Information retrieval (IR)

By: **Dr. LOUNNAS Bilal**

What is Internet?

The Internet is essentially a global network of computing resources.

- A network of networks based on the TCP/IP communications protocol.
- A community of people who use and develop those networks.

Internet-Based Services

Some of the basic services available to Internet users are:

- Email: A fast, easy, and inexpensive way to communicate with other Internet users around the world.
- Telnet: Allows a user to log into a remote computer as though it were a local system.
- FTP: Allows a user to transfer virtually every kind of file that can be stored on a computer from one Internet-connected computer to another.
- UseNet news: A distributed bulletin board that offers a combination news and discussion service on thousands of topics.
- World Wide Web (WWW): A hypertext interface to Internet information resources.

What is WWW?

WWW stands for World Wide Web. A technical definition of the World Wide Web is - All the resources and users on the Internet that are using the Hypertext Transfer Protocol (HTTP).

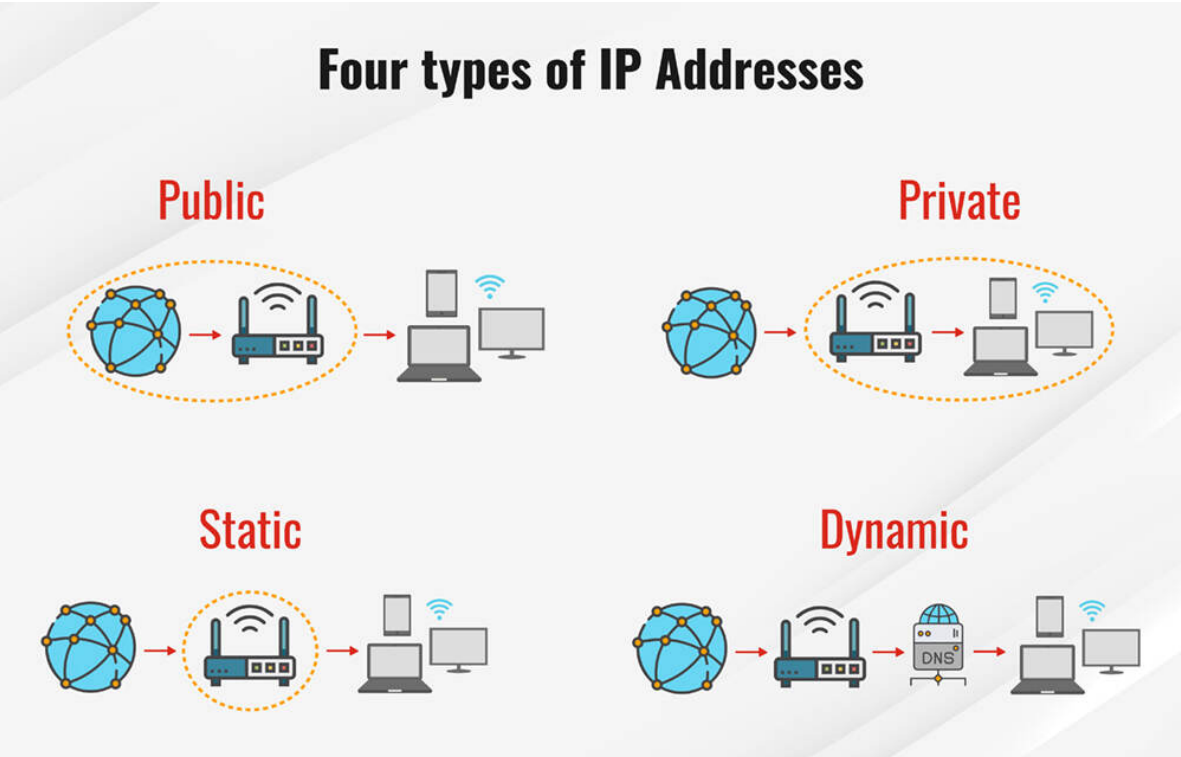
World Wide Web Consortium (W3C): The World Wide Web is the universe of network-accessible information, an embodiment of human knowledge.

- billions of documents
- authored by millions of diverse people
- distributed over millions of computers, connected by variety of media

What is Web Server?

Every Website sits on a computer known as a Web server. This server is always connected to the internet

Every Web server that is connected to the Internet is given a unique IP Address (Internet Protocol).

