database- bibliographic, full text or multimedia- the system presupposes that there is a group of users from whom the system is designed. Users are considered to have certain queries or information needs, and when they put forward their requirement to the system, the latter should be able to provide the necessary bibliographic references of those documents containing either the required information, or the actual text in the case of a full-text retrieval system. An information retrieval system is designed to retrieve the documents or information required by the user community. Thus, an information retrieval system aims at collecting and organizing information in one or more subject areas in order to the user as soon as asked for. The following situation clearly reflects the purpose of information retrieval systems:
i. A writer presents a set of ideas in a document using a set of concepts

ii. Somewhere there will be some users who require the ideas but may not be able to identify those.

iii. Information retrieval systems serve to match the writer's ideas expressed in the document with the users' requirements or demands for those.

Thus, an information retrieval system serves as a bridge between the worlds of creators or generators of information and the users of that information.

3.3.1 Trends in Information Retrieval Systems

Information retrieval covers a vast area of study, and it is therefore difficult to keep track of the latest developments and consequently the trends in research in this field. Thousands of research papers are published every year in a number of research journals, in seminar, conference and workshop volumes on different areas of library and information science, computer science, electronic engineering, export systems and artificial intelligence, linguistics, psychology, and so on.

Mechanization of information storage and retrieval using computer technology has a history that is about as long as that of the computer itself. Probably the first person to think about computer-aided solutions of information retrieval problems was Robert Fairthrone who, in the early 1950s, investigated the use of punched card equipment for the retrieval of bibliographic information. In the early 1960s people proposed techniques, systems and models for information retrieval but most of this research did not produce fruitful results because the computational power available then was not adequate for all these tasks. A substantial amount of scientific work
was done in the late 1960s and early 1970s experimenting on a number of information retrieval problems. In the 1970s a number of mathematically sophisticated information retrieval models were developed, such as vector space and probabilistic models for retrieval. These experimental investigations yielded concrete evidence that relevance feedback was an effective way of enhancing into the 1980s. During that decade it was realized that retrieval methods such as string searching, keyword searching or searches using keyword frequency information can be computationally efficient but they may not always produce the desired results. So interest shifted back to examining natural language processing techniques as a way of improving retrieval performance when the stored information is largely textual. Several research studies have been undertaken that aim to improve the process of searching and retrieval in an information retrieval environment.

3.3.2 Components of an Information Retrieval System

An information retrieval system has three major components, i.e. the document subsystem, the user's subsystem, and the searching/retrieval subsystem. These divisions are quite broad and each one is designed to serve one or more functions, such as, analysis of documents and organization of information creation of a document database, analysis of users' queries, preparation of a strategy to search the database, actual searching or matching of users' queries with the database, and finally, retrieval of items that fully or partially match the search statement. It is evident that on the one side of an information retrieval system there are the documents or sources of information and the other there are the users' queries. These two sides are linked through a series of tasks. In order to carry out these series
of tasks an information retrieval system comprises the following six major subsystems:

i) the document subsystem  
ii) the indexing subsystem  
iii) the vocabulary subsystem  
iv) the searching subsystem  
v) the user-system interface, and  
vi) the matching subsystem.

The broad outline of an information retrieval system is shown in the following figure.

**Figure - 3.2**  
Board outline of an Information Retrieval System

![Diagram of Board outline of an Information Retrieval System]

All the tasks mentioned in the above figure can be brought under two major groups—subject/concept analysis, and search and retrieval. Subject or Content analysis includes the task related to the analysis, organization and storage of information.
The process of search and retrieval includes the tasks of analyzing users' queries, creation of a search formula, the actual searching, and retrieval of information. Researchers in the information retrieval world are engaged in developing suitable methodologies for both sets of operation. Developments in the technological world, especially in computer and communication technology, have provided additional inputs to the development of information retrieval systems. Researchers who are working on the storage side of the information retrieval system are engaged in designing sophisticated methods for identification and representation of the various bibliographic elements essential for documents, automatic content analysis, and text processing, and so on. On the other hand, researchers working on the retrieval side are attempting to develop sophisticated techniques, user interfaces, and various techniques for producing output for local as well as remote users. The recent emergence of the Internet, particularly the World Wide Web has made a significant impact on the information retrieval environment.

