ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

Anatomy of Web Search Engines

Trilok Gupta[1], Archana Sharma[2]

Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, India[1]

Associate Professor, Department of Computer Science, Gurukul Institute of Engineering & Techchnology Institutional Area, Ranpur, Kota, Rajasthan, India[2]

Abstract: Today the main source of an information system is search engine. With the rapid development of Internet, network information sweeping the globe, produced a large amount of text, images, multimedia, and other forms of electronic information resources. This paper describes in detail the basic tasks a search engine performs. An overview of how the whole system of a search engine works. It provides and focuses on the structure of the retrieval system and to improve the performance, efficiency of the retrieval system, to speed up the inspection speed, and constantly adapt to the development of network information.

Keywords: World Wide Web, Web searching, Search engines, Web Crawlers, Indexing, metadata.

I. INTRODUCTION

Internet search engines are 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, but they all perform in four-step process[1]:

1. Crawling the Web, following links to find pages.
2. Indexing the pages to create an index from every word to every place it occurs.
3. Ranking the pages so the best ones show up first.
4. Displaying the results in a way that is easy for the user to understand.

Early search engines held an index of a few hundred thousand pages and documents, and received maybe one or two thousand inquiries each day. Today, a top search engine will index hundreds of millions of pages, and respond to tens of millions of queries per day.

In 1990, Tim Berners-Lee created the first Web browser (and Web editor) originally called the World Wide Web and later renamed to Nexus in order to avoid "confusion between the program and the abstract information space (which is now spelled World Wide Web with spaces)" [3]; it was written in Objective-C using the NeXT Computer. And at the time, this was the only way to browse the web. You can see a screenshot of the first browser in Figure 1.1 below.

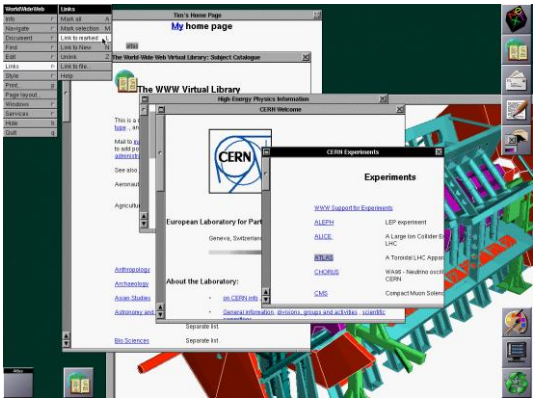

Figure 1.1: Screenshot of the first web browser called World Wide Web launched in 1990.

1993 marked an important turning point for the World Wide Web. The National Centre for Supercomputing Applications (NCSA) at the University of Illinois, led by Marc Andreessen, introduced the Mosaic browser. It quickly became popular due to its graphical support and its ability to "display images inline with text instead of displaying images in a separate window. Mosaic made it much easier for people to navigate hyperlinked pages and it made the Web "easy to use and more accessible to the average person. Andreesen's browser sparked the internet boom of the 1990s".

A year later, Andreesen "started his own company, named Netscape, and released the Mosaic-influenced Netscape Navigator in 1994, which quickly became the world's most popular browser, accounting for 90% of all web use at its peak". Then in 1995, Microsoft got involved in the web browser business and released Internet Explorer which was "heavily influenced by Mosaic, initiating the industry's first browser war. Bundled with Windows, Internet Explorer gained dominance in the web browser market". Till date so many search engines has been developed and some of them are active whiles others due to various reasons are now not active. Here from the first date of development the list of active web browsers can be seen in Table 1.2 [12] –

List of Active Search Engines from 1993

Year	Engine	Current Status
1993	W3Catalog	Active
1994	WebCrawler	Active, Aggregator
	Go.com	Active, Yahoo Search
	Lycos	Active
1995	Daum	Active
	Excite	Active
	SAPO	Active
	Yahoo! 2008	Active, Launched as a directory
1996	Dogpile	Active, Aggregator
	HotBot	Active (lycos.com)

