SWE 363: Web Engineering & Development

Module 1-2

Search Engines

Objectives

- ❑ Identify the search engine concept and its types
- ❑ Explain Search Engine mechanism
- ❑ Discuss Optimization techniques

Outline

- ❑ Introduction
- ❑ Search engines types
- ❑ Stats of Search Engines
- ❑ Major components of a search engine
- ❑ Challenges Faced by Search Engines
- ❑ How does Search Engine works?
  - ▪ PageRank algorithm
- ❑ Search engines optimization
  - ▪ White-hat SEO
  - ▪ black-hat SEO
- ❑ Invisible web

Used Resources

- ❑ Connolly, R. (2015). *Fundamentals of web development*. Pearson Education.
- ❑ Henzinger, Monika Rauch, Rajeev Motwani, and Craig Silverstein. "*Challenges in web search engines*." In IJCAI

- ❑ Other links:
  - ▪ http://www.googleguide.com/google_works.html
  - ▪ https://www.netmarketshare.com
  - ▪ https://en.wikipedia.org/wiki/Web_search_engine
  - ▪ http://searchengineland.com
  - ▪ http://computer.howstuffworks.com/internet/basics/search-engine1.htm
  - ▪ http://libguides.astate.edu/c.php?g=14516&p=78177
  - ▪ https://library.laguardia.edu/invisibleweb
  - ▪ http://searchengineland.com/write-meta-description-gets-clickthroughs-207922

Introduction

- Search engine is the popular term for an information retrieval (IR) system.
- The Web is a repository of information on almost any topic – huge volume of online content
- To help find information quickly on a specific topic, use
  - Index of topics (directory or catalog)
    - No central catalog is possible for the web
  - Search engines search and retrieve documents for specified keywords

- A web search engine is a software system that is designed to search information of the web
  - It uses the keywords to search for documents that related to the specified keywords and then put the result in order of relevance to the topic that was searched for.

Search Engines

- Without Search Engine and a huge space of web → It would be impossible to search for the information that is specifically needed
  - This is why search engines are used to filter the information that is on the internet and transform it into results (in the order of relevance to the topic that was searched for) where each individual can easily access and use

- Search engines differ in their capabilities and the way they work
  - Examples: Google, Yahoo, Bing, Baidu, Ask.com, etc.
    - See a list of them at http://www.thesearchenginelist.com/

- There are many search engines
  - Generic search engines - used for general search
  - Vertical search engines - focus on specific topics (e.g. bioinformatics, medical, jobs, business, real estate, travel, etc)

Types of Search Engines

Directories-based

- Directories depend on human editors to create their listings or the database.
  - A website is submitted to the directory and must be approved for inclusion by editorial staff.
- How the indexing in Directories-based works?
- Site owner submits a short description of the site to the directory along with category it is to be listed.
  - Submitted site is then manually reviewed and added in the appropriate category or rejected for listing.
- Keywords entered in a search box will be matched with the description of the sites. This means the changes made to the content of a web pages are not taken into consideration as it is only the description that matters.
  - A good site with good content is more likely to be reviewed for free compared to a site with poor content.

Types of Search Engines

Directories-based…

- ❑ Human-powered directories are good when you are interested in a general topic of search.
  - Examples: Yahoo Directory, DMOZ

- ❑ **Advantages:**
  - Each page is reviewed for relevance and content before being included.
  - Less results sometimes means finding what you need quicker.

- ❑ **Disadvantages:**
  - Unfamiliar design and format.
  - Delay in creation of a website and it's inclusion in the directory.
  - May have trouble with more ambiguous searches.

Types of Search Engines..

Crawler based

- ❑ Crawler based search engines use a “spider” or “crawler” to search the internet.

- ❑ Web Crawlers refer to a class of software that digs through individual web pages, identifies the hyperlinks, pulls out keywords and then adds the pages to the search engine’s database.

- ❑ The crawlers could download a page and parse out all the links (by a scraper) to other pages (backlinks), building a list of new pages to visit. This created the ability to aggregate many URLs at a time, with the end goal of capturing every link on the WWW.

- ❑ Scrapers are programs that identify certain pieces of information from the web to be stored in databases.

- ❑ Examples of crawler search engines: Google and Yahoo

Types of Search Engines..

