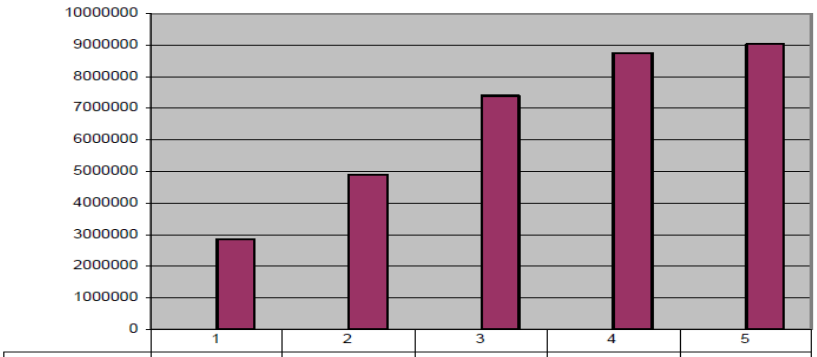
LECTURE NINE

WEB SEARCH BASICS

9.1 Internet

The Internet has become the largest source of information. Today, millions of Websites exist and this number continuous to grow.



	1	2	3	4	5
Year	1998	1999	2000	2001	2002
No. of Websites	2,851,000	4,882,000	7,399,000	8,745,000	9,040,000

(Source : OCLC)

Internet research is the practice of using Internet information, especially free information on the World Wide Web, in research. It is:

- focused and purposeful (so not recreational browsing),
- uses internet information or internet-based resources (like internet discussion forums),
- tends towards the immediate (drawing answers from information you can access without delay)
- and tends to access information without a purchase price.

The most popular search tools for finding information on the internet is Web search engines.

The end user uses the Internet heavily in accessing information in their day to day needs since the Internet is the biggest repository of knowledge in the history of mankind. Thus, the Internet has become the world's widely accessed information source and Websites are added, updated, and obsolete daily.

Generally, people use search engines for one of **three things**: *research*, *shopping*, or *entertainment*. Most people who are using a search engine are doing it for research purposes. They are generally looking for answers or at least to data with which to make a decision. A smaller percentage of people, but still very many, use a search engine in order to shop. Research and shopping aren't the only reasons to visit a search engine. The Internet is a vast, addictive, reliable resource for entertainment.

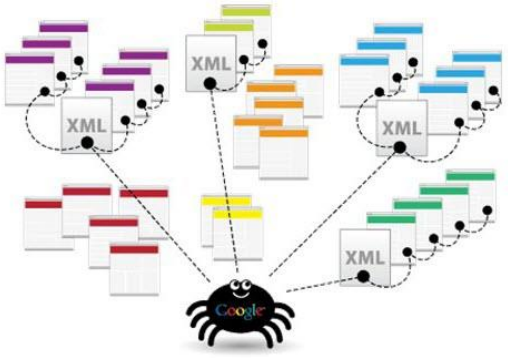
A web search engine is a type of website that helps computer user find information on the Internet. It does this by looking through other *web pages* for the text the user wants to find. The software that does this is known as a *search engine*.

9.2 What is a Search Engine?

- Search engine is a tool, which helps in retrieving information from the Internet
- It indexes the web and accordingly builds its databases
- Each search engine has its own set rules to index websites
- When the query or keyword is entered in the search box it checks its ‘index’ with the query
- Relevant matches are retrieved and returned as 'hits' or 'search results ‘

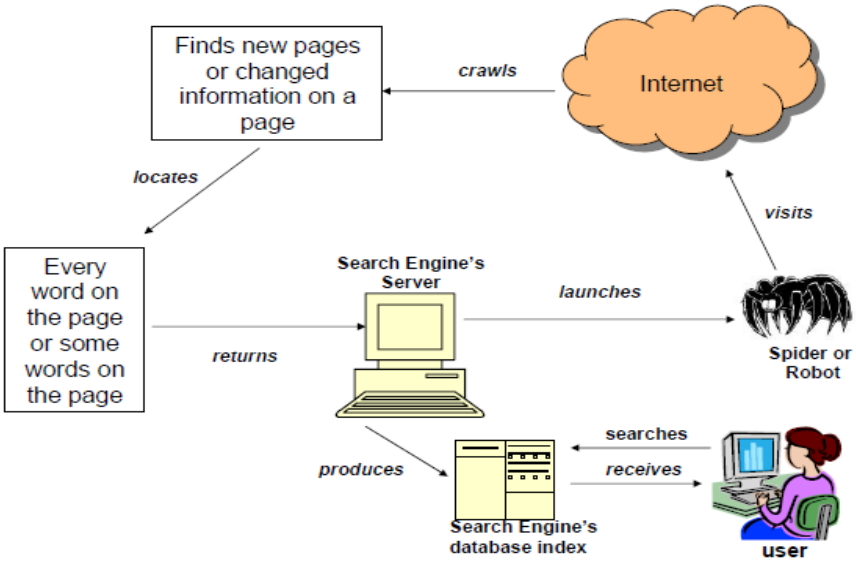
A Search engine has three parts:

• **Spider**: Deploys a robot program called a spider or robot designed to track down web pages. It follows the links these pages contain, and add information to search engines’ database. It is also called "*web crawler*". Example: Googlebot (Google’s robot program)



• **Index**: Database containing a copy of each Web page gathered by the spider.

• **Search engine software**: Technology that enables users to query the index and that returns results in a schematic order.



9.3 Types of search engines:

In broad sense, search engines can be divided into two categories.

1. **Individual search engines**:

An individual search engine uses a spider to collect its information regarding websites for own searchable index. There are two types of individual search engines.

i . *General search engines*

Examples: Google, AltaVista, HotBot, Lycos

ii. *Subject specific search engines*

Examples: MetaPhys, Chritech, ReligionExplorer.

2. **Meta search engines**

A Meta search engine searches multiple individual engine simultaneously. It does not have its own index, but uses the indexes collected by the spiders of other search engines.

Example: metacrawler, Ixquick, mamma

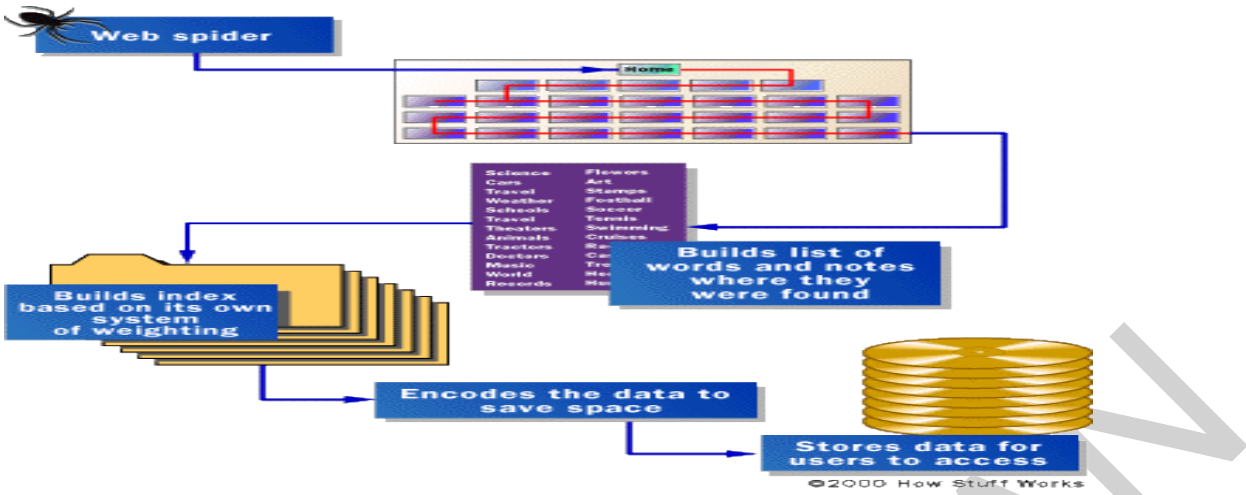
9.4 How Internet Search Engines Work?

Before a search engine can tell you where a file or document is, it must be found. 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*. In order to build and maintain a useful list of words, a search engine's spiders have to look at a lot of pages.

Crawler ("spider" or "bot"): A crawler is a program that visits Web sites and reads their pages and other information in order to create entries for a search engine index.

Crawlers apparently gained the name because they crawl through a site a page at a time, following the links to other pages on the site until all pages have been read.

It does not (or cannot) go through *firewalls*. And it uses a special *algorithm* for waiting between successive server requests so that it doesn't affect response time for other users



"Spiders" take a Web page's content and create key search words that enable online users to find pages they're looking for.

How does any spider start its travels over the Web? The usual starting points are lists of heavily used servers and very popular pages. 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.