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

	Ask Jeeves	Active (rebranded ask.com)
1997	Yandex	Active
1998	Google	Active
	MSN Search	Active as Bing
	empas	Inactive (merged with NATE)
1999	GenieKnows	Active, re branded Yellowee.com
	Naver	Active
	Teoma	Active
2000	Baidu	Active
	Exalead	Active
	Gigablast	Active
2003	Info.com	Active
2004	Yahoo! Search	Active, Launched own web search (see Yahoo! Directory, 1995)
	Sogou	Active
2005	AOL Search	Active
	Ask.com	Active
	GoodSearch	Active
2006	Quaero	Active
	Ask.com	Active
	Live Search	Active as Bing, Launched as re branded MSN Search
	ChaCha	Active
	Guruji.com	Active as BeeMP3.com
	Quaero	Active
2007	Blackle.com	Active, Google Search
2008	DuckDuckGo	Active
2009	Bing	Active, Launched as rebranded Live Search
	Scout (Goby)	Active
	NATE	Active
2010	Blekko	Active
	Yandex	Active, Launched global (English) search
2011	YaCy	Active, P2P web search engine
2012	Cloud Kite	Active, formerly Open Drive cloud search

Table 1.2 – Active Search Engines

WWW user survey indicate that about 86% of people now find a useful Web site through search engines, and 85% find them through hyperlinks in other Web pages; people now use search engines as much as surfing the Web to find information.

The paper is organized as follows:

Section 2 Architecture of Web Search Engines

Section 3 Tools for Web Based Retrieval

Section 4 Literature Review

Section 5 concludes the paper while references are shown in section 6.

II. ARCHITECTURE OF WEB SEARCH ENGINES

Search engines are programs that search web pages or documents for a particular keywords or where the keywords were found. A Search engine is really a collection of programs; however, the term is often used to specifically describe systems like Google, Bing and Yahoo! Search that enable users to search for documents on the World Wide Web. The components and tasks of web search engines, Crawling or Spidering is an automated process to gather the data with web spiders. They can be pictured as little spiders and are also known as crawlers, robots, software agents, web agents, wanderers, walkers, or know bots [Clay & Esparza, 2009]. They are named after special software robots, this type of search service is called "spider-based" or "crawler-based" search engine. Spiders process the web page and give us information. The web pages are found by them by URL which is given by a web page holder to notify their web page, or through hypertext links embedded in most web pages [Sherman & Price, 2001]. In the latter case, spiders start by crawling a few web pages and follow the links on those pages. After fetching the pages they point to, they follow the links that are on the last pages. The same process will be continued until they have indexed a certain part of the web that includes pages they store across many machines, what leads to the next task. Indexing is the second part of search engines. It is the process of "taking the raw data and categorizing it, removing duplicate information, and generally organizing it all into an accessible structure" [Clay & Esparza, 2009].The stored full-text indexes of the crawled web pages are organized in a database, typically in an inverted index data structure [Sherman & Price, 2001]. It is ultimate for keyword based queries, so that documents that include the typed keywords can be quickly retrieved. Webmasters have taken many advantages of the web, especially for business commitments. A lot of power will be put into search engine optimization (SEO) or maximizing search engine visibility, online marketing strategies [Clay & Esparza, 2009].

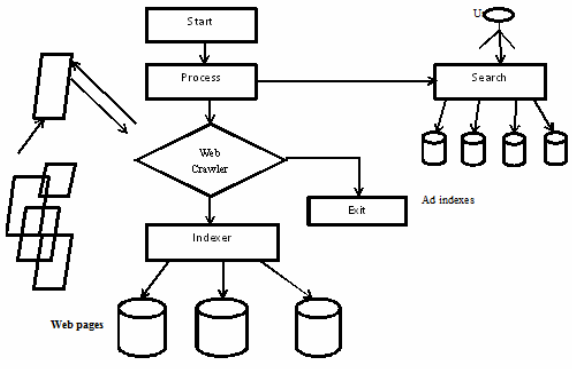

Figure 2.1: Architecture of Web Search Engines

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

III. Tools for Web Based Retrieval

3.1 Web Crawlers/Spider/Robots.

A web crawler is a program that, given one or more seed URLs, downloads the web pages associated with these URLs, extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks in a methodical way [14]. Web crawlers is also called ant, bot, worm or Web spider. A Web crawler usually starts with a list of URLs to visit (called the seeds). As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit (crawl frontier). URLs from the frontier are then recursively visited according to a set of policies.