Crawler based

❑ **Advantages**:

- They contain a huge amount of pages.
- Ease of use.
- Familiarity. Most people who search the Internet are familiar with Google.

❑ **Disadvantages:**

- Sometimes, it's just too much information.
- It is easy to trick the crawler. Websites have hidden data that can be manipulated to make the page appear like it's something it's not.
- Page rank can be manipulated. There are ways to improve where your page appears on the list of results.

Types of Search Engines..

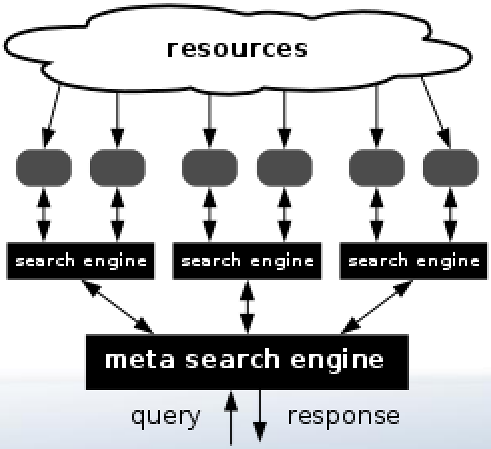
Hybrid & Mata Search Engine

❑ Hybrid Search Engine

- Hybrid search engines use both crawler based searches and directory searches to obtain their results
- Sometimes, you have a choice when you search whether to search the Web or a directory.
- Other times, you may receive both human powered results and crawler results for the same search.
  - ➢ In this case, the human results are usually listed first

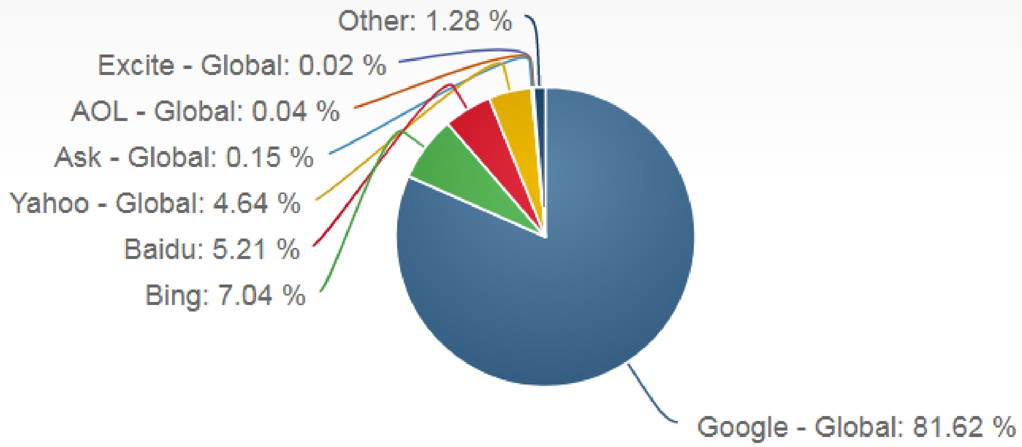
❑ Meta Search Engine

- Meta search engines are ones that search several other search engines at once and combines the results into one list. It works to
  - ➢ integrate the search results returned from all the search engines
  - ➢ eliminate the duplicates, and
  - ➢ implement additional features such as clustering by subjects within the search results
- Examples of meta search engines: Dogpile and Clusty are.