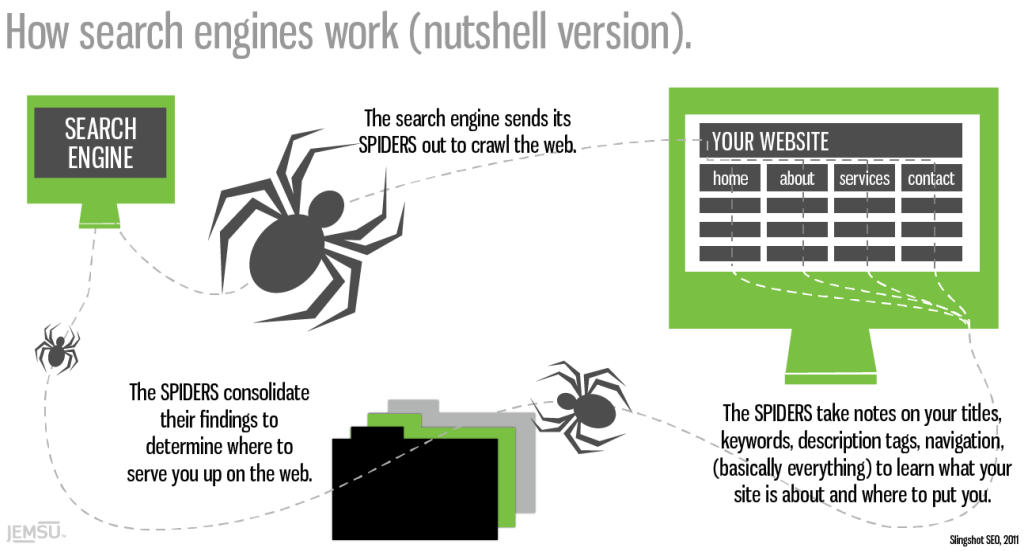
Keeping everything running quickly meant building a system to feed necessary information to the spiders.

The early Google system had a server dedicated to providing URLs to the spiders. Rather than depending on an Internet service provider for the domain name server (DNS) that translates a server's name into an address, Google had its own DNS, in order to keep delays to a minimum.

When the Google spider looked at an HTML page, it took note of two things:

- The words within the page
- Where the words were found

Words occurring in the title, subtitles, **meta tags** and other positions of relative importance were noted for special consideration during a subsequent user search. The Google spider was built to index every significant word on a page, leaving out the articles "a," "an" and "the." Other spiders take different approaches.



9.5 Primary Goals of Search Engines

• **Effectiveness** (quality): to retrieve the most relevant set of documents for a query

– Process text and store text statistics to improve relevance

• **Efficiency** (speed): process queries from users as fast as possible

– Use specialized data structures

Advantages of using search engines

- Search engines are best at finding unique keywords, phrases, quotes, and information buried in the full-text of web pages since they normally index WWW documents word by word.
- Search engines allow the user to enter keywords, and then they are searched against its database. Users can use advanced search techniques such as phrase searching, truncation/wildcard searching, as well as for Boolean operators (AND, OR, NOT combinations).

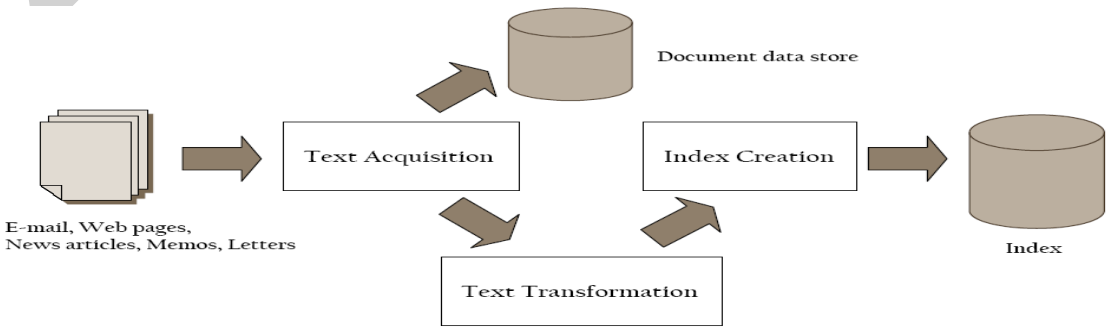
Disadvantages of Search Engines

- Creates information overload
- Semantic search is not possible
- Privacy and security is of concern
- Makes everyone to dependent