Here is a Figure 3.1.1 that shows the architecture of a Web Crawler [14]:

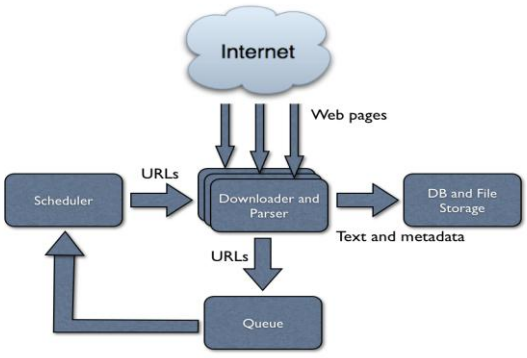

Figure 3.1.1: Architecture of Web Crawler

Web crawlers are an important component of web search engines, where they are used to collect the corpus of web pages indexed by the search engine. Moreover, they are used in many other applications that process large numbers of web pages, such as web data mining, comparison shopping engines, and so on. Despite their conceptual simplicity, implementing high-performance web crawlers poses major engineering challenges due to the scale of the web. The process of scanning the WWW is called Web crawling or Spidering. Scientists have recently been investigating the use of intelligent agents for performing specific tasks, such as indexing on the Web. There is some ambiguity concerning proper terminology to describe these agents. They are most commonly referred to as crawlers, but are also known as ants, automatic indexers, bots, spiders, Web robots and worms. It appears that some of the terms were proposed by the inventors of a specific tool, and their subsequent use spread to more general applications of the same genre.

To find information on the hundreds of millions of Web pages that exist, a search engine employs special software robots, called spiders, to build lists of the words found on Web sites. When a spider is building its lists, the process is called Web crawling. The spider will begin with a popular site, indexing the words on its pages and following every link found within the site. In this way, the spidering system quickly begins to travel, spreading out across the most widely used portions of the Web. These different approaches usually attempt to make the spider operate faster, allow users to search more efficiently, or both. For example, some spiders will keep track of the words in the title, sub-headings and links, along with the 100 most frequently used words on the page and each word in the first 20 lines of text. Lycos is said to use this approach to spidering the Web.

The amount of data that a search engine can store is limited by the amount of data it can retrieve for search results. Google can index and store about 3 billion web documents. This capacity is far more than any other search engine during this time [6].

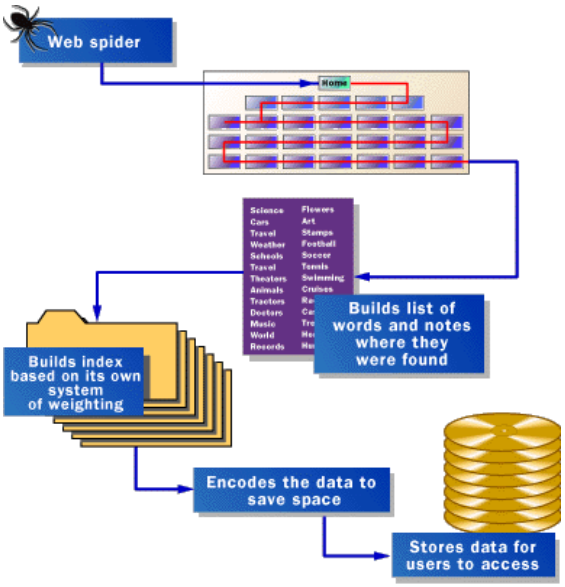

Figure 3.1.2: "Spiders" take a Web page's content and create key search words that enable online users to find pages they're looking for (Franklin, 2002).

