UNIT 15  SEARCH ENGINES

Structure
15.0  Objectives
15.1  Introduction
15.2  Search Engines
15.3  Types of Search Tools
   15.3.1  Search Directory
   15.3.2  Search Engines
   15.3.3  Meta-search Engines
15.4  Features of Search Tools
   15.4.1  Keyword Search
   15.4.2  Boolean Search
   15.4.3  Proximity Search
   15.4.4  Truncation Search
   15.4.5  Case Sensitive Search
   15.4.6  Limiting Search
   15.4.7  Fields Search
   15.4.8  File Types Search
   15.4.9  Stop Words
   15.4.10  Ranking
   15.4.11  Family Filters
   15.4.12  Fuzzy Search
15.5  Architecture of Search Tools
   15.5.1  Web Crawler
   15.5.2  Metadata Storage
   15.5.3  Search Agent
   15.5.4  User Interface
15.6  Challenges
15.7  Summary
15.8  Answers to Self Check Exercises
15.9  Keywords
15.10  References and Further Reading

15.0  OBJECTIVES
After going through this Unit, you will be able to:

- describe different search tools available to locate information on web;
- discuss search engines as an effective information retrieval tool;
- know different search strategies to increase the performance of search tools; and
- explain search engine architecture.
**15.1 INTRODUCTION**

Internet has created revolutionary changes in this era of Information Technology. For many, it is one stop platform to find or locate any information they are interested in. Traditionally, librarians had the job to assist their users to locate the information they needed. But, now the scenario has changed a lot. Internet has offer a variety of search tools such as search engines, search directories to locate the information on web.

A search on web is a simple process and can be conducted by simply issuing a query to the search tool. The search tool in return will look for the information in its web based information databases and retrieves those, which are relevant to the query. Searching is an iterative process i.e. one needs to keep working on their query unless the exact information is located.

The very first tool used for searching on the Internet was Archie. The name stands for “archive” without the “v.” It was created in 1990 by Alan Emtage, a student at McGill University in Montreal. Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerised Archives) and Jughead (Jonsy’s Universal Gopher Hierarchy Excavation And Display) were two other popular search programs.

There are three basic types of search tools that most people use to find what they are looking for on the Web: Search Engines, Subject Directories and Meta Search Tools. Search Engines are more generic and much larger than Subject Directories. Meta Search Tools get their results from several search engines. The following sections will provide an elaboration on these search tools.

**15.2 SEARCH ENGINES**

Search engine is a tool for locating information from a collection. Search engines uses information about the information (such as metadata, catalogue) stored in the database to locate information. Sometimes they perform full text search within the document from first character to last character.

The search is done on pattern matching algorithm whether it is a database or full text.

**15.3 TYPES OF SEARCH TOOLS**

**15.3.1 Search Directory**

Search directories are classified collections of documents. They are good for searching with a context. These directories are good for browsing. In subject directories, documents are pre classified by a person. Librarians’ Internet Index; Google Directory; Yahoo!; dmoz are some of the examples of subject directories.

There are two basic types of directories:

- Academic and Professional Directories: These are often created and maintained by subject experts to support the needs of researchers. INFOMINE, from the University of California, is a good example of an academic directory.

- Commercial Directories: These cater to the needs of general public. Directories of Yahoo! and Google are examples of commercial directories.
15.3.2 Search Engines

World Wide Web is a network of several information databases. In recent years, an exponential growth in these databases has made it difficult to locate a particular piece of information. Internet offers a powerful tool known as search engine to manage, filter and retrieve the information for their users.

Search engines are automated tools for searching information from a collection using metadata stored in the database of search engine. In other words, it is an information retrieval system and assists in locating information on web.

Google and Yahoo! are most popular search engines.
15.3.3 Meta-Search Engines

Meta Search engines are online tools (search engines) which performs simultaneous search on more than one search engine at a time. These search engines aggregates the results into a single list and displays them according to their source. e.g. Dogpile is a metasearch engine and gets its results from Google, Yahoo, MSN Search, Ask, About, MIVA, LookSmart, and more.