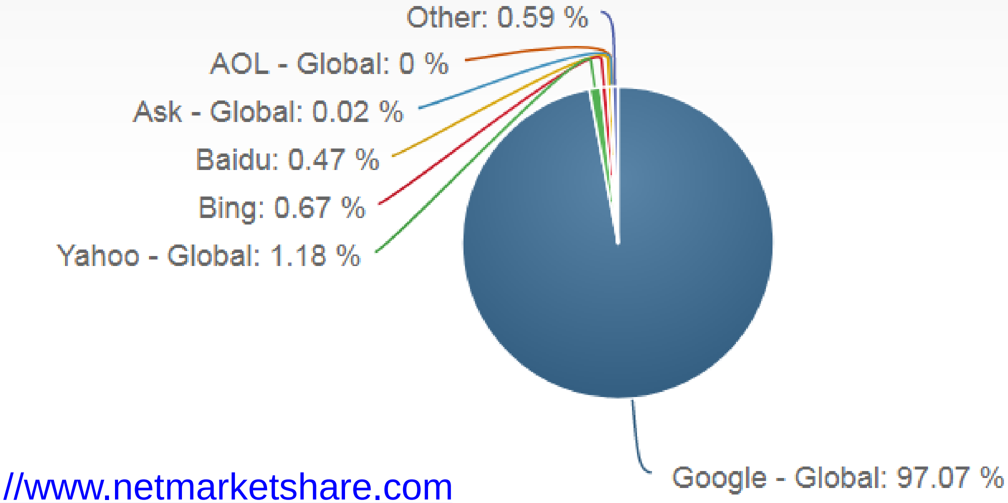


Stats of Search Engines

- ❑ Desktop Search Engine Market Share (Aug 2017)



- ❑ Mobile/Tablet Search Engine Market Share (Aug 2017)



Apple switches back to Google search results for iOS & Mac (Sep 2017)

Source: https://www.netmarketshare.com

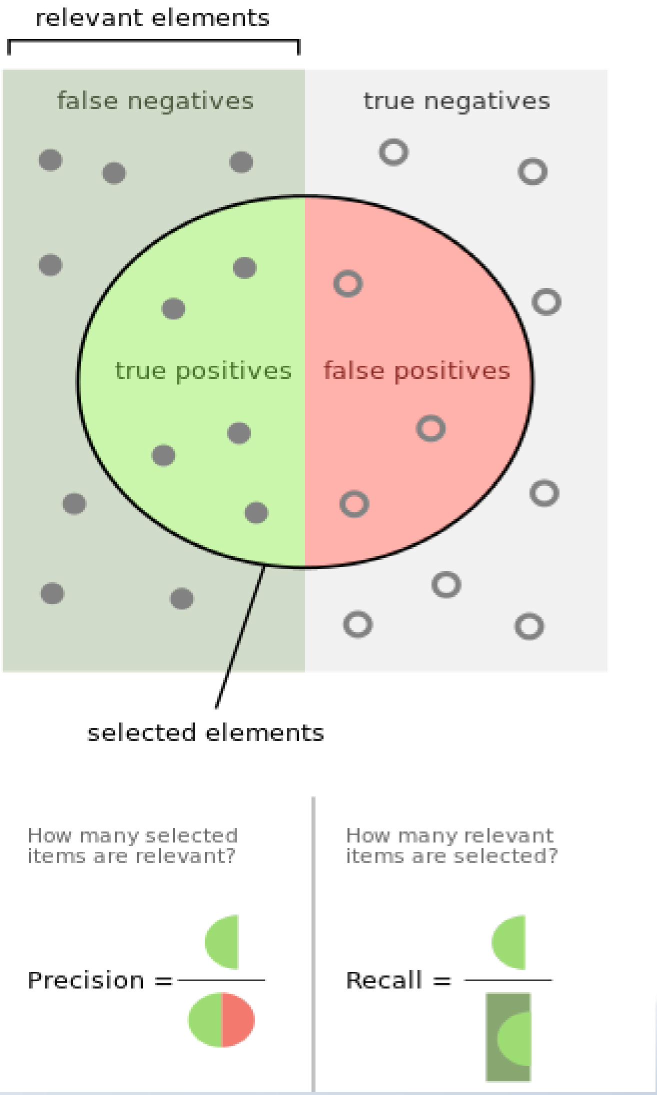
Retrieving relevant results

- ❑ The usefulness of a search engine depends on the relevance of the result set it gives back.

