RGPVNOTES.IN

Subject Name: **Modern Information Retrieva**

Subject Code: **CS-7004**

Semester: **7th**

Unit IV lecture notes

Introduction:

Web Search: Crawling and Indexes, Search Engine architectures, Link Analysis and ranking algorithms such as HITS and PageRank, Meta searches, Performance Evaluation of search engines using various measures, Introduction to search engine optimization.

A web search engine is a software system that is designed to search for information on the World WideWeb. The search results are generally presented in a line of results often referred to as search engineresults pages (SERPs).

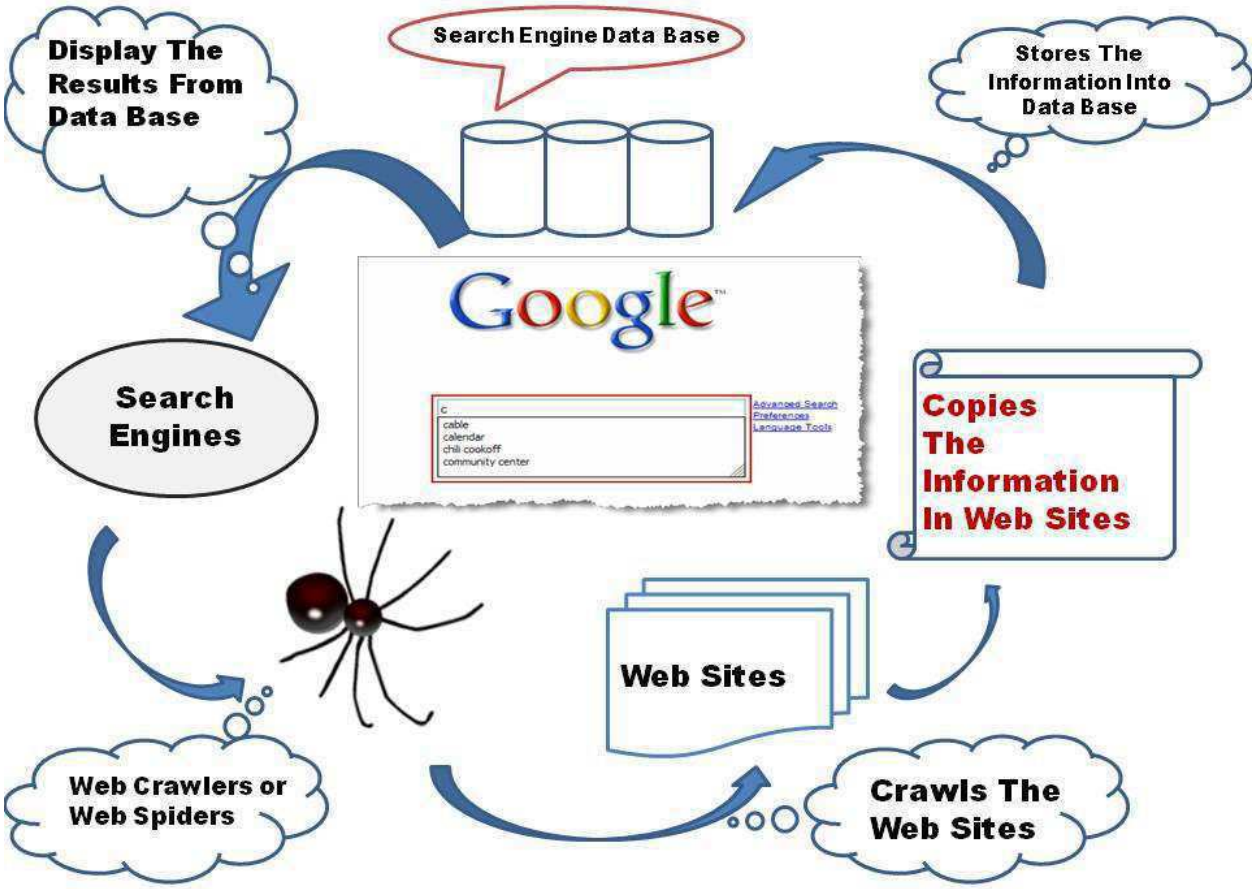


Fig.1 Web Search Engine Information Retrieval Process

Crawling and indexing the web

In layman's terms, indexing is the process of adding webpages into Google search. Depending upon which meta tag you used (index or NO-index), Google willcrawl and index your pages. A no-index tag means that that page will not be added into the web search'sindex. By default, every WordPress post and page isindexed.

