Web Technology
MCA-124

Rajkumar
Research Scholar,
I.T.C.A.

MMMUT, Gorakhpur

September 14, 2020

Outline

Unit-I

- World Wide Web ✓
- WWW Architecture ✓
- Web Search Engine ✓
- Web Crawling ✓
- Web Indexing ✓
- Web Mining

References Books

1. Web Technologies, 1/e -Uttam K. Roy ,Oxford University Press, USA
2. Web Technology: Theory and Practice -M. Srinivasan, Pearson Education India,
3. Deitel, Deitel and Nieto, Internet and Worldwide Web -How to Program, 5th Edition, PHI, 2011.
4. Developing Web Application-Second Editon -Ralph Moseley & M. T. Savaliya, Wiley
5. Web Programming Step by Step, Stepp/Miller/Kirst, 2nd edition, 2009

World Wide Web

1. World Wide Web, which is also known as a Web, is a collection of websites or web pages stored in web servers and connected to local computers through the internet.
2. websites contain text pages, digital images, audios, videos, etc.
3. A web page is given an online address called a Uniform Resource Locator (URL). A particular collection of web pages that belong to a specific URL is called a website, e.g., www.facebook.com, www.google.com, etc.

Difference between World Wide Web and Internet

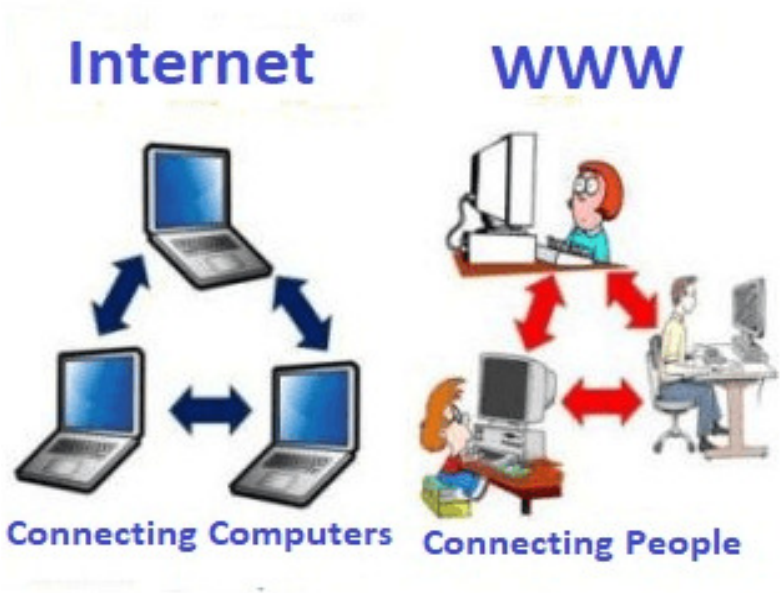

Figure: 1 World Wide Web and Internet

How the World Wide Web Works?

WWW is a collection of websites connected to the internet so that people can search and share information. Now, let us understand how it works!

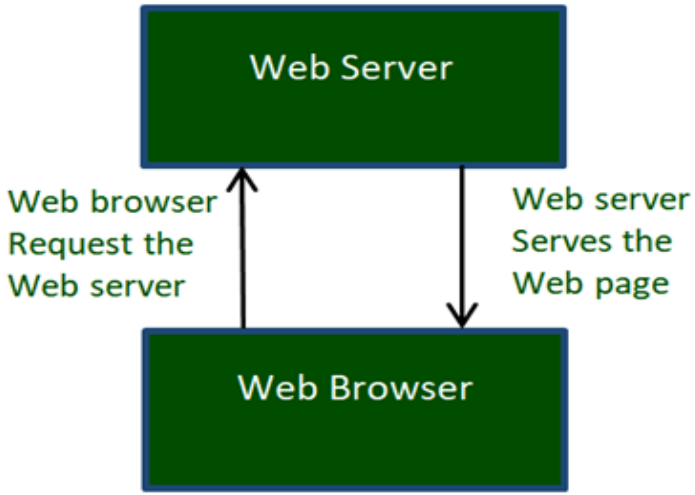

Figure: 2 World Wide Web Working process

How the World Wide Web Works?