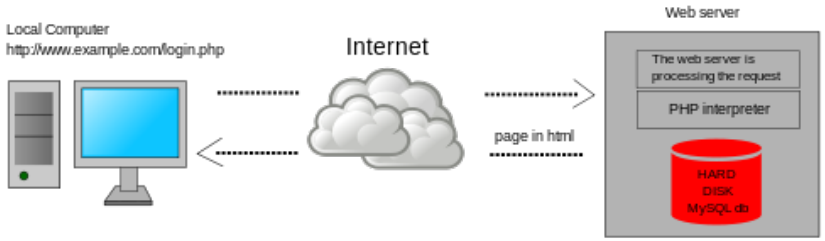


What is Web Site?

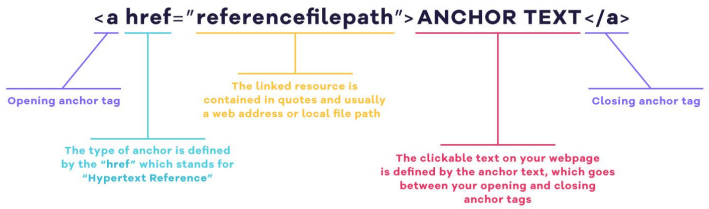
A website is a collection of web pages and related content that is identified by a common domain name and published on at least one web server.

- Domain name (Location)
- Pages (Content)

What is Web Page?



HTML stands for Hyper Text Markup Language. This is the language in which we write web pages for any Website



A hyperlink or simply a link is a selectable element in an electronic document that serves as an access point to other electronic resources.

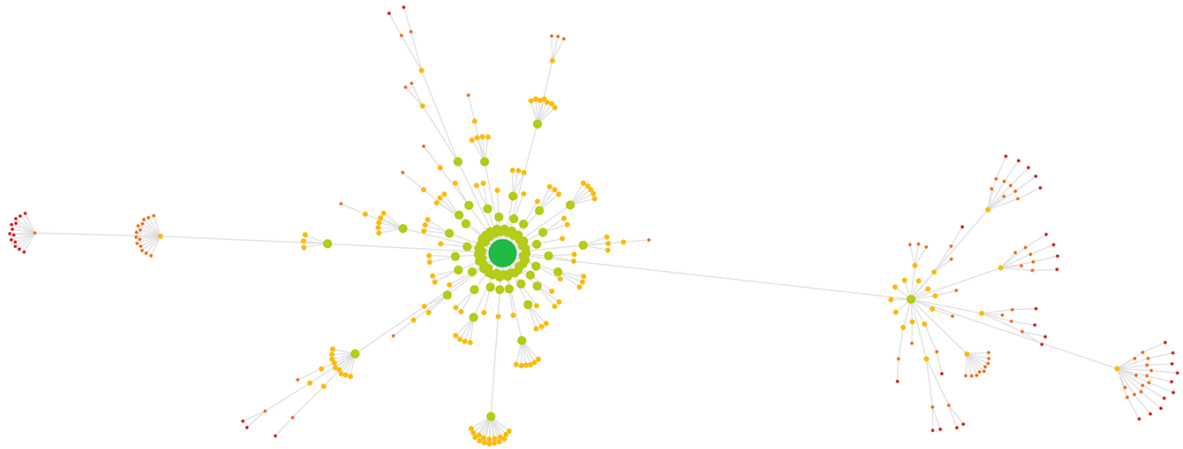
More concept

- DNS (Domain Name System)
- W3C (World Wide Web Consortium)
- SMTP Server (Simple Mail Transfer Protocol)
- ISP (Internet Service Provider)
- Web Browser
- HTTP
- ...

Building Search Engine

- Building Index (Crawling)
  web pages are particularly easy to copy, since they are meant to be retrieved over the Internet by browsers. This instantly solves one of the major problems of getting information to search
- Searching over this obtained resources (Freshness)

Crawl definition



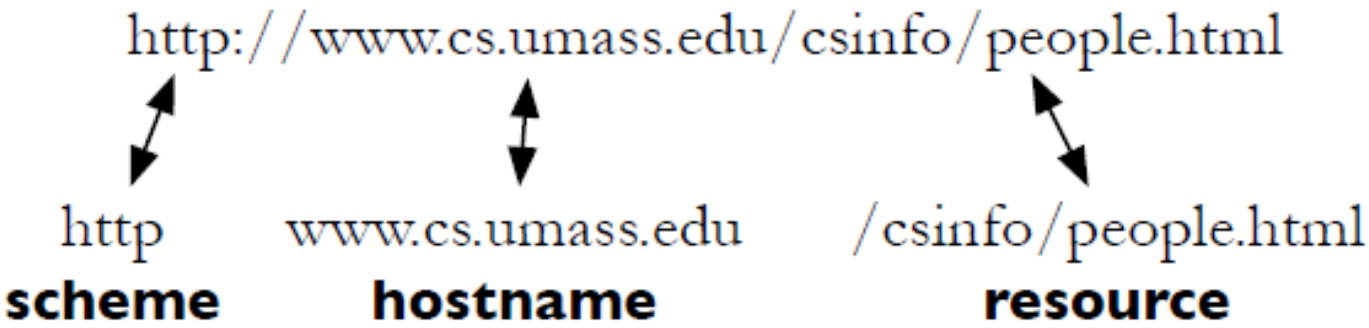
Crawl describes a bot, script, or software program that visits a web page and grabs its content and links. Once completed, the program visits the next link in the list or a link obtained from the web page it had recently visited.