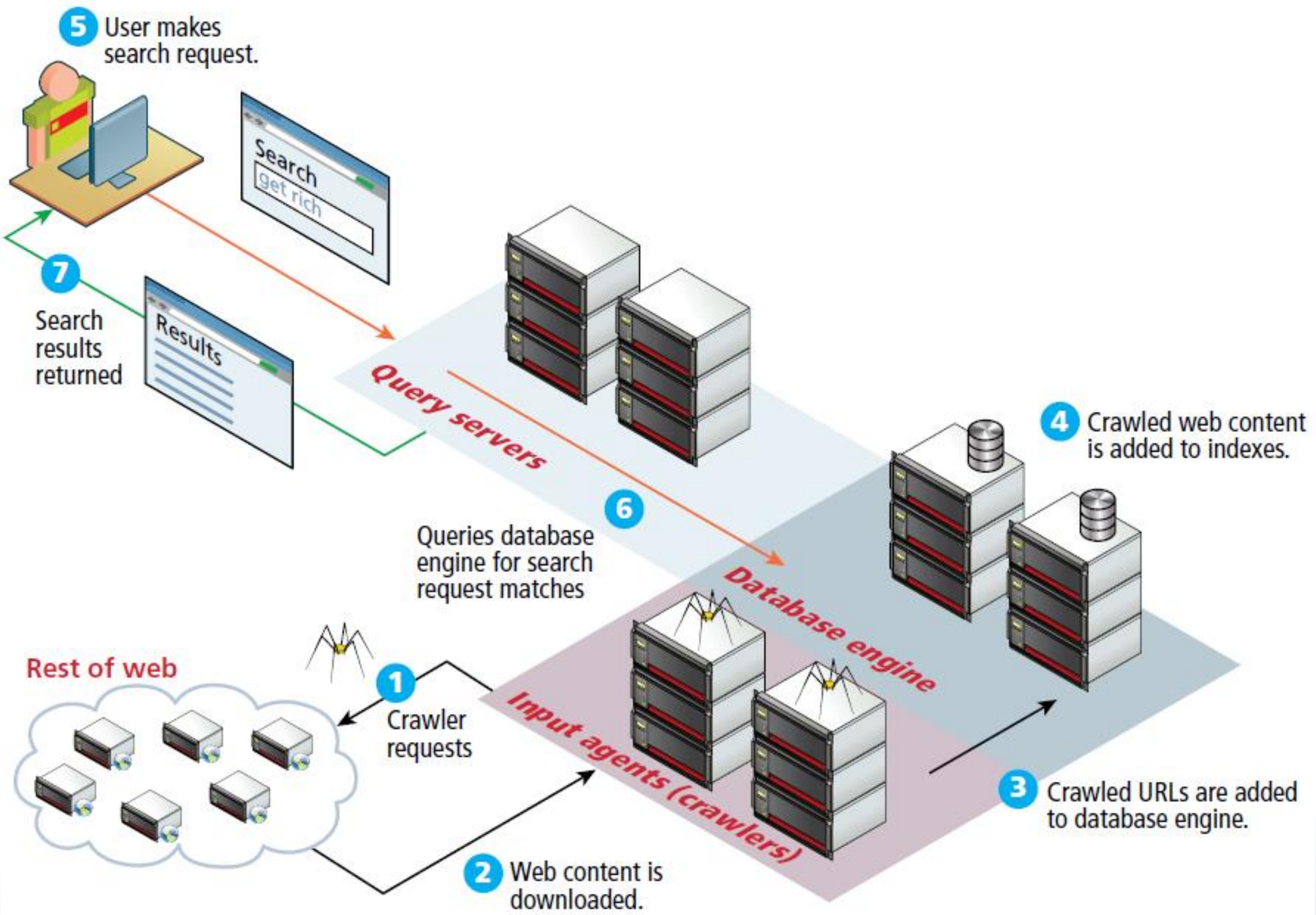
- ❑ Large search engines, such as Google, index trillions of web pages involving a comparable number of distinct terms, and answer billions of queries every day.

- ❑ While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or trustworthy than others.
  - >> Most search engines employ methods to rank the results to provide the "best" results first.

- ❑ **How** a search engine decides which pages are the best matches, and **what** order the results should be shown in, varies widely from one engine to another.
  - The methods also change over time as Internet usage changes and new techniques evolve.



Major components of a search engine



5 User makes search request.
Search
get rich
7
Search results returned
Results
Query servers
4 Crawled web content is added to indexes.
6
Queries database engine for search request matches
Database engine
Rest of web
1
Crawler requests
Input agents (crawlers)
3 Crawled URLs are added to database engine.
2 Web content is downloaded.

Major components of a search engine..