Many search engines rely on automatically generated indices, either by them-selves or in combination with other technologies, e.g., AltaVista; Excite; Harvest ,harvest.transarc.com.; HotBot; Infoseek; Lycos; Webcrawler,webcrawler.com/.; and World Wide Web Worm. Although most of Yahoo!'s entries are indexed by humans or acquired through submissions, it uses a robot to a limited extent to look for new announcements. Examples of highly specialized crawlers include Argos,argos. Evansville.edu. Crawlers that index documents in limited environments include Look Smart, looksmart.com/. For a 300,000 site database of rated and reviewed sites;

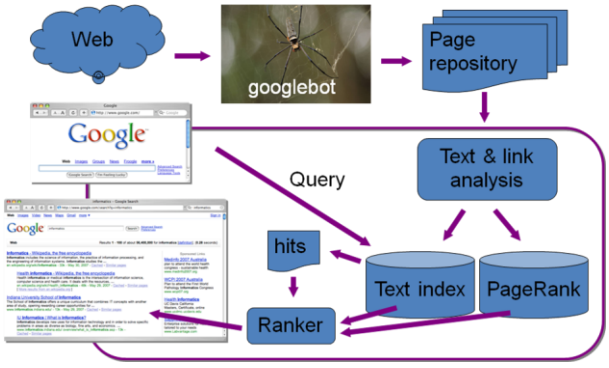

Figure 3.1.3: A crawler within a search engine

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

3.2 Indexing

The American Heritage Dictionary (1976) defines index as follows:
(in z dex) 1. Anything that serves to guide, point out or otherwise facilitate reference, as: a. An alphabetized listing of names, places, and subjects included in a printed work that gives for each item the page on which it may be found. b. A series of notches cut into the edge of a book for easy access to chapters or other divisions. c. Any table, file, or catalogue.

Although the term is used in the same spirit in the context of retrieval and ranking, it has a specific meaning. Some definitions proposed by experts are "The most important of the tools for information retrieval is the index—a collection of terms with pointers to places where information about documents can be found" [Manber 1999]; "indexing is building a data structure that will allow quick searching of the text" [Baeza-Yates 1999]; or "the act of assigning index terms to documents, which are the objects to be retrieved" [Korfhage 1997]; "An index term is a (document) word whose semantics helps in remembering the document's main themes" [Baeza-Yates and Ribeiro-Neto 1999]. Four approaches to indexing documents on the Web are:

1. Human or manual indexing;
2. Automatic indexing;
3. Intelligent or agent-based indexing; and
4. Metadata, RDF, and annotation-based
5. Indexing

3.2.1 Classical Methods.

Manual indexing is currently used by several commercial, Web-based search engines, e.g., Galaxy, galaxy.einet.net. KidsClick!, sunsite.berkeley.edu/ kidsclick!/. ; LookSmart , looksmart.com.; Web Developer's Virtual Library ,stars.com.; World-Wide Web Virtual Library Series Subject Catalog,w3.org/ hypertext/datasources/bysubject/overview.html.; and Yahoo!. The practice is unlikely to continue to be as successful over the next few years, since, as the volume of information available over the Internet increases at an ever greater pace, manual indexing is likely to become obsolete over the long term. Another major drawback with manual indexing is the lack of consistency among different professional indexers; as few as 20% of the terms to be indexed may be handled in the same manner by different individuals [Korfhage 1997, p. 107], and there is noticeable inconsistency, even by a given individual.

Though not perfect, compared to most automatic indexers, human indexing is currently the most accurate because experts on popular subjects organize and compile the directories and indexes in a way which (they believe) facilitates the search process. Notable references on conventional indexing methods, including automatic indexers, are Part IV of Soergel [1985]; Jones and Willett [1977]; van Rijsbergen [1977]; and Wittenet al. [1994, Chap. 3]. Technological advances are expected to narrow the gap in indexing quality between human and machine-generated indexes. In the future, human indexing will only be applied to relatively small and static (or near static) or highly specialized data bases, e.g., internal corporate Web pages.

3.2.2 Metadata, RDF, and Annotations.

"What is metadata? The Macquarie dictionary de- fines the prefix 'meta-' as meaning 'among,' 'together with,' 'after' or 'behind.' That suggests the idea of a 'fellow traveller ': that metadata is not fully fledged data, but it is a kind of fellow- traveller with data, supporting it from the side- lines. My definition is that 'an element of meta- data describes an information resource or helps provide access to an information resource.'" [Cathro 1997]

In the context of Web pages on the Internet, the term "metadata" usually refers to an invisible file attached to a Web page that facilitates collection of information by automatic indexers; the file is invisible in the sense that it has no effect on the visual appearance of the page when viewed using a standard Web browser.