1. The Web works as per the internet's basic client-server format as shown in the following image.
2. The servers store and transfer web pages or information to user's computers on the network when requested by the users.
3. A web server is a software program which serves the web pages requested by web users using a browser.
4. The computer of a user who requests documents from a server is known as a client.

The moment you open the browser and type a URL in the address bar or search something on Google, the WWW starts working. There are three main technologies involved in transferring information (web pages) from servers to clients (computers of users). These technologies include Hypertext Markup Language (HTML), Hypertext Transfer Protocol (HTTP) and Web browsers.

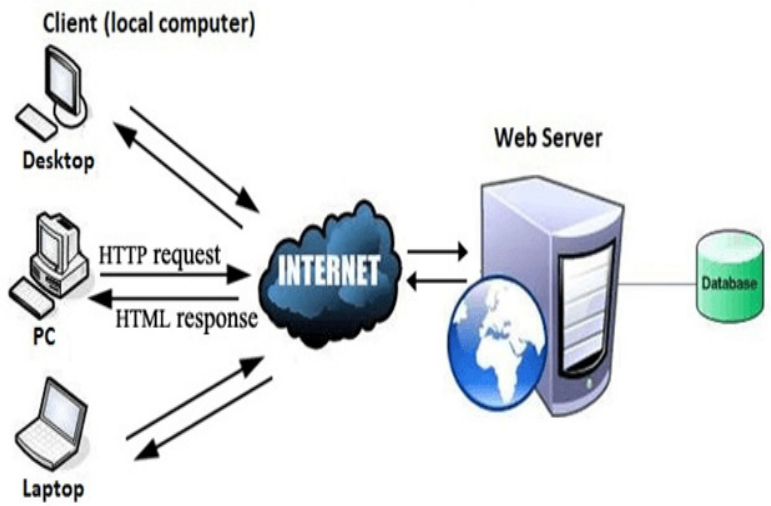

Figure: 3 World Wide Web Working process

WWW Architecture

WWW architecture is divided into several layers as shown in the following diagram:

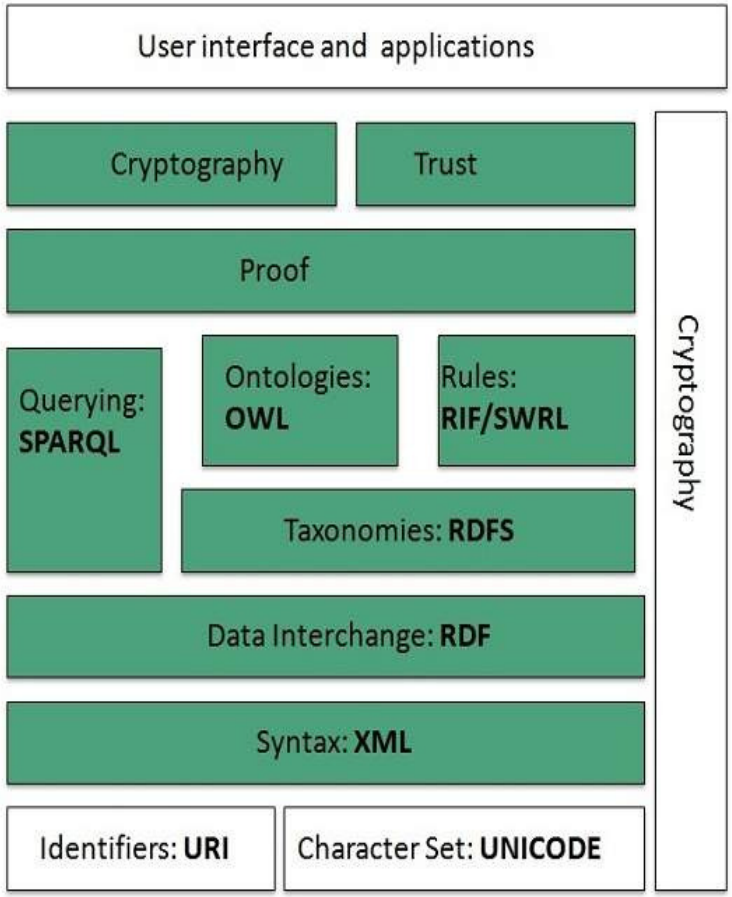

Figure: 4 World Wide Web Architecture

WWW Architecture I

Identifiers and Character Set
Uniform Resource Identifier (URI) is used to uniquely identify resources on the web and UNICODE makes it possible to built web pages that can be read and write in human languages.