Example: Dogpile, WebCrawler, Browsys

Fig. 15.3: Search Interface of “Dogpile”, a Metasearch Engine

Self-Check Exercise

Note:

i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of this Unit.

1) Write a short note on Search Tools?

2) Name any four search engines and two metasearch engines?

....................................................................................................................
....................................................................................................................
....................................................................................................................
....................................................................................................................

15.4 FEATURES OF SEARCH TOOLS

The searching tools follows pattern matching algorithm. There are many types of searches can be done using search tools,
15.4.1 **Keyword Search**

When searching is done using a keyword it is known as Keyword searching. Keyword may occur at any place in the document or in the metadata field. This kind of search has higher recall value.

15.4.2 **Boolean Search**

Logical AND, OR and NOT are known as Boolean operators. When Boolean operators are used for searching it is known as Boolean search. The operators are used for combining more than one word with certain conditions. These kind of searching also known as Combinatorial search.

➢ **AND**

This operator will retrieve all the documents which contains all the keywords occurring at both ends of the AND operator.

*Syntax:* `<Search Term A> AND <Search Term B>`

*Example:* Library AND Information

*Output:*

1. The above query will retrieve only those documents which contains both the terms Library and Documentation
2. The precision in search is more. The number of documents retrieved will be less hence less is the recall value.

➢ **OR**

This operator will retrieve all the documents which contains all the keywords occurring at both ends of the OR operator.

*Syntax:* `<Search Term A> OR <Search Term B>`

*Example:* Library OR Information

*Output:*

1. The above query will retrieve all documents which contains both the terms Library and Documentation
2. The recall in search is more. The number of documents retrieved will be more but the precision in retrieved documents will be less.

➢ **NOT or AND NOT**

These operators increase the precision of the search result. The query can be made more specific by using these operators. Using the capitalised AND NOT operator preceding a search term eliminates documents that contain that term.

*Syntax*

<Words to be searched> AND NOT <Words not to be searched>

*Example:*

If user is looking for information on Drivers and do not want documents that include information relating to the Screw Drivers the query could be “Driver” AND NOT Screw.
15.4.3 Proximity Search

This is another kind of Combinatorial search where the proximity of two words is checked. The term proximity means ‘nearness of words’. Proximity is given in terms of number of words by which two words should be separated. There are two kinds of proximities,

1) Near Proximity

2) Exact Proximity

1) Near Proximity

Near Proximity brings range of search results where the number of proximity is from adjacent to the mentioned number proximity. For example, for two keywords COLLEGE and LIBRARIANS, if the proximity of 3 is applied between them it will bring the results as follows,

COLLEGE LIBRARIANS
COLLEGE FOR LIBRARIANS
COLLEGE OF LIBRARIANS
COLLEGE WITHOUT LIBRARIANS
COLLEGE WITH THE LIBRARIANS
COLLEGE OF INDIAN LIBRARIANS
COLLEGE OF THE LIBRARIANS

The Near Proximity would bring the results where the search terms would be separated by no word to n-1 word (where n is number of proximity). In other words, near proximity brings all the proximities which are lesser than the mentioned number.

2) Exact Proximity

Exact proximity brings the results with exact number of proximity mentioned. It does not bring the results which have lesser number of proximity. For example, if the number of proximity is set to 3 between two keywords i.e. COLLEGE and LIBRARIANS. The retrieved result would be,

COLLEGE WITH THE LIBRARIANS
COLLEGE OF INDIAN LIBRARIANS
COLLEGE OF THE LIBRARIANS

15.4.4 Truncation Search

Truncation means concatenation of words. In other words, if the root string of the words is searched it brings all the derivatives derived out of the given root string. Truncation is of three types based on truncation techniques:

1) Left Truncation

When the root string is concatenated from the left side, it is known as left truncation. For example, if the left truncation is implemented for the root string ISM, it will bring all the words which ends with the string ISM, like
2) **Right Truncation**

When the root string is concatenated from the right side it is known as right truncation. For example, right truncation is used with the root string CLASS, it will bring all the words which starts with the root string CLASS, like

CLASS
CLASSIFICATION
CLASSIFICATIONIST
CLASSIFIER

15.4.5 **Case Sensitive Search**

One of the major features of search tools is their support to search words based on their case. In other words, search tools can differentiate between Upper and lower cases. For example, DUKE and duke will bring different search results based on the case. In an ordinary/plain search, search tool performs searching irrespective of their cases. However, if case sensitive search is invoked, search tool brings exact search string based on the case of search string.