One of the major drawbacks of the simplest type of metadata for labelling. HTML documents, called metatags, they can only be used to describe contents of the document to which they are attached, so that managing collections of documents (e.g., directories or those on similar topics) may be tedious when updates to the entire collection are made. Since a single command cannot be used to update the entire collection at once, documents must be updated one-by-one. Another problem is when documents from two or more different collections are merged to form a new collection. When two or more collections are merged, inconsistent use of metatags may lead to confusion, since a Meta tag might be used in different collections with entirely different meanings.

Metadata places the responsibility of aiding indexers on the Web page author, which is reasonable if the author is a responsible person wishing to advertise the presence of a page to increase legitimate traffic to a site. Unfortunately, not all Web page authors are fair players. Many unfair players maintain sites that can increase advertising revenue if the number of visitors is very high or charging a fee per visit for access to pornographic, violent, and culturally offensive materials. These sites can attract a large volume of visitors by attaching metadata with many popular keywords. Development of reliable filtering services for parents concerned about their children's surfing venues is a serious and challenging problem. Spamming, i.e., excessive, repeated use of key words or "hidden" text purposely inserted into a Web page to promote retrieval by search engines, is related to, but separate from, the unethical or deceptive use of metadata. Spamming is a new phenomenon that appeared with the introduction of search engines, automatic

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

indexers, and filters on the Web [Flynn 1996; Libera- tore 1997].

3.3 Clustering

Grouping similar documents together to expedite information retrieval is known as clustering [Anick and Vaithyanathan 1997; Rasmussen 1992; Sneath and Sokal 1973; Willett 1988]. During the information retrieval and ranking process, two classes of similarity measures must be considered: the similarity of a document and a query and the similarity of two documents in a database. The similarity of two documents is important for identifying groups of documents in a database that can be retrieved and processed together for a given type of user input query. Several important points should be considered in the development and implementation of algorithms for clustering documents in very large databases. These include identifying relevant attributes of documents and determining appropriate weights for each attribute; selecting an appropriate clustering method and similarity measure; estimating limitations on computational and memory resources; evaluating the reliability and speed of the retrieved results; facilitating changes or updates in the database, taking into account the rate and extent of the changes; and selecting an appropriate search algorithm for retrieval and ranking. This final point is of particularly great concern for Web-based searches.

There are two main categories of clustering: hierarchical and non-hierarchical. Hierarchical methods show greater promise for enhancing Internet search and retrieval systems. Although details of clustering algorithms used by major search engines are not publicly available, some general approaches are known. For instance, Digital Equipment Corporation's Web search engine Alta- Vista is based on clustering. Anick and Vaithyanathan [1997] explore how to combine results from latent semantic indexing and analysis of phrases for context-based information retrieval on the Web.

3.4 User Interfaces

Currently, most Web search engines are text-based. They display results from input queries as long lists of pointers, sometimes with and sometimes without summaries of retrieved pages. Future commercial systems are likely to take advantage of small, powerful computers, and will probably have a variety of mechanisms for querying non-textual data (e.g., hand-drawn sketches, textures and colors, and speech) and better user interfaces to enable users to visually manipulate retrieved information. Hearst [1999] surveys visualization interfaces for information retrieval systems, with particular emphasis on Web-based systems.