3.4 Web Information Retrieval

The goal of information retrieval is to find all documents relevant for a user query in a collection of documents. Decades of research in information retrieval were successful in developing and refining techniques that are solely word-based. With the advent of the web new sources of information became available, one of them being the hyperlinks between documents and records of user behavior. Hyperlinks provide a valuable source of information for web information retrieval. The introduction and growth of the WWW have brought significant changes in the way we access information. The development of the Web began in 1989 by Tim Berners-Lee and his colleagues at CERN. They created a protocol, called the Hyper Text
Transfer Protocol (HTTP), which standardized communication between servers and clients. The Web gained rapid acceptance with the creation of a Web browser called Mosaic. Mosaic allowed people to use the Web using the 'point-and-click' graphical manipulations. Subsequently the Mosaic staff started their own company and developed one of the most popular Web browsers, Netscape Navigator. Another popular browser soon appeared, Internet Explorer from the Microsoft Corporation.

Web searching is becoming an indispensable part of the daily life of many groups of people. Among all web-based applications, Web searching is one of the most common and important ones. However, satisfying the information needs on the Web is not always an easy and straightforward process. Information retrieval through the Web is a multifaceted process and many side elements may affect on this procedure. In spite of all remarkable progress in developing more sophisticated information retrieval tools on the Web, people may still encounter many difficulties for locating information they need for meeting their information needs. Although developing web-based search tools has been inspired by pre-web online information retrieval systems, a couple of unique features of the Web environment and the wide range of Web users characteristics impose new condition for searching and retrieval situation.

Retrieving information from the Web is becoming a common practice for internet users. However, the size and heterogeneity of the Web challenge the effectiveness of classical information retrieval techniques. Since the late nineteen-nineties Web Information Retrieval has attracted the attention of both researchers and commercial companies and has been fueled by theoretical results, methods, experiments, and implementations. There are different methods for information retrieval through the Web environment; however the most common way for information retrieval in this
environment is searching through different web-based search facilities including various search engines.

3.4.1 Traditional vs Web Information Retrieval

Information retrieval systems have arguably now become a global tool for information access. Traditionally, the primary users of IR systems have been information professionals who have acted as intermediaries for clients with information needs. Much of the content of these systems has related to published works in the form of document surrogates, such as bibliographic citations with abstracts or, increasingly, full text documents. This is the traditional context for and utilization of IR systems. Web information retrieval is significantly different from traditional text retrieval systems. There differences mainly stem from a number of typical characteristics of the Web such as its distributed information and users, and so on. This section discusses some of these issues with a view to highlighting the complexities of Web information retrieval.

i) Distributed nature of the Web: Web resources are distributed all over the world, so complex measures are required to locate, index and retrieve them. The fact that the computers that are interconnected have different architecture and the information resources are created using different platforms, software and standards, makes the matter more complex. Most text retrieval systems deal with a set of information resources that is several times smaller in volume than the web. In addition, text retrieval systems usually deal with a set of documents that have been created using a set of standards- hardware, software and processing standards. When OPACs retrieve distributed information, they use several standards to process it, such as the MARC
ii) Size and growth of the Web: The Web has grown exponentially from 1990 to 2007. The processes of identifying, indexing and retrieving information become more complex as the size of the web, and hence the volume of information on the web, increases. Conventional text retrieval systems have to be tested and modified to make them suitable for handling the large volume of data on the web.

iii) Deep vs the Surface Web: Information resources on the Web can be accessed at two different levels. While millions of Web information resources can be accessed by anyone, a lot of information is accessible either through authorized accessed (information that is password protected, say) or can be generated only by activating an appropriate program. Researchers call the former the surface Web and the latter the deep web, with a note that the deep Web is several time larger than the surface web.

iv) Type and format of the documents: Text retrieval systems deal with textual information only; the Web contains a much wider variety, from simple text to multimedia information. Again these information resources appear in a variety of formats thereby making the task of index and retrieval more complex.

v) Quality of information: Since anyone can publish almost anything on the web, it is very difficult to assess the quality of information resources. As opposed to conventional text retrieval systems, which deal with published information resources
that have some quality control, Web information retrieval systems have to deal with controlled information resources.

vi) Frequency of change: Web pages change quite frequently. This is in sharp contrast with the input of conventional text retrieval systems, which deal with relatively static information. Once an information resource is added to a text retrieval system it does not change its content. Keeping track of the changes in the millions of Web pages and making necessary change in the information retrieval system is a major challenge. Another major problem with the Web is that the resources (web pages) often move. This information needs to be tracked by the retrieval system in order to facilitate proper retrieval.

vii) Ownership: Information resources that are accessible through the Web have different access requirements; some information can be accessed and used freely, others require specific permission or access rights, often through payment of fees. Identifying the right to access is a major challenge for Web information retrieval.