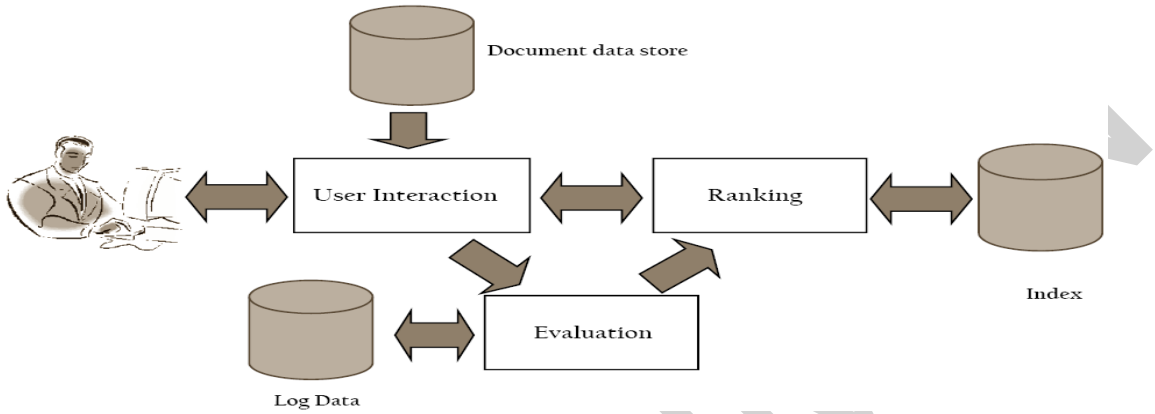
9.6 Search engine major functions

Search engine components support *two major functions*:

• **The index process**: building data structures that enable searching.



• **The query process**: using those data structures to produce a ranked list of documents for a user's query.



User Interaction

• Providing the interface between users and the search engine

• **Tasks**

– Accepting a user query and transforming it into index terms

– Taking the ranked list of documents from the search engine and organizing it into the results shown to the user

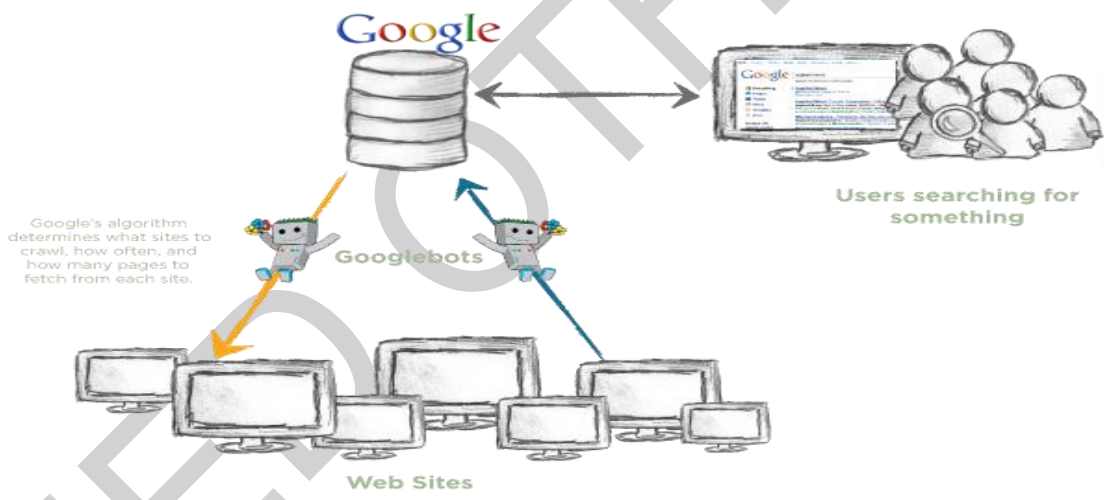
– Refining a user query to better represent the information need

Search Engine Optimization (SEO):

Is simply the optimization of your site, so that search engines can find the content YOU would like them to find.

There are three fundamental strategies in regards to optimizing your site so that users will find you.

1. Content is king. The better quality your content is, the better off your site will be. Write for your users, not the search engines.
2. Build your site with good page semantics. This means clean, well structured code, so that search engines such as Google will find what they need quickly, and the way you want them to.
3. Identify your keywords. How are / will your users be finding you online? Think about what users will be searching on at a search site such as Google. Those letters that users type into Google's search box are called "keywords" or "keyword phrases" - it's pure gold if you can accurately identify what key words users will find you with.