- The input agents refer mostly to web crawlers, which surf the WWW requesting and downloading web pages, all with the intent of identifying new URLs.
  - Additional input agents include URL submission systems, ratings systems, and administrative backends, but web crawlers are the most important.
- A database engine stores and manages the resulting URLs
- URLs are broken down into their components (domain, path, query string, fragment). This allows the engine to prioritize domains and URLs for more intelligent downloading (Indexing)
  - The Indexes speed up searches by storing B-trees or hashes in memory so queries can be executed quickly on those indexes to recover complete records.
- The query server handles requests from end users for particular queries in the database engine with the pages crawled and fully indexed,
- This final part it contains the algorithms, such as PageRank which determines what order to list the search results in and makes use of the database engine's indexes

Challenges Faced by Search Engines

- Spam
  - Users of web search engines tend to examine only the first page of search results. (for 85% of the queries only the first result screen is requested).
  - For commercially-oriented websites, it is their interest to be ranked within top 10 results
- Size of the Web
  - The Indexed Web contains at least 130 Trillion pages (Nov 2016).
- Concurrency
  - Many Web pages are updated frequently, which forces the search engine to revisit them periodically.
- Relevancy
  - Because the queries one can make are currently limited to searching for keywords, may result in many false positives

Challenges Faced by Search Engines..

- ❑ Problem with dynamically-generated Web pages
  - ▪ Dynamically generated web pages in response to a query are either left un-indexed by search engine spiders if indexed results in excessive results.
- ❑ Search engines can be tricked
  - ▪ To return pages, in favor of the trick makers, which contain little or no information about the matching phrases.
  - ▪ Making the more relevant Web pages pushed further down in the results list
- ❑ Indexing secured pages
  - ▪ Content hosted on HTTPS URLs pose a challenge for crawlers which either can't browse the content for technical reasons or won't index it for privacy reasons.
  - ▪ Google indexes secure pages.

To read more about this topic: "*Challenges in Web Search Engines*" posted in the blackboard

How Web Search Engines Work?

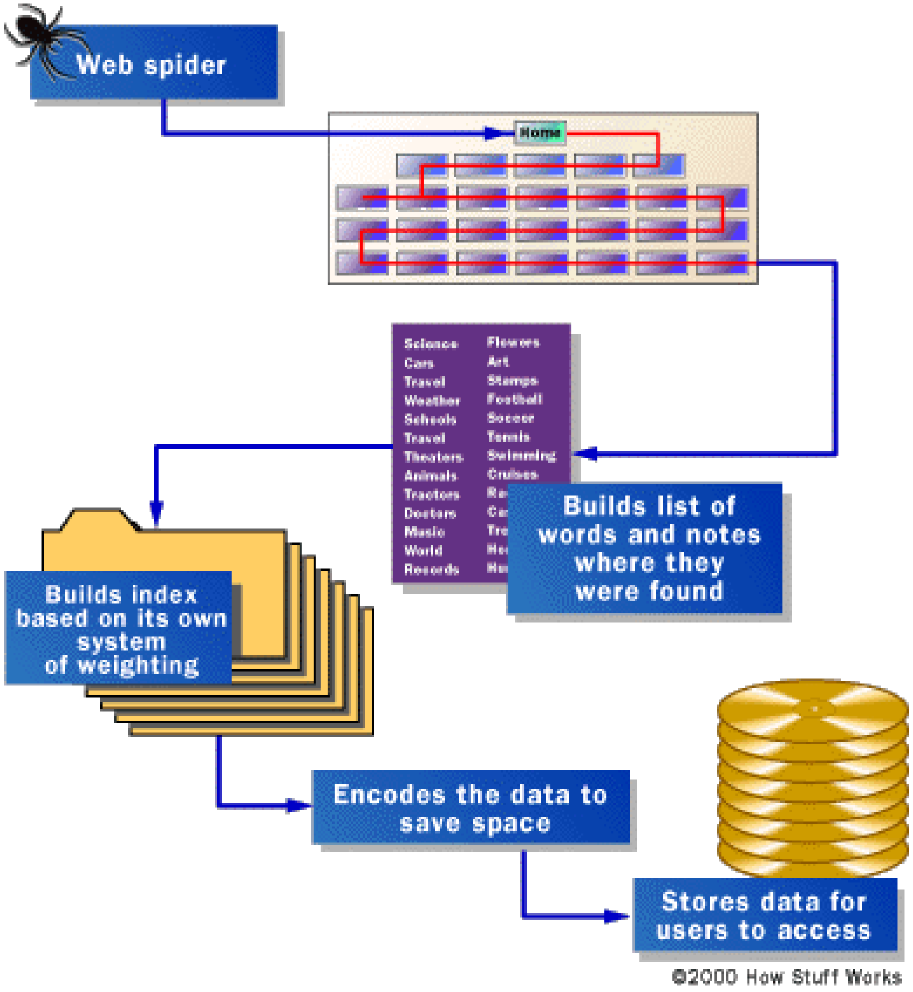
How Web Search Engines Work

Before a search engine can tell you where a file or document is, it must be found

- It 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.