viii) Distributed users: Most text retrieval systems are designed to meet the information needs of a specific user community. Hence text retrieval systems usually have an idea of the nature, characteristics, information needs, search behaviors, and so of the target user community. Web information is in sharp contrast with this. Ideally the user of an information resource on the Web may be anyone, located anywhere in the world. This imposes a significant challenge since the designer of a Web information retrieval system will have no idea about the target users, their nature, characteristics, location, information search behaviors, and so on.
ix) Multiple languages: Since the Web is distributed all over the world, the language of information resources as well as users varies significantly. An ideal Web information retrieval system should be able to retrieve the required information irrespective of the language of the query or the source information. This diversity of language poses a tremendous challenge for Web information retrieval.

x) Resource requirements: Massive amount of resource are required to build and run an effective and efficient Web information retrieval system. The matter is worsened by the fact that there is no single body who would fund for these resources, and yet everyone wants a good information retrieval system for access to Web information resources.

The dynamic nature of the Web in terms of its growth, and also as a result of changes in the Web pages from additions and deletions to the removal of the entire page, constitutes a major difference between Web and traditional information retrieval. The retrieval services of the Web generally lack the same level of search and display features found in traditional information retrieval systems used to access bibliographic databases. Recent studies have shown that the search characteristics of these systems catering to general audiences are different than traditional information retrieval systems.

There is another important point on which Web information retrieval significantly differs from traditional information retrieval systems, that is the volume of use. The following table shows the number of searches handles per day by eight search
engines in the month of January-February 2003. This may be noted that no conventional information retrieval system is design to handle this many searches per day.

### Table 3.3
Number of searches per day on search engines

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Searches per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>250 million</td>
</tr>
<tr>
<td>Overture</td>
<td>167 million</td>
</tr>
<tr>
<td>Inktomi</td>
<td>80 million</td>
</tr>
<tr>
<td>Looksmart</td>
<td>45 million</td>
</tr>
<tr>
<td>Findwhat</td>
<td>33 million</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>20 million</td>
</tr>
<tr>
<td>Alta Vista</td>
<td>18 million</td>
</tr>
<tr>
<td>Fast</td>
<td>12 million</td>
</tr>
</tbody>
</table>

*Source: (www.searchengine.com).*

#### 3.5 Search Facilities Offered by Most Text Retrieval Systems

The text retrieval software available in the market is characterized by the availability of a wide variety of search and retrieval provisions. However, most software is limited to Boolean and proximity searching, though some offers relevance feedback, and so on. With the introduction of the WWW, a number of retrieval systems, called Web search engines or simply search engines, have been developed that help users search and retrieve information from the WWW. A number of Web search engines are now available, and several publications describe their feature. The following are some of the common search facilities available in all text retrieval packages; it illustrates how the software packages vary in terms of accepting the search expressions, search operators, and so on.
a) **Boolean Query Formulation**

A text retrieval system should provide for query formulation by using the Boolean AND, OR, and NOT operators, and also provide nested Boolean searching. Boolean search facilities allow a user to combine search terms in a given search prescription, with certain conditions imposed. These conditions specify whether more than one search term should simultaneously be present in the desired records, whether any one of some chosen words should be present, or whether one or more terms should be present while another term should not be present in the desired records, and so on. Nested Boolean search facilities allow more complex conditions to be imposed along with the search terms. For Boolean OR queries, the system first finds hits for each search term and then matches the output sets to find out those records that contain both or either of the search terms. For Boolean AND queries, the system produces one set for each term is produced and then the results are matched to find out those record numbers that are common in each set. For Boolean NOT queries, a set(s) is produced with the search term(s) is produced with the search term(s)and another for the NOT term(s), and then the items which appear in the former set but not in the NOT set are retrieved.

b) **Proximity Searching**

In a text retrieval system there should be provision for adjacency/proximity searching, the purpose of which is to refine search statements by permitting the searcher to specify the context in which a term must occur. This search facility allows the user to specify whether two search terms should occur adjacent to each other, whether one or more words occur in between the search terms, whether the search terms should occur in the same paragraph irrespective of the intervening words, and so on. Proximity searching is a common feature of most information
retrieval system, including online and CD-ROM databases and Web search engines. Different types of proximity search facilities are available, ranging from distance between two search terms to the location of one or more search terms in a given sentence, a given paragraph, and so on. Conceptually, the simplest way to implement a query specification containing a proximity search parameter is to process the query as Boolean AND search and then to discard those items that do not match the specified criteria set by the proximity operator. The later part is performed by using the location information associated with the terms in the inverted file. This makes it possible to verify the nearness conditions without actually accessing the record information in the main record file. Nevertheless, the increase in the complexity of the file organization can be justified here because proximity searching is extremely important and is therefore available, with varying degrees of sophistication, in all the major text retrieval systems.

c) Range Searching

Range searching is most useful with numerical information. It is important in selecting records within certain data ranges. The following options are usually available for range searching. These operators are used do prescribe a precise condition in a given search statement.