15.4.6 **Limiting Search**

There are certain conditions based on which the searches can be narrowed down, for example, by Date, by Domain, by media type, by Document Directory Depth, by Page Depth and so on. This kind of condition reduces the number of search results and increases the relevancy of final output.

15.4.7 **Fields Search**

Field Search is a kind of limiting search to a particular field of the database. Searching can be done within a given context. For example, searching within Title or searching within Author or searching within both the fields. This kind of searching is known as Field search.

15.4.8 **File Types Search**

When the searching is restricted to a particular file type like, MS-Word, PDF, PPT etc. it is known as file type search.

15.4.9 **Stop Words**

While searching documents in a collection or over Web, some frequently occurring words like prepositions, conjunction etc. should be avoided. In order to avoid such redundant words from the search results, the tools contain a file called stop word file. This file lists all the words which are to be avoided from being indexed. This saves space of storage and reduces time of search.

15.4.10 **Ranking**

Search tools present the search results in some order. Normally, when system is small
the presented results are arranged in alphabetical order. But when results run in several pages, it becomes important to present the most relevant document on the top followed by less relevant one. Thus, it is important to rank the retrieved documents based on their relevancy to their users. Search engines have an automated mechanism to rank the retrieved results according to the relevancy of each retrieved search result.

For example, PageRank of Google is an algorithm for measuring weightage of results based on link analysis.

15.4.11 Family Filters

Family filters are used to reduce, if not remove, the objectionable matter to appear on search results. Search engines do provide functionality for setting family filters as safe search. Family filters are used by Google (as Safe Search), AltaVista, Yahoo and so on. Apart from the search engines there are tools (e.g. Naomi, which is a freeware) which can be loaded on computers to stop display of obscene matters.

15.4.12 Fuzzy Search

Fussy search is one of the major features of today’s information retrieval system. It brings out results based on approximations. In other words, these are error correction algorithms. For example, if a keyword is miss-spelt search algorithm used in searching attempts to render the search result according to correct spelling. Such algorithms are known as Soundex and Metaphone algorithms. Levenshtein distance algorithm is one such kind of algorithm used by Lucene search engine, an open source search tool.

Self-Check Exercise

Note:  i) Write your answers in the space given below.
      ii) Check your answers with the answers given at the end of this Unit.

3) Write short note on Boolean Search?
4) What is a Proximity Search?
5) What is “PageRank”?

   ....................................................................................................................
   ....................................................................................................................
   ....................................................................................................................
   ....................................................................................................................

15.5 ARCHITECTURE OF SEARCH TOOLS

WWW is a huge source of information and search engines are tools or agents for locating information. Everything and anything can be located over Internet using search engines. Search Engines are tools which provide a kind of interface for users to search the web. A Search Engine basically has three components:

- Web Crawler
- Metadata storage
- Search Agent
15.5.1 Web Crawler

Web crawler is also known as robot or spider. It is a program which goes to each and every site over Internet and indexes the content of the webpage. The content includes metadata information and the text from the webpage. Text of the page can be indexed as a whole or only few lines or bytes of data are stored. This index is stored within search engines database with corresponding URL (Uniform Resource Locator).

Hence, Web crawler is a program used by search engine in order to extract data from the web pages so that pages can be searched using the search engine’s interface. Following are the names of web crawlers used by popular search engines:

<table>
<thead>
<tr>
<th>Search engine</th>
<th>Robot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Googlebot</td>
</tr>
<tr>
<td>Yahoo</td>
<td>Slurp</td>
</tr>
<tr>
<td>MSN</td>
<td>MSNbot</td>
</tr>
</tbody>
</table>

15.5.2 Metadata Storage

Metadata is data about data. Web Crawlers extract metadata like title, author, filename, file type, file size, links and so on from search engines. The metadata information is extracted from meta-tag of webpage, file name, file extension, etc. The collected metadata is stored in the repository of search engine, which is a database, in the form of index. Normally, many of the system follow a kind of keyword indexing. But the Keyword indexing is good for recall not for precision. In such cases the context for search is lost. However, the use of metadata is good for preserving the context of search term. The metadata is stored in the form of metadata index inside the search engine database.