❑ Web crawling

- Navigates the web, retrieves web pages that satisfy certain criteria
- Starts with a popular Web site containing lots of links, such as Yahoo then continues until it finds a logical stop, e.g. a dead end with no external links or reaching a number of levels inside the Web site’s structure

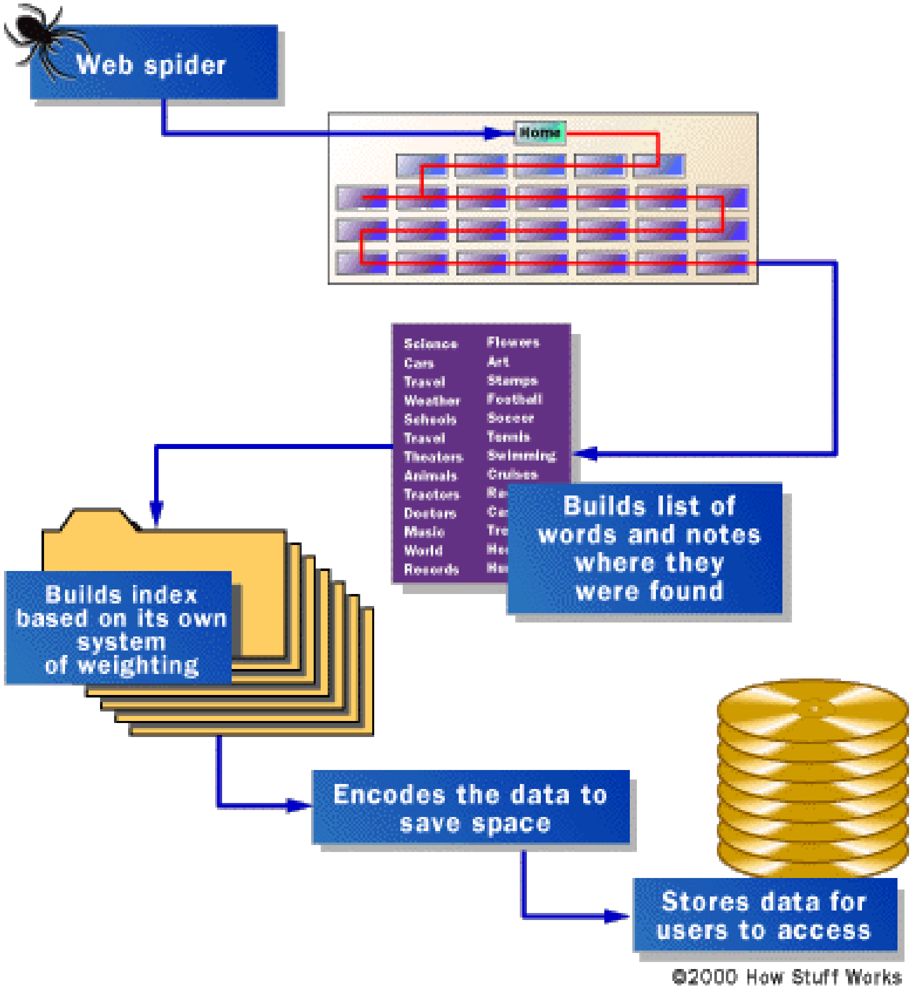

Web spider
Home
Science
Cars
Travel
Weather
Schools
Travel
Theaters
Animals
Tractors
Doctors
Music
World
Records
Flowers
Art
Stamps
Football
Soccer
Tennis
Swimming
Cruises
Builds list of words and notes where they were found
Builds index based on its own system of weighting
Encodes the data to save space
Stores data for users to access

©2000 How Stuff Works

Source: http://computer.howstuffworks.com

How Web Search Engines Work..