Indexing a web page

Web indexing (or Internet indexing) refers to various methods for indexing the contents of a website or of the Internet as a whole. ... With the increase in the number of periodicals that have articles online, web indexing is also becoming important for periodical websites.

Crawling in SEO

Crawling is the process performed by search engine crawler, when searching for relevant websites on the index. For instance,Google is constantly sending out "spiders" or "bots" which is a search engine's automatic navigator to discover which websites contain the most relevant information related to certain keywords.

Search Engine architectures

Search Engine refers to a huge database of internet resources such as web pages, newsgroups, programs, images etc. It helps to locate information on World Wide Web.

User can search for any information by passing query in form of keywords or phrase. It then searches for relevant information in its database and return to the user.

Search Engine Components

Generally there are three basic components of a search engine as listed below:

1. Web Crawler

2.Database

3. Search Interfaces

Web crawler

It is also known as spider or bots. It is a software component that traverses the web to gather information.

Database

All the information on the web is stored in database. It consists of huge web resources.

Search Interfaces

This component is an interface between user and the database. It helps the user to search through the database.

Architecture

The search engine architecture comprises of the three basic layers listed below:

Content collection and refinement.

Search core

User and application interfaces

Technology used by search engine to crawl websites:

A Web crawler, sometimes called a spider, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing (webspidering). Web search engines and some other sites use Web crawling orspidering software to update their web content or indices of others sites' web content.

Search Engine Architecture:

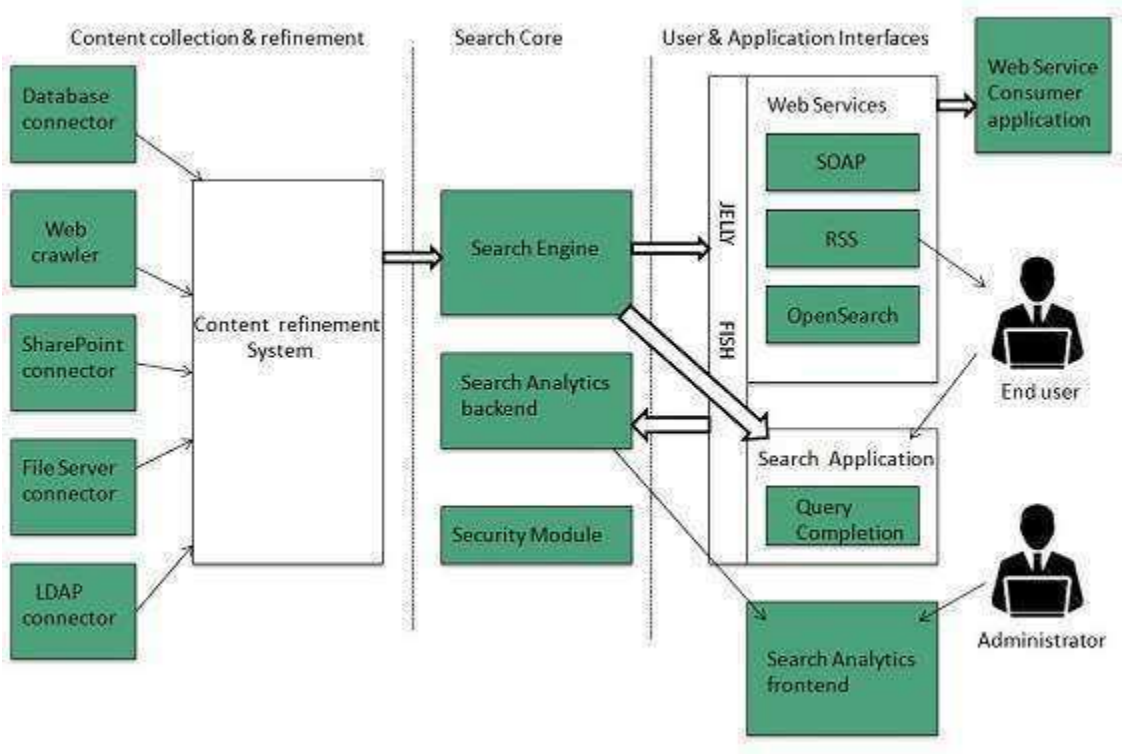


Fig.2 Search Engine archiecture

Software architecture consists of software components, the interfaces provided by those components and the relationships between them.