Syntax
XML (Extensible Markup Language) helps to define common syntax in semantic web.

Data Interchange
Resource Description Framework (RDF) framework helps in defining core representation of data for web. RDF represents data about resource in graph form.

Taxonomies
RDF Schema (RDFS) allows more standardized description of taxonomies and other ontological constructs.

WWW Architecture II

Ontologies Web Ontology Language (OWL) offers more constructs over RDFS. It comes in following three versions:

1. OWL Lite for taxonomies and simple constraints.
2. OWL DL for full description logic support.
3. OWL for more syntactic freedom of RDF.

Rules RIF and SWRL offers rules beyond the constructs that are available from RDFs and OWL. Simple Protocol and RDF Query Language (SPARQL) is SQL like language used for querying RDF data and OWL Ontologies.

Proof All semantic and rules that are executed at layers below Proof and their result will be used to prove deductions.

Cryptography Cryptography means such as digital signature for verification of the origin of sources is used.

User Interface and Applications On the top of layer User interface and Applications layer is built for user interaction.

WWW Architecture III

How it work

WWW works on client- server approach. Following steps explains how the web works:

1. User enters the URL (say, http://www.xyz.com) of the web page in the address bar of web browser.
2. Then browser requests the Domain Name Server for the IP address corresponding to www.xyz.com.
3. After receiving IP address, browser sends the request for web page to the web server using HTTP protocol which specifies the way the browser and web server communicates.
4. Then web server receives request using HTTP protocol and checks its search for the requested web page. If found it returns it back to the web browser and close the HTTP connection.
5. Now the web browser receives the web page, It interprets it and display the contents of web page in web browser's window.

WWW Architecture IV

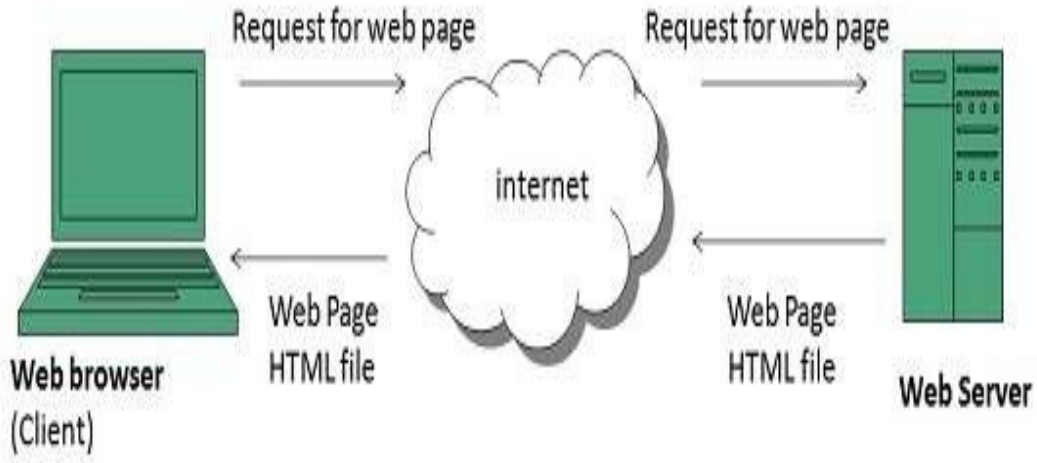

Figure: 5 World Wide Web Architecture

Web Search Engine

Search engines are answer machines. They exist to discover, understand, and organize the internet's content in order to offer the most relevant results.