Crawling issues

Challenges

- There are at least tens of billions of pages on the Internet.
  Even if the number of pages in existence today could be measured exactly, that number would be immediately wrong, because pages are constantly being created
- web pages are usually not under the control of the people building the search engine database.
  - there is no easy way to find out how many pages there are on the site.
  - The owners may not want you to copy (later robots.txt) some of the data.

Retrieving Web Pages



- Each web page on the Internet has its own unique URL (uniform resource locator).
- Use a protocol called Hypertext Transfer Protocol, or HTTP, to exchange information with client software (Web browsers Or web crawlers).

Web browsers Or web crawlers

Steps of connecting web browsers Or web crawlers

1. Connects to a domain name system (DNS) server
2. The DNS server translates the hostname into an internet protocol (IP) address.
3. Once the connection is established, the client program sends an HTTP request to the web server to request a page.
   GET /csinfo/people.html HTTP/1.0
   This simple request asks the server to send the page called /csinfo/people.html back to the client, using version 1.0 of the HTTP protocol specification

Web browsers Or web crawlers

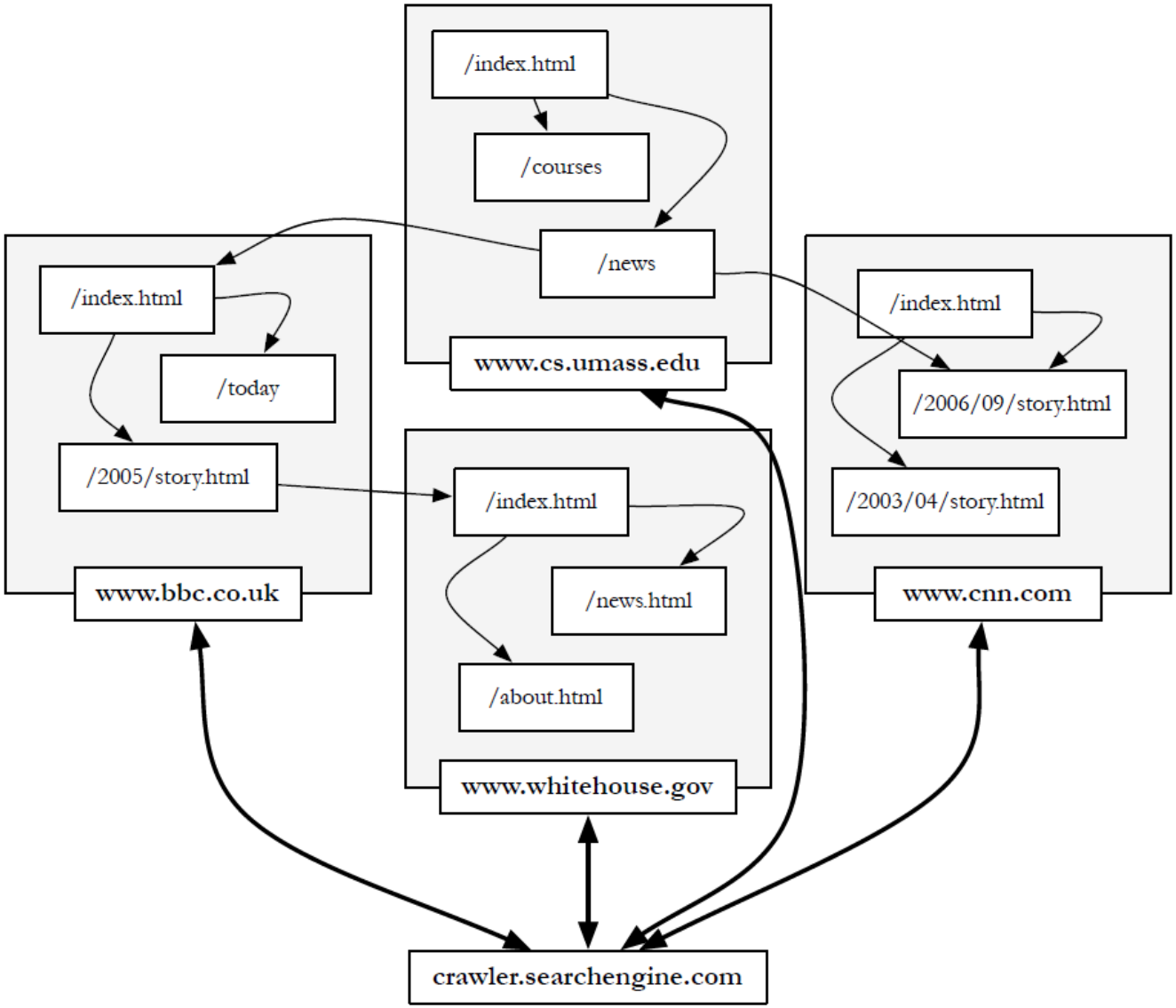
Steps of connecting web browsers Or web crawlers