Architecture of a search engine determined by two requirements:

i.)Effectiveness(quality of results)

ii.)Efficiency(speed: response time and throughput)

Search Engine Processing

Indexing Process

Indexing process comprises of the following three tasks:

Text acquisition

Text transformation

Index creation

Text acquisition

It identifies and stores documents for indexing.

Text Transformation

It transforms document into index terms or features.

Index Creation

It takes index terms created by text transformations and create data structures to suport fast searching.

Query Process

Query process comprises of the following three tasks:

User interaction

Ranking

Evaluation

Link Analysis:

Use of hyperlinks for ranking web search results.

Such link analysis is one of many factors considered by web search engines in computing a composite score for a web page on any given query.

The web as a graph:

Link analysis builds on two intuitions:

i.)The anchor text pointing to page B is a good description of page B.

ii.) The hyperlink from A to B represents an endorsement of page B, by the creator of page A. This is not always the case; for instance, many links amongst pages within a single website stem from the user of a common template. For instance, most corporate websites have a pointer from every page to a page containing a copyright notice - this is clearly not an endorsement. Accordingly, implementations of link analysis algorithms will typical discount such ``internal'' links.

Ranking:

Ranking of query results is one of the fundamental problems in information retrieval(IR), the scientific/engineering discipline behind search engines. Given a query q and a collection D of documents that match the query, the problem is to rank, that is, sort, the documents in D according to some criterion so that the "best" results appear early in the result list displayed to the user. Classically, ranking criteria are phrased in terms of relevance of documents with respect to an information need expressed in the query.

Ranking is often reduced to the computation of numeric scores on query/document pairs; a baseline score function for this purpose is the cosine similarity between tf–idf vectors representing the query and the document in a vector space model.

Ranking functions are evaluated by a variety of means; one of the simplest is determining the precision of the first k top-ranked results for some fixed k; for example, the proportion of the top 10 results that are relevant, on average over many queries.

HITS algorithm:

Hyperlink-Induced Topic Search (HITS; also known as hubs and authorities) is a link analysisalgorithm that rates Web pages, developed by Jon Kleinberg. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that they held, but were used as compilations of a broad catalog of information that led users direct to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs.

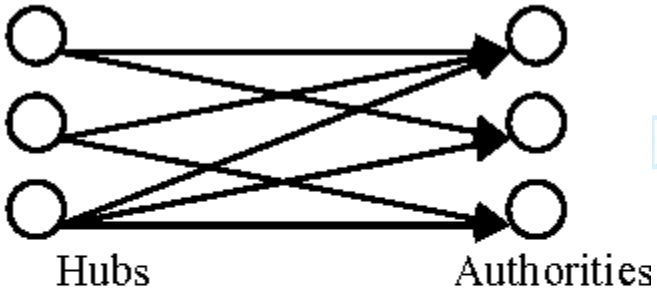


Figure 1. Hubs and authorities

$$a_i = \sum_{j \in B(i)} h_j \qquad \text{and} \qquad h_i = \sum_{j \in F(i)} a_j$$

Algorithm:

In the HITS algorithm, the first step is to retrieve the most relevant pages to the search query. This set is called the root set and can be obtained by taking the top pages returned by a text-based search algorithm. A base set is generated by augmenting the root set with all the web pages that are linked from it and some of the pages that link to it. The web pages in the base set and all hyperlinks among those pages form a focused subgraph. The HITS computation is performed only on this focused subgraph. According to Kleinberg the reason for constructing a base set is to ensure that most (or many) of the strongest authorities are included.