9.7 Types of search results:

The Google, Yahoo! and Bing Search engines combine advertising and search results on their search results pages. In each case, the ads are designed to look similar to the search results, though differ in formatting enough for readers to make distinctions between organic results and ads, such as their background color and/or placement on the page.

Further, the appearance of the ads on all major search engines is so similar to the genuine search results that a large majority of search engine users cannot effectively distinguish between them.

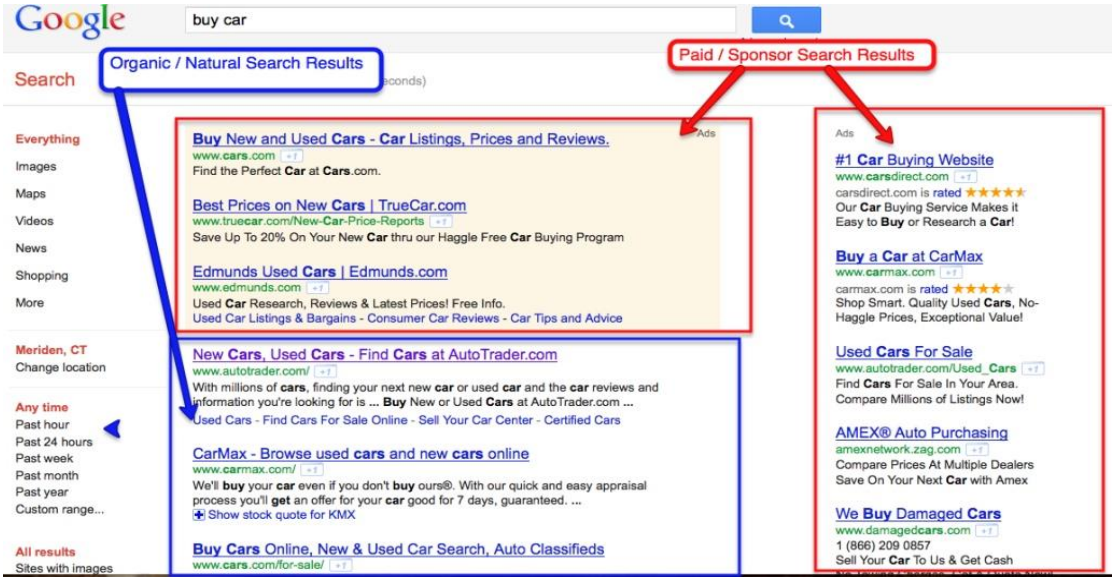
Organic search results are listings on search engine results pages that appear because of their relevance to the search terms, as opposed to their being advertisements. In contrast, non-organic search results may include pay per click advertising.

What is the Difference between Organic and Paid Search Engine Results?

When a search engine returns its search results, it gives you two types: organic and paid. **Organic search** results are the Web page listings that most closely match the user's search query based on relevance. Also called "*natural*" search results, ranking high in the organic results is what SEO is all about.

Paid results are basically advertisements — the Web site owners have paid to have their Web pages display for certain keywords, so these listings show up when someone runs a search query containing those keywords.

On a search results page, you can tell paid results from organic ones because search engines set apart the paid listings, putting them above or to the right of the organic results, or giving them a shaded background, border lines, or other visual clues. The following figure shows the difference between paid listings and organic results.



Results page from Google with organic and paid results

The typical Web user might not realize they're looking at apples and oranges when they get their search results. Knowing the difference enables a searcher to make a better-informed decision about the relevancy of a result. Additionally, because the paid results are advertising, they may actually be more useful to a shopping searcher than a researcher (as search engines favor research results).

9.8 Search engine indexing

Meta search engines reuse *the indices* of other services and do not store a local index, whereas cache-based search engines permanently store the index along with the corpus.

The *purpose of storing an index* is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power.

Index design factors:

Major factors in designing a search engine's architecture include:

- **Merge factors**

  How data enters the index, or how words or subject features are added to the index during text corpus traversal, and whether multiple indexers can work asynchronously. The indexer must first check whether it is updating old content or adding new content.

- **Storage techniques**

  How to store the index data, that is, whether information should be data compressed or filtered.

- **Index size**

  How much computer storage is required to support the index.

- **Lookup speed**

  How quickly a word can be found in the inverted index. The speed of finding an entry in a data structure, compared with how quickly it can be updated or removed, is a central focus of computer science.

- **Maintenance**

  How the index is maintained over time.