![Fig. 15.4: Architecture of Search Engine](image)

The most commonly used metadata schema is Dublin Core Metadata Initiative (DCMI) over Internet. The standard is developed and maintained by DCMI and DCMI Task Groups. There are 15 elements given in Dublin Core. Apart from these 15 elements there are other metadata set vocabularies which should be used with 15 elements.
15.5.3 Search Agent

Search Agent is a set of search programs which receives query from the user interface. The received key word is passed through several algorithms. Some commonly used algorithms are as follows:

- Boolean operator
- Data clustering algorithms
- Error correction algorithms.

Once the query or search term is processed, the agent performs search within the repository of the search engine and retrieves the search result and send to user interface.

### Boolean Operators

The Boolean operators are AND, OR and NOT. These operators are used to generate combinatorial search. AND and NOT operators increase precision where as OR increases recall of search results. The shaded area represents retrieved records in the following example (Fig. 15.6). Almost all the search engines provide facility of using Boolean operators. These operators can also be used to combine keywords present in different fields.

<table>
<thead>
<tr>
<th>Water</th>
<th>Plant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Water AND Plant

<table>
<thead>
<tr>
<th>Water</th>
<th>Plant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Water OR Plant

<table>
<thead>
<tr>
<th>Water</th>
<th>Plant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Water NOT Plant

### Clustering Algorithms

Cluster is group of like entities. Search engines use clustering technique to classify like concepts for more meaningful retrieval. Clustering is a technique by which similar kind
of data are grouped based on certain characteristics. This technique is very useful when there is large set of data for effective and efficient retrieval. For example, the faculty member of a university can be clustered according to their area of specialisation.

The clustering algorithm attempts to identify groups in a given set of data or population based on likeness of certain characteristics or traits. Thus it creates a picture of a big group and then inside several sub groups. The algorithms attempts to identify core entity which is also known as Centroid. Hence, centroid is the centre of the concept or the core concept. The other concepts are placed around the core concept. The likeness decreases with the increasing distance of the concept from the core concept or the centroid. Hence, to determine cluster membership and size, the algorithm evaluates the distance between concept and the core. Such techniques are highly dependent on use of statistics for generating clusters.

Error correction Algorithms

While typing the keywords often the searcher commits mistake. Search engines are equipped with the algorithms which correct the spelling and yield meaningful results. Google’s “Did you mean” uses these kind of algorithms. In case of incorrect spelling or if a word is not found in its database, Google suggests alternate spellings for the keyword.

Soundex and metaphone algorithms are examples of such kind of algorithms. Both algorithms are based on the pronunciation of a word.

In soundex algorithm a numeric code is assigned to each character used in a word and when search is performed, words with similar codes are also brought out in search result. Metaphone also works on same algorithm but unlike soundex which encodes a word on letter-by-letter basis, it encodes groups of letters i.e. a word.

Metaphone algorithm embodies more accurately the rules of pronunciation in language. Such algorithms are well established for English as a language. Both algorithms return all words that exactly match the desired word as well as all similar sounding words (homophones).

15.5.4 User Interface

User Interface is the part which interacts with the user. It is like a switch board for the user for invoking the system to perform search for needed information. There are two parts of User Interface:

- Search Interface
- Result Interface

Search Interface

This is the end from where users enter his/her search terms. It is one of the major components which initiate the communication between users and the system. The Search Interface performs following tasks:

1) Capturing user input/query

It is also known as front-end. The interface captures keywords given by users. It passes it to the search agent. The look and feel of the front-end should be easy to operate.

2) Search refinement

Search interface should have facility to refine the search. The refinement facility should
be given within the displayed search results. Hence, user interface should provide facilities for modifying search statement. Sometimes user interface gives facility for browsing pre-classified categories.