The algorithm performs a series of iterations, each consisting of two basic steps:

Authority update: Update each node's authority score to be equal to the sum of the hub scores of each node that points to it. That is, a node is given a high authority score by being linked from pages that are recognized as Hubs for information.

Hub update: Update each node's hub score to be equal to the sum of the authority scores of each node that it points to. That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.

The Hub score and Authority score for a node is calculated with the following algorithm:

- Start with each node having a hub score and authority score of 1.
- Run the authority update rule
- Run the hub update rule
- Normalize the values by dividing each Hub score by square root of the sum of the squares of all Hub scores, and dividing each Authority score by square root of the sum of the squares of all Authority scores.
- Repeat from the second step as necessary.

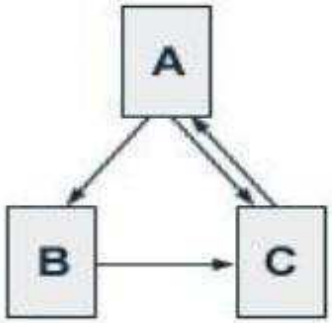
PageRank algorithm:

The Hub score and Authority score for a node is calculated with the following algorithm:

- Start with each node having a hub score and authority score of 1.
- Run the authority update rule
- Run the hub update rule
- Normalize the values by dividing each Hub score by square root of the sum of the squares of all Hub scores, and dividing each Authority score by square root of the sum of the squares of all Authority scores.
- Repeat from the second step as necessary.

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best-known.



Consider an imaginary web of 3 web pages.
And the inbound and outbound link structure is as shown in the figure. The calculations can be done by following method :

PR(A) = 0.5 + 0.5 PR(C)
= 0.5 + (0.5*1)
= 1

PR(B) = 0.5 + 0.5 (PR(A) / 2)
= 0.5 + 0.5 (1/2)
= 0.5 + (0.5 * 0.5)
= 0.5 + 0.25
= 0.75

PR(C) = 0.5 + 0.5 ((PR(A) / 2)+ PR (B))
= 0.5 + 0.5 (1/2 + 0.75)
= 0.5 + 0.5 (1.25)
= 0.5 + 0.625
= 1.125

Description:

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element *E* is referred to as the *PageRank of E* .Other factors like *Author Rank* can contribute to the importance of an entity.

Algorithm:

The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

A probability is expressed as a numeric value between 0 and 1. A 0.5 probability is commonly expressed as a "50% chance" of something happening. Hence, a PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank.

Meta-Searches:

A **metasearch engine** (or aggregator) is a search tool that uses another search engine's data to produce its own results from the Internet.

Metasearch engines take input from a user and simultaneously send out queries to third party search engines for results. Sufficient data is gathered, formatted by their ranks and presented to the users.

Metasearch engines have their own sets of unique problems. All of the websites stored on search engines are different, which draws irrelevant content. Problems such as spammingreduce result accuracy. The process of fusion aims to tackle this issue and improve the engineering of a metasearch engine.

Advantages:

By sending multiple queries to several other search engines this extends the search coverage of the topic and allows more information to be found. They use the indexes built by other search engines, aggregating and often post-processing results in unique ways. A metasearch engine has an advantage over a single search engine because more results can be retrieved with the same amount of exertion.It also reduces the work of users from having to individually type in searches from different engines to look for resources.

Metasearching is also a useful approach if the purpose of the user's search is to get an overview of the topic or to get quick answers. Instead of having to go through multiple search engines like Yahoo! or Google and comparing results, metasearch engines are able to quickly compile and combine results. They can do it either by listing results from each engine queried with no additional post-processing (Dogpile) or by analyzing the results and ranking them by their own rules (IxQuick, Metacrawler, and Vivismo)

Disadvantages:

Metasearch engines are not capable of decodingquery forms or able to fully translate query syntax. The number of links generated by metasearch engines are limited, and therefore do not provide the user with the complete results of a query. The majority of metasearch engines do not provide over ten linked files from a single search engine, and generally do not interact with larger search engines for results

Metasearching also gives the illusion that there is more coverage of the topic queried, particularly if the user is searching for popular or commonplace information. It's common to end with multiple identical results from the queried engines. It is also harder for users to search with advanced search syntax to be sent with the query, so results may not be as precise as when a user is using an advanced search interface at a specific engine. This results in many metasearch engines using simple searching.