- Indexing
  - Pages are analyzed and a list of words and notes (extracted from titles, headings and other special meta tags) are stored in indexes to facilitate quick information retrieval
- Searching
  - A search engine stores information about web pages in a database
  - Records that best match the search criteria are returned to the user



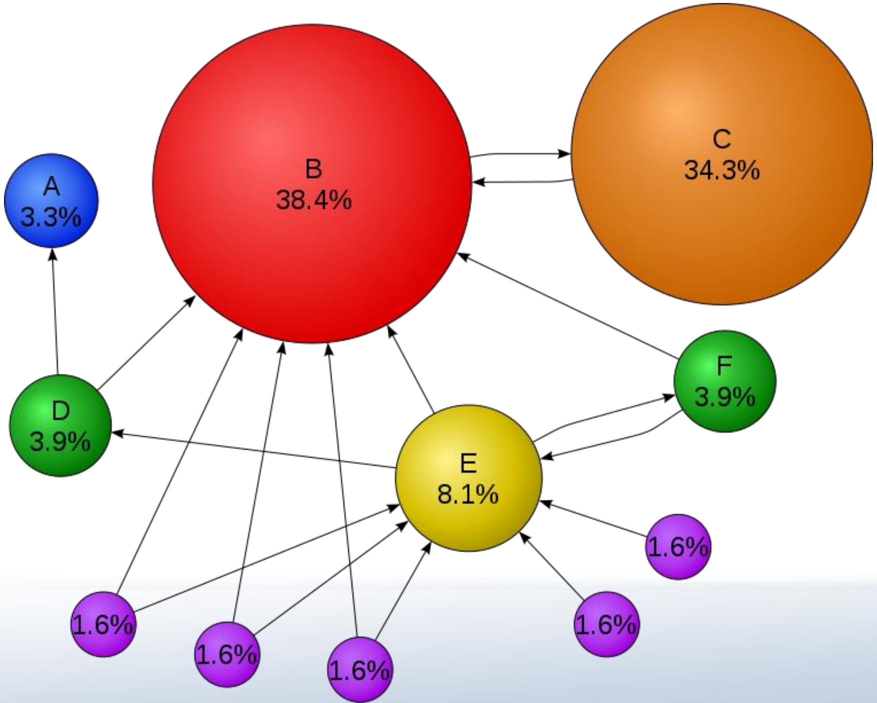
Source: http://computer.howstuffworks.com

PageRank algorithm

- PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results
  - by measuring the importance of website pages
- 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

  - Webgraph: all Web pages as nodes and hyperlinks as edges

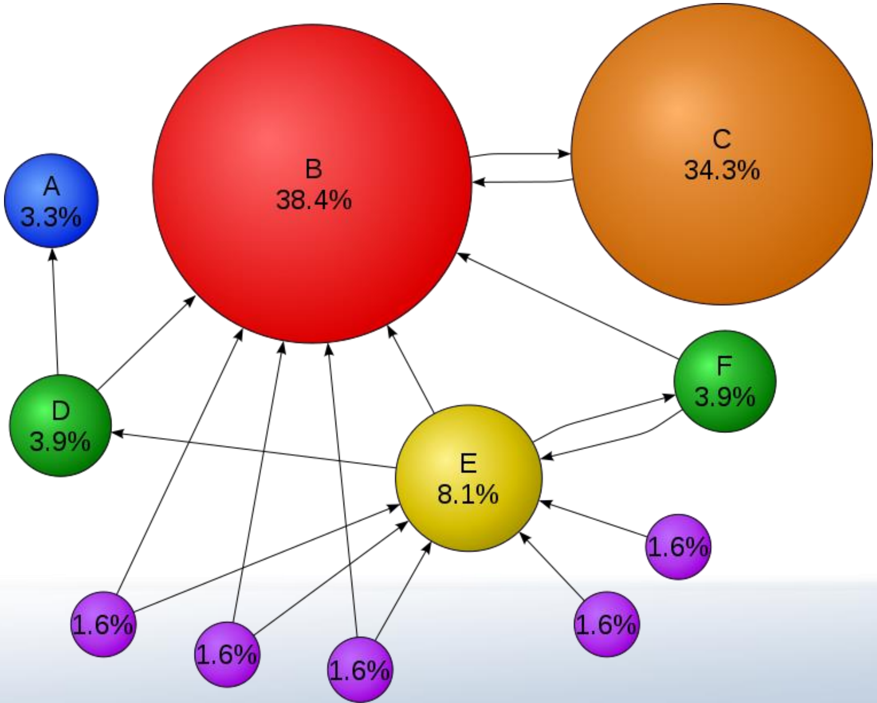
  - PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.