How do search engines work?
Search engines have three primary functions:

- **Crawl:** Scour the Internet for content, looking over the code/content for each URL they find.
- **Index:** Store and organize the content found during the crawling process. Once a page is in the index, it's in the running to be displayed as a result to relevant queries.
- **Rank:** Provide the pieces of content that will best answer a searcher's query, which means that results are ordered by most relevant to least relevant.

Web Crawling

A key motivation for designing Web crawlers has been to retrieve Web pages and add their representations to a local repository.

- What is the "Web Crawling
- What are the uses of Web Crawling
- How API are used

-**A Web crawler** *(also known as a Web spider, Web robot, or—especially in the FOAF community—Web scutter)* is a program or automated script that browses the World Wide Web in a
-methodical
-automated manner

-Other less frequently used names for Web crawlers are ants, automatic indexers, bots, and worms.

What the Crawlers are...

Crawlers are computer programs that roam the Web with the goal of automating specific tasks related to the Web.The role of Crawlers is to collect Web Content.

Basic crawler operation

1. Begin with known "seed" pages
2. Fetch and parse them
3. Extract URLs they point to
4. Place the extracted URLs on a Queue
5. Fetch each URL on the queue and repeat URLs on a Queue

Several Types of Crawlers

- **Batch Crawlers**- Crawl a snapshot of their crawl space, until reaching a certain size or time limit.
- **Incremental Crawlers**- Continuously crawl their crawl space, revisiting URL to ensure freshness.
- **Focused Crawlers**- Attempt to crawl pages pertaining to some topic/theme, while minimizing number of off topic pages that are collected.
- ...

The challenges of "Web Crawling"

There are three important characteristics of the Web that make crawling it very difficult:

1. Its large volume
2. fast rate of change
3. Dynamic page generation

The working of a web crawler is as follows...

- Initializing the seed URL or URLs.
- Adding it to the frontier.
- Selecting the URL from the frontier.
- Fetching the web-page corresponding to that URLs.
- Parsing the retrieved page to extract the URLs.
- Adding all the unvisited links to the list of URL i.e. into the frontier.
- Again start with step 2 and repeat till the frontier is empty.

Architecture of a Web Crawler

Architecture of a web crawler Figure shows the generalized architecture of web crawler. It has three main components:a **frontier** which stores the list of URL's to visit, **Page Downloader** which download pages from WWW andWeb **Repository** receives web pages from a crawler and stores it in the database.