IV. LITERATURE REVIEW

Our main goal is to improve the quality of web search engines. In 1994, some people believed that a complete search index would make it possible to find anything easily. According to Best of the Web 1994 -- Navigators, "The best navigation service should make it easy to find almost anything on the Web (once all the data is entered)." However, the Web of 1997 is quite different. Anyone who has used a search engine recently can readily testify that the completeness of the index is not the only factor in the quality of search results. "Junk results" often wash out any results that a user is interested in. In fact, as of November 1997, only one of the top four commercial search engines finds itself (returns its own search page in response to its name in the top ten results). One of the main causes of this problem is that the number of documents in the indices has been increasing by many orders of magnitude, but the user's ability to look at documents has not. People are still only willing to look at the first few tens of results. Because of this, as the collection size grows, we need tools that have very high precision (number of relevant documents returned, say in the top tens of results). Indeed, we want our notion of "relevant" to only include the very best documents since there may be tens of thousands of slightly relevant documents. This very high precision is important even at the expense of recall (the total number of relevant documents the system is able to return). There is quite a bit of recent optimism that the use of more hyper textual information can help improve search and other applications. In particular, link structure and link text provide a lot of information for making relevance judgments and quality filtering. Google makes use of both link structure and anchor text.

V. CONCLUSION

This paper presents an overview of search engines and its techniques. In order to improve retrieval accuracy of Web search, we studied its architecture and tools for web based retrieval. Our proposed approaches described in this paper contribute for indexing a target Web page more accurately and allowing each user to understand, perform more fine-grained search that satisfy his/her information need.

REFERENCES

[1] http://norvig.com/InternetSearching.pdf
[2] N. Fuhr. Probabilistic Models in Information Retrieval. The Computer Journal, 35(3): pages 243–255, 1992.
[3] http://www.w3.org/People/Berners-Lee/ WorldWideWeb.html
[4] http://en.wikipedia.org/wiki/Web_browser
[5] http://www.webopedia.com.
[6] Franklin, Curt. How Internet Search Engines Work, 2002. www.howstuffworks.com
[7] http://docs.google.com
[8] S. E. Robertson and K. S. Jones. Relevance Weighting of Search Terms. Journal of the American Society for Information Sciences, 27(3): pages 129–146, 1976.
[9] J. Rocchio. Relevance Feedback in Information Retrieval. In G. Salton, editor, The Smart Retrieval System: Experiments in Automatic Document Processing, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, 1971.
[10] G. Salton. The Smart Retrieval System: Experiments in Automatic Document Processing. Prentice - Hall, Englewood Cliffs, NJ, 1971.
[11] C. J. van Rijsbergen. Information Retrieval. Butterworths.
[12] http://en.wikipedia.org/wiki/Search_engine
[13] http://www.pewinternet.org/Reports/2012/ Search-Engine-Use-2012/Summary-of- findings.aspx

[14] http://research.microsoft.com/pubs/102936/ eds-webcrawlerarchitecture.pdf.
[15] G. Jeh and J. Widom. Scaling Personalized Web Search. In Proc. of the 12th International World Wide Web Conference (WWW 2003), pages 271–279, 2003.
[16] T. Hofmann. Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis. In Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03), pages 259–266, 2003.

BIOGRAPHIES

Trilok Gupta was born in Kota (Rajasthan). He has done Diploma in Computer Science and received his Master Degree in Computer Science from JRNRV University, Udaipur, Rajasthan-India. He is pursuing Ph.D. Computer Science from Faculty of Computer Application, Pacific University, Udaipur-PAHER, Rajasthan - India. His area of interest includes Data Handling, Data Mining, Web Applications, Search Engines Optimization and Information Exploring. He is working in the field of education for last 14 years. He has published several research papers in National and International Journals. He is currently exploring the anatomy of Search Engines and Web Mining.

Archana Sharma was born in Ajmer (Rajasthan). She is Ph.D in Computer Science and Engineering with specialization in Simulation and Modeling. She completed her M.Tech in Computer Science from Banasthali Vidyapith, India. Her field of study is Simulation and modeling, data mining, database, Artificial Intelligence.
She is working in the field of education for last 15 years. She has taught many subjects at undergraduate and postgraduate level. She has published several research papers in national and International Journals. She is currently working in the field of Cloud Computing, Artificial Intelligence and Educational Data Mining. She is Senior Member of International Association of Computer Science and Information Technology (IACSIT). She is also the Board member in Seventh Sense Research Group Journals. P Professor Sharma is member of Indian Society of Theoretical and Applied Mechanics. She worked as Editor in Journal of Management and IT 'OORJA'.