Backend Part..
PageRank algorithm

- The underlying assumption is that more important websites are likely to receive more links from other websites.
- A hyperlink to a page counts as a vote of support.
  - >> A page that is linked to by many pages with high PageRank receives a high rank itself.

  - Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value.



Google Search Engine

- ❑ Backend Part
  - Crawling & indexing (via Google webmaster tools)
  - Algorithms (Uses search behavior)
  - Fighting Spam (Flood Protection)

- ❑ Frontend Part
  - Title (Keywords play the key role)
  - URL (keywords play the key role)
  - Meta Description (Keywords play the key role)
  - Others..

Backend Part
Crawling and Indexing

- Find information by crawling
  - Crawler is called “Googlebot”
    - The Googlebot is simply the search bot software that Google sends out to collect information about documents on the web to add to Google’s searchable index.
  - Crawlers look at webpages and follow links on these pages
  - Crawling is the process where the Googlebot goes around from website to website, finding new and updated information to report back to Google.
  - The crawl process begins with a list of web addresses from past crawls and sitemaps provided by website owners
    - > Crawlers pays special attention to new sites, changes to existing sites and dead links

Backend Part..
Crawling and Indexing

- ❑ Organize information by indexing
  - Indexing is the processing of the information gathered by the Googlebot from its crawling activities.
  - Once documents are processed, they are added to Google's searchable index if they are determined to be quality content.
  - During indexing, the Googlebot processes the words on a page and where their locations.
  - >> Information such as title tags and attributes are also analyzed during indexing.
  - The indexer sorts every word on every page and stores the resulting index of words in a huge database

- ❑ So how does the Googlebot find new content on the web such as new websites, blogs, pages, etc.?
  - It starts with web pages captured during previous crawl processes and adds in sitemap data provided by webmasters.
  - As it browses web pages previously crawled, it will detect links upon those pages to add to the list of pages to be crawled.
  - If you want more details, you can read about them in Webmaster Tools Help.

Backend Part.. Algorithms

- ❑ Algorithms are the computer processes and formula that take your questions and turn them into answers
  - The algorithms compare your search query to the index and recommends the documents that it considers most relevant

- ❑ Google has to ensure the results it serves on its SERPs (search engine results pages) are of the highest quality possible.
  - Google's algorithms rely on more than 200 signals to guess what you are looking for.
    - ➢ These signals include things like your web history, location, PageRanke etc.

- ❑ Here are some of the ways Google uses Search algorithms to return useful information:
  - Analyzing your words
  - Matching your search
  - Ranking useful pages
  - Considering context

Backend Part..
Fighting Spam

- ❑ Spam sites attempt to game their way to be the top of search results through techniques like repeating keywords over and over, buying links that pass PageRank , etc.

- ❑ Google's algorithms can detect vast majority of spam and demote it automatically

- ❑ Spam Types:
  - Cloaking and/or sneaky redirects
  - Hacked site
  - Hidden text and/or keyword stuffing
  - Thin content with little or no added value
  - and many others...

Search Engine Optimization

Search engine optimization (SEO)

- SEO is the process undertaken by a webmaster to make a website more appealing to search engines
  - >> to increases its ranking in search results for terms the webmaster is interested in targeting

- Sites that appear high in a search engine's rankings are more likely to attract new potential customers, and therefore contribute to the core business of the site owner.

- SEO techniques can be broken down into two major categories:
  - White-hat SEO that tries to honestly and ethically improve your site for search engines, and
  - Black-hat SEO that tries to game the results in your favor.

White-hat SEO

- White-hat SEO – also called *Ethical* SEO- is a practice used to improve search performance that is in line with terms and conditions of a search engine. e.g.,:
  - Offering quality content
  - Using proper metadata and effective keywords
  - Having incoming links from relevant high-quality pages

- White-hat SEO is more frequently used by those who intend to make a long-term investment on their website.

- Although the white-hat SEO techniques are not particularly challenging, yet many websites do not apply these simple principles.
  - These techniques include page title, meta tags, URLs, site design, anchor text, images, and content.

White-hat SEO
Title page