4. The server sends the contents of that file back to the client
5. If the client wants more pages, it can send additional requests; otherwise, the client closes the connection.

A client can also fetch web pages using POST requests. A POST request is like a GET request, except that it can send additional request information to the server.

- Downloading pages.
- Finding URLs.

Web crawlers example



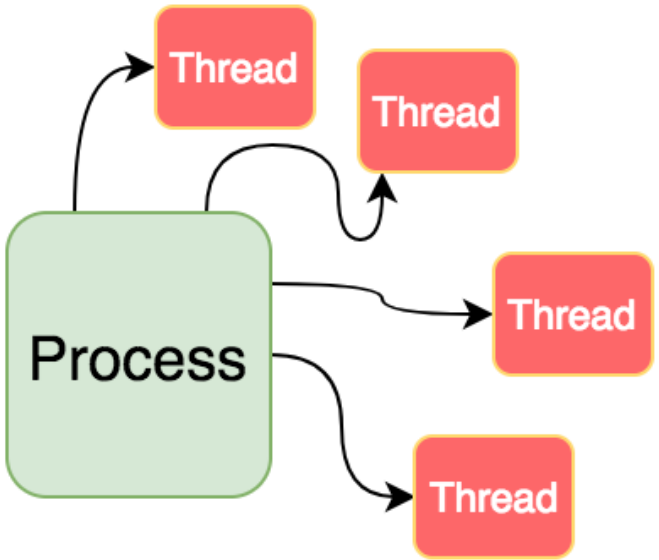
Web crawlers

Steps of web crawlers tasks - Downloading pages, Finding URLs.

1. Starts with a set of seeds
   set of URLs given to it as parameters
2. These seeds are added to a URL request queue (frontier)
   The frontier may be a standard queue, or it may be ordered so that important pages move to the front of the list
3. The crawler starts fetching pages from frontier. Once a page is downloaded, it is parsed to find link tags that might contain other useful URLs to fetch, if the crawler finds a new URL than add to frontier

This process continues until the crawler either runs out of disk space to store pages or runs out of useful links to add to the request queue.

Web crawlers threads



1. One thread - it would not be very efficient
   because web crawler spends a lot of its time waiting for responses
2. Multiple thread - is good for person running web crawler
   but not necessarily good for the person running the web server on the other end.

Web crawlers policies

The behavior of a web crawler is the outcome of a combination of policies:

1. a selection policy that states which pages to download.
   Robots.txt
2. a re-visit policy that states when to check for changes to the pages.
   Freshness
3. a duplication policy
4. a politeness policy that states how to avoid overloading web sites.
5. a parallelization policy that states how to coordinate distributed web crawler.

Web crawlers robots.txt example

```
User-agent: *
Disallow: /private/
Disallow: /confidential/
Disallow: /other/
Allow: /other/public/

User-agent: FavoredCrawler
Disallow:

Sitemap: http://mysite.com/sitemap.xml.gz
```

An example robots.txt file

Web crawlers Freshness example

Client request:

```
HEAD /csinfo/people.html HTTP/1.1
Host: www.cs.umass.edu
```

Server response:

```
HTTP/1.1 200 OK
Date: Thu, 03 Apr 2008 05:17:54 GMT
Server: Apache/2.0.52 (CentOS)
Last-Modified: Fri, 04 Jan 2008 15:28:39 GMT
ETag: "239c33-2576-2a2837c0"
Accept-Ranges: bytes
Content-Length: 9590
Connection: close
Content-Type: text/html; charset=ISO-8859-1
```