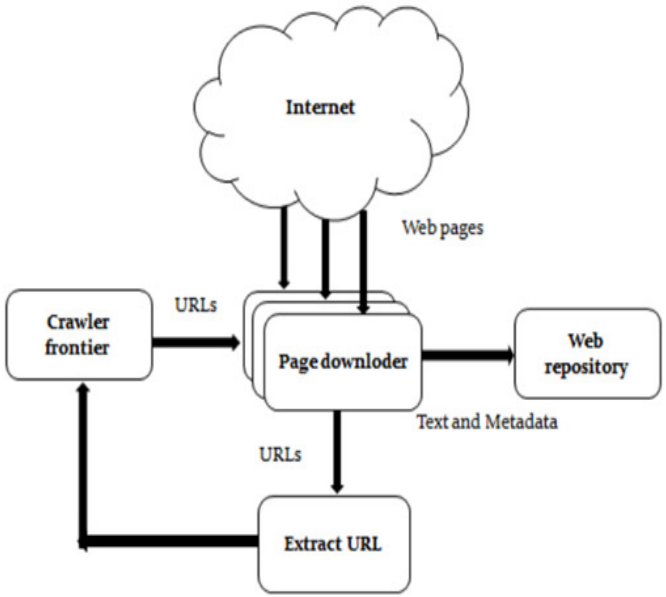

Figure: 6 Web Crawling Working process

Flow of web crawler

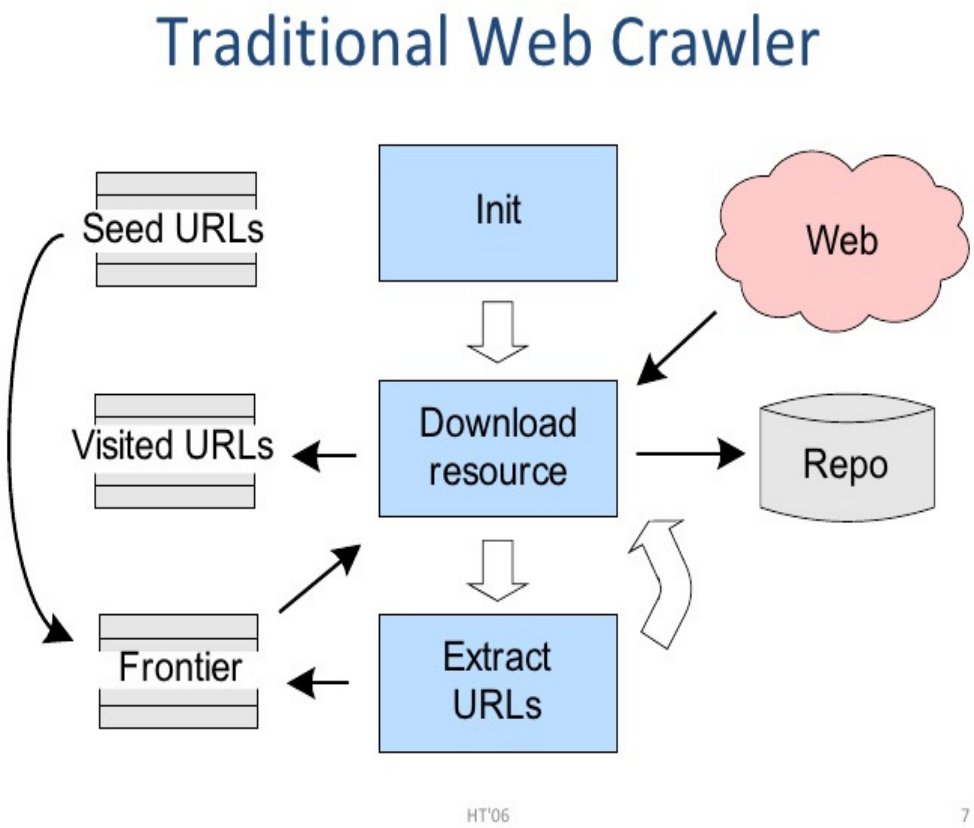

Figure: 7 Flow of web crawler

Search Engine Index

Search engine Index
Search engines process and store information they find in an index, a huge database of all the content they've discovered and deem good enough to serve up to searchers.

Search engine Ranking
When someone performs a search, search engines scour their index for highly relevant content and then orders that content in the hopes of solving the searcher's query.

This ordering of search results by relevance is known as **ranking**.

Web Mining

Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services.

The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

Applications of Web Mining:

- Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages.
- It is used for Web Searching e.g., Google, Yahoo etc and Vertical Searching e.g., FatLens, Become etc.
- Web mining is used to predict user behavior.
- Web mining is very useful of a particular Website and e-service e.g., landing page optimization.

Web Mining I

Web mining can be broadly divided into three different types of techniques of mining:

1. Web Content Mining
2. Web Structure Mining
3. Web Usage Mining

These are explained as following below.

Web Content Mining:

- Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc.
- The Content data is the group of facts that a web page is designed. It can provide effective and interesting patterns about user needs.
- The Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining.
- This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.

Web Structure Mining:

- Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages.
- Structure mining basically shows the structured summary of a particular website. It identifies relationship between web pages linked by information or direct link connection.
- To determine the connection between two commercial websites, Web structure mining can be very useful.

Web Usage Mining:

- Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets.
- These patterns enable you to understand the user behaviors or something like that. In web usage mining, user access data on the web and collect data in form of logs.
- So, Web usage mining is also called log mining.

References

- Web Enabled Commercial Application Development Using HTML, DHTML, Java Script, Perl & CGI -Ivan Bayross (BPB Publication), 2005.
- Internet and WebTechnologies, RAJ KAMAL, Tata McGraw Hill.
- Java Server Pages–Hans Bergsten, SPD O'Reilly.
- www.w3c.org

Thank You for Your Attention!

For any query contact me:

rajkumargaur[AT]mmmut.ac.in
8299722827