Performance evaluation of search engine:

Many metrics exist to perform the task of search engine evaluation that are either looking for the experts judgments or believe in searchers decisions about the relevancy of the web documents. However, search logs can provide us information about how real users search

There are six criteria for search engine evaluation. These criteria are web coverage, dwell time, recall, precision, presentation and user efforts. We can explore the user efforts in the form of session duration, Ranked Precision and Clicks hits

Search Engines have become an integral part of daily internet usage. The search engine is the first stop for web users when they are looking for a product. Information retrieval may be viewed as aproblem of classifying items into one of two classes corresponding to interesting and uninteresting items respectively. A natural performance metric in this context is classification accuracy, defined as the fraction of the system's interesting/uninteresting predictions that agree with the user's assessments. On the other hand, the field of information retrieval has two classical performance evaluation metrics: precision, the fraction of the items retrieved by the system that are interesting to the user, and recall, the fraction of the items of interest to the user that are retrieved by the system.

Criteria for Evaluating Search Engine's Performance:

Many publications compare or evaluate Web search engines (e.g. Notess, 2000). Perhaps the best known of these are Search Engine Watch (http://www.searchenginewatch.com) and Search Engine Showdown (http://www.searchengineshowdown.com).

Recall in Search Engine Performance Evaluation:

Recall has always been a difficult measure to calculate because it requires the knowledge of the total number of relevant items in the collection.

Precision in Search Engine Performance Evaluation:
Precision is always reported in formal information retrieval experiments. However, there are variations in the way it is calculated depending on how relevance judgments were made.

Precisionmeasures the ability of Search Engine to produce only relevant results. Precision is the ratio between the number of relevant documents retrieved by the system and the total number of documents retrieved. An ideal system would produce a precision score of1, i.e. every document retrieved by the system is judged relevant. Precision is relatively easy to calculate, which mainly accounts for its popularity. But a problem with precision in the search engine context is the number of results usually given back in response to typical queries.
In many cases, search engines return thousands of results. In an evaluation scenario, it is not feasible to judge so many results. Therefore, cut-off rates (e.g. 20 for the first 20 hits) are used in retrieval tests.
The coverage of a search enginecan be determined as the total number of pages returned by the search engine.

Search engine Optimization:

Search engine optimization (**SEO**) is the process of affecting the online visibility of a website or a web page in a web search engine's unpaid results—often referred to as "natural", "organic", or "earned" results.

As an Internet marketing strategy, SEO considers how search engines work, the computer programmed algorithms which dictate search engine behavior, what people search for, the actual search terms or keywords typed into search engines, and which search engines are preferred by their targeted audience. Optimizing a website may involve editing its content, adding content, doing HTML, and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines.

Search Engine Optimization (SEO) is the activity of optimizing web pages or whole sites in order to make them search engine friendly, thus getting higher positions in search results.

SEO stands for Search Engine Optimization. SEO is all about optimizing a website for search engines. SEO is a technique for:

designing and developing a website to rank well in search engine results.

improving the volume and quality of traffic to a website from search engines.

marketing by understanding how search algorithms work, and what human visitors might search.

Search engines such as Google and Yahoo! often update their relevancy algorithm dozens of times per month. When you see changes in your rankings, it is due to an algorithmic shift or something else beyond your control. Although the basic principle of operation of all search engines is the same, the minor differences between their relevancy algorithms lead to major changes in the relevancy of results

What is On-Page and Off-Page SEO?

Conceptually, the re are two ways of optimization:

•On-Page SEO –

It includes providing good content, good keywords selection ,putting keywords on correct places, giving appropriate title to every page ,etc.

•Off-Page SEO-

It includes link building, increasing link popularity by submitting open directories, search engines, link exchange,etc

RGPVNOTES.IN

We hope you find these notes useful.

You can get previous year question papers at https://qp.rgpvnotes.in .

If you have any queries or you want to submit your study notes please write us at rgpvnotes.in@gmail.com

LIKE & FOLLOW US ON FACEBOOK
facebook.com/rgpvnotes.in