An HTTP HEAD request and server response

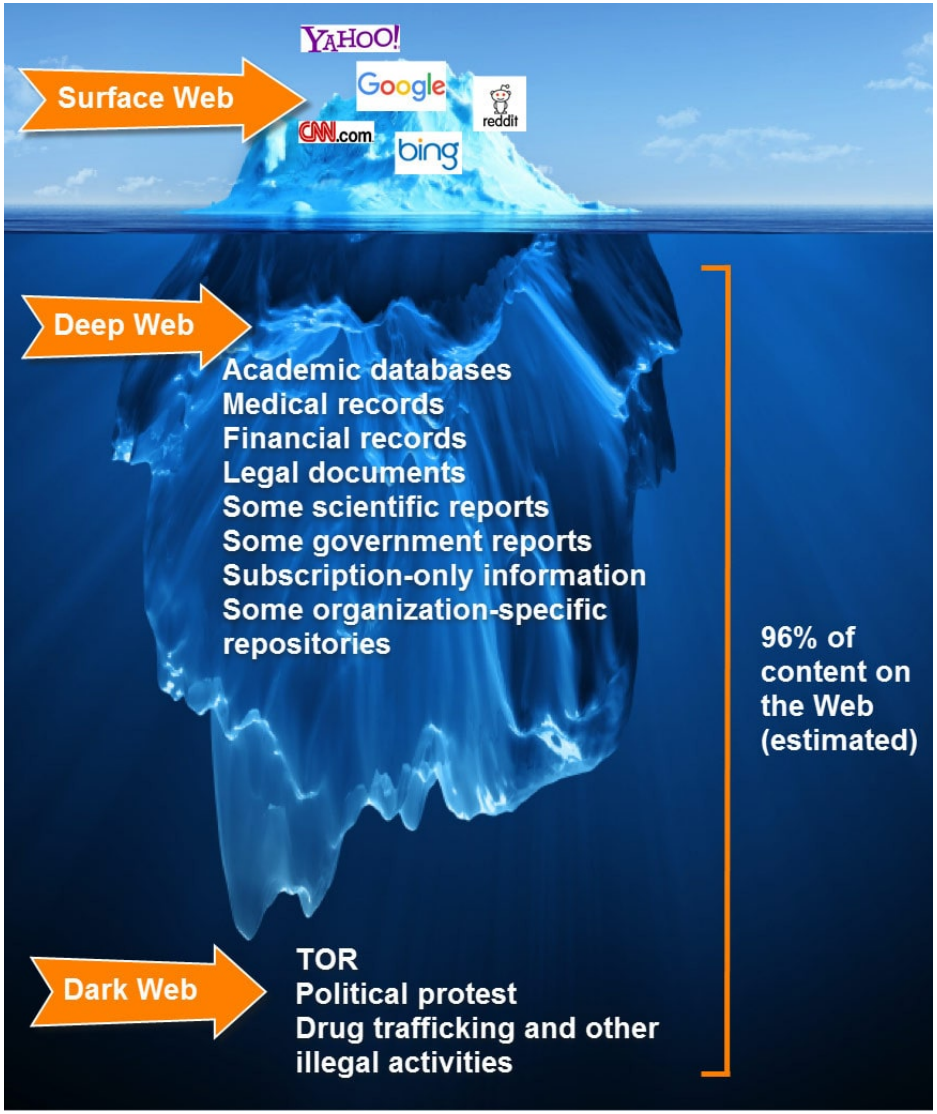
Web crawlers issues

1. Unknown pages.
   Both crawler and site owner
2. Frequency of update of a page.
   Both crawler and site owner
3. Solution: sitemap

Web crawlers site map

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.company.com/</loc>
    <lastmod>2008-01-15</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.7</priority>
  </url>
  <url>
    <loc>http://www.company.com/items?item=truck</loc>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.company.com/items?item=bicycle</loc>
    <changefreq>daily</changefreq>
  </url>
</urlset>
```

Deep web


YAHOO!
Google
reddit
CNN.com
bing
Surface Web
Deep Web
Academic databases
Medical records
Financial records
Legal documents
Some scientific reports
Some government reports
Subscription-only information
Some organization-specific repositories
96% of content on the Web (estimated)
Dark Web
TOR
Political protest
Drug trafficking and other illegal activities

Other crawler

Crawler type and Challenges

1. Crawling Documents and Email.
2. Document Feeds (RSS)
3. The Conversion Problem
   Many type: HTML, PDF, XLSX, RTF
4. Storing the Documents.
   Using a Database System, Compression and Large Files, Update
5. Detecting Duplicates.
   Hashage, CheckSUM
6. Removing Noise.
   80% of page content is noise