- **Fault tolerance**

  How important it is for the service to be reliable. Issues include dealing with index corruption, determining whether bad data can be treated in isolation, dealing with bad hardware, partitioning, and schemes such as hash-based or composite partitioning, as well as replication.

9.9 Challenges in parallelism

A major challenge in the design of search engines is the management of serial computing processes. For example, a new document is added to the corpus and the index must be updated, but the index simultaneously needs to continue responding to search queries. This is a collision between two competing tasks. The indexer is the producer of searchable information and users are the consumers that need to search. The challenge is magnified when working with distributed storage and distributed processing.

Compression

Generating or maintaining a large-scale search engine index represents a significant storage and processing challenge. Many search engines utilize a form of compression to reduce the size of the indices on disk

What are the challenges in natural language processing?

1-Word Boundary Ambiguity

Native English speakers may at first consider tokenization to be a straightforward task, but this is not the case with designing a multilingual indexer. In digital form, the texts of other languages such as Chinese, Japanese or Arabic represent a greater challenge, as words are not clearly delineated by whitespace. The goal during tokenization is to identify words for which users will search.

2-Language Ambiguity

To assist with properly ranking matching documents, many search engines collect additional information about each word, such as its language or lexical category (part of speech). These techniques are language-dependent, as the syntax varies among languages.

3--Language recognition

If the search engine supports multiple languages, a common initial step during tokenization is to identify each document's language; Language recognition is the process by which a computer program attempts to automatically identify, or categorize, the language of a document.

Other challenges in search engines:

1-Format analysis

If the search engine supports multiple document formats, documents must be prepared for tokenization. For example, HTML documents contain HTML tags, which specify formatting information such as new line starts, bold emphasis, and font size or style. Format analysis is the identification and handling of the formatting content embedded within documents which controls the way the document is rendered on a computer screen or interpreted by a software program. Common, well-documented file formats that many search engines support include:

- HTML ,ASCII text files (a text document without specific computer readable formatting), Adobe's Portable Document Format (PDF), PostScript (PS), LaTeX, Multimedia meta data formats like ID3, Microsoft Word, Microsoft Excel, Microsoft PowerPoint.

Some search engines support inspection of files that are stored in a compressed or encrypted file format. Commonly supported compressed file formats include:

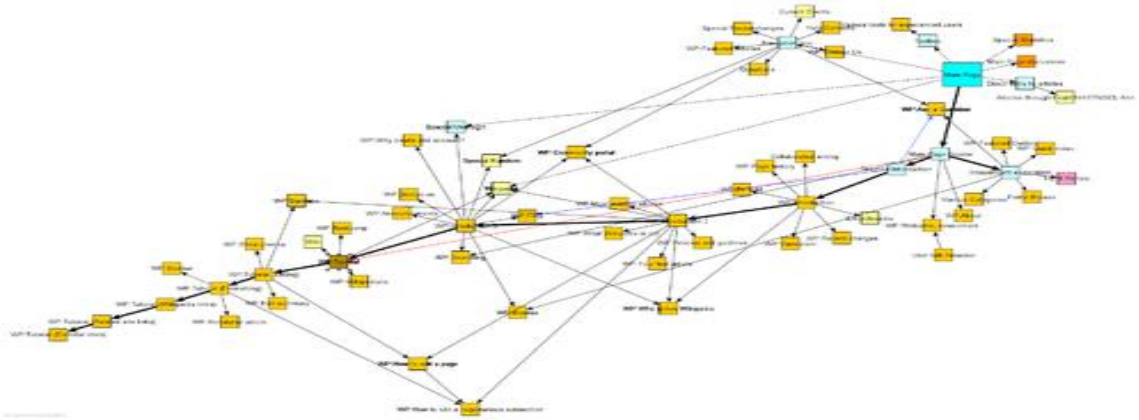
- ZIP - Zip archive file, RAR - Roshal ARchive file, CAB - Microsoft Windows Cabinet File, Gzip - File compressed with gzip, BZIP - File compressed using bzip2.

2-Section recognition

Some search engines incorporate section recognition, the identification of major parts of a document, prior to tokenization. Section analysis may require the search engine to implement the rendering logic of each document, essentially an abstract representation of the actual document, and then index the representation instead.

9.10 Site Map

A site map (or sitemap) is a list of pages of a web site accessible to crawlers or users. It can be either a document in any form used as a planning tool for Web design, or a Web page that lists the pages on a Web site, typically organized in hierarchical fashion.



There are two popular versions of a site map: An XML Sitemap, and HTML sitemaps.