3) **Advance Search statement**

User interface should have another interface for advance level search. The advance level search includes, use of Boolean operators, image search, file-type search, language search, date-wise search and so on.

**Result Interface**

Display of search results is another important aspect of searching. It should be in user friendly format and customisable by the user.

Ranking of search result is an important feature. The search engines use algorithms for ranking. Google uses an algorithm called PageRank. In this, link analysis is done for ranking of retrieved documents. Some search engines rank their results based on the frequency of occurrence of search terms. Statistical techniques are widely used for this purpose.

**Self-Check Exercise**

**Note:**

i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of this Unit.

6) What is a “Web Crawler”?

7) What is Clustering?

8) What is Soundex Algorithm?

.......................................................................................................................

.......................................................................................................................

.......................................................................................................................

.......................................................................................................................

15.6 **CHALLENGES**

Searching on Internet has gone a long way. In the era of Web 2.0, search engines are well equipped with several algorithms like soundex, metaphones, page rank, etc. for improved retrieval of search results. One area which is still required to be improved is context based searching. Though clustering techniques are applied for setting the context but these methods are not fools proof. The implementation of Semantic web is ahead which promises search agents with specialised search strategies and features. These agents would not only perform the search, they will also guide the users in taking decision.

The idea of Semantic web is still evolving and it has still long way to go. But there are products available in the market which has started making mark on the Internet arena like, Wikipedia, Social networking sites, blogs and so on.

15.7 **SUMMARY**

This Unit highlights the features of search tools. Search engines are web based search tools to search the documents or objects over Internet. The Unit has discussed different
types of search engines i.e. search engine, meta-searching engine and search directory. Search engines have two kinds of search interfaces, simple and advance. Advance search facilitates different kinds of searches such as Boolean search, proximity search, truncation search, case sensitive search, fields level search, file types search, stop words, sorting, etc.

The Unit elaborates the architecture of search engines with reference to user interface, search agents and web crawler.

15.8 **ANSWERS TO SELF CHECK EXERCISES**

1) To locate information on web, search tools play an important role. The very first tool used for searching was Archie.

Internet offers various types of search tools:

   a) search engine
   b) subject directories
   c) metasearch engines

2) Search Engine:

   a) Google
   b) Yahoo!
   c) Alta-Vista
   d) Vivisimo (vivisimo.org)

Meta-search Engine:

   e) Dogpile
   f) Browsys

3) When Boolean operators are used in a search it is known as Boolean Search. The Logical operators such as AND, OR, NOT are known as Boolean Operators.

4) Proximity means ‘nearness’. Hence, when a search is based on the proximity or nearness between two words is known as Proximity Search.

5) PageRank is an algorithm used by Google for measuring weightage of results based on link analysis to rank the retrieved results.

6) Web crawler is a program used by search engine in order to extract data from the web pages so that pages can be searched using the search engine’s interface. It is also known as robot or spider.

7) Clustering is a technique by which similar kind of data are grouped based on certain characteristics. This technique is very useful when there is large set of data for effective and efficient retrieval.

8) Soundex is an error-correction phonetic algorithm. If there is minor error in the spelling of the search term or there are cases of homophones, this algorithm helps in retrieving the results.
15.9 KEYWORDS

Directories : Lists of pages classified into useful categories, (like Yahoo or Looksmart).

Exact Match : If a document contains exact match to the query, then only it will get retrieved. Increases precision of result but low recall value.

Query : A query is the combination of the word or words used for searching.

Recall : Total retrieved records against a query.

Precision : Total relevant records retrieved against a query.

Relevance : The extent to which retrieved records against a query satisfies the end-user.

Search engine : A program that indexes web documents and facilitates user to perform search on them.

Weighting : Weighting is a heuristic technique designed to improve the relevance ranking algorithms.

Index term : It is a pre-defined term which can be used to refer to the content of a document.

Full Text Search : It is a methodology in which all the words which compose the text of the document are used as indexing terms.

Fussy Model : It is “set theoretic model” of document retrieval based on fuzzy theory.

Inverted file : An index composed of vocabulary and list of occurrences.

15.10 REFERENCES AND FURTHER READING


7. Meta Search Engines. <www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html>
Internet Tools and Services