- greater than(>)
- less than(<)
- equal to(=)
- not equal to (/=or <>)
- greater than equal to (>=)
- less than or equal to (<=)
d) Limiting Searches

The database in a text retrieval system comprise different fields containing different items of information. The user in his or her query formulation should be able to limit the search in one or more fields, and text retrieval software usually provides this facility. This is also known as field searching, whereby the user can specify that the search terms are to be looked for in one more fields. Limiting search or field-specific is a common information retrieval feature availing in most information retrieval systems, including search engines. If the system maintains a separate index file for each field, then it will only look for the term in the index file of the specified field and will build the search set(s) accordingly. On the other hand, if the retrieval system maintains only one index file, then it has to have the field information associated with each term in the index file. However, in the latter case, the retrieval time will be more because the system first has to produce a set where the search term(s) occur, and then from that set has to select only those items where the specified term occurs in the specified field.

e) Truncation

Truncation allows a search to be conducted for all the different forms of a word having the same common root. As an example, the truncated term COMPUT* will retrieve items on COMPUTER, COMPUTATION, COMPUTE, etc. A number of different options are truncation, e.g. right truncation, left truncation, and masking of letters in the middle of the word. Left truncation retrieves all words having the same characters at the right-hand part, e.g. '*hyl' will retrieve words like methyl, 'ethyl'. etc. Similarly, middle truncation retrieves all words having the same characters at the left and right-hand part. For example, a middle-truncated search 'colo*r' will
retrieve both ‘colour’ and ‘color’. A ‘wild card’ is used to allow any letter to appear in a specific location within a word.

f) String Searching
String searching is the ability to search on character strings within the body of the text in a record, which is usually available for those fields whose text has not been included in an inverted file, and is, therefore, not pre-indexed. In other words, string searching allows one to search those terms that have not been indexed. However, as the process of string searching matches the search term character by character with the stored records, the search process is tremendously slow for a relatively large database. Typically, text string searching may be used on a subset of the database that subset having been retrieved by searching on indexed fields. Some text retrieval systems provide a facility for string searching.

g) Search Refinements
The ability to specify a search strategy and to elicit a response in terms of numbers of posting or records retrieved on the basis of that strategy, is fundamental to most text retrieval systems. This facility supports the narrowing or broadening of search strategies. Once a search statement is formulated, there should be provisions for refining it in a step-by-step fashion so as to arrive at an optimum level of retrieval.

3.6 Conclusion
To sum up, information retrieval research has gradually become more confident and has established strong links with a number of related fields such as artificial intelligence, export systems, human-computer interaction, psychology and cognitive
science. A recent trend is towards the developments of intelligent knowledge-based systems in a global information infrastructure. Intelligent search agents are now available in information retrieval, particular in the Internet environment and perhaps a user-agent for every user. The other question is how to ensure that each agent communicates effectively with other. Researches are actively engaged in solving these problems and are also working on new standard for agents communications and knowledge transfer.

How would users access an ‘ideal’ computer-based information retrieval system? What strategies would they use in seeking information if they had access to a truly expert knowledge base that could respond effectively to any kind of questioning, phrased in any way? There is no such system exists. However, researchers are working on this context to give a best technique to retrieve information. Several measures, such as recall, precision, term overlap and efficiency, have been used to evaluate searching in bibliographic databases. There have been attempts to design and promote the notion of a standard interface, such as the Common Command Language, which would be made available to all hosts. User interfaces have been largely standardized by the use of common browser features. Many user interfaces now have attractive visualization features. In general, online search services and search engine interface have improved significantly; they have become more intuitive and demanding. Most advanced search interfaces allow users to formulate fairly complex search queries without having to learn the typical search syntax.

The retrieval services of the Web do not generally have the same level of search and display features found in traditional information retrieval systems used to access
bibliographic databases. Researchers have noted that the search characteristics of these systems catering to general audiences are different from traditional information retrieval systems. The future of research on information retrieval context lies in its increasing depth and integrations. We need a deeper understanding of the import contextual elements that encompass information retrieval processes and generalize able process models across situations. We need further integration of information retrieval and information-seeking models within the broader human information behavior context. Integrative models will work to help people develop their information seeking behaviors into effective searches, involving query formulation and reformulation, and to formulate relevance judgments effectively. A relatively new area of research with respect to context focuses on deriving contextual information from user queries themselves. Work in this area has been conducted primarily in the Web environment, but it has general application to interactive information retrieval more generally and we can expect more work in this area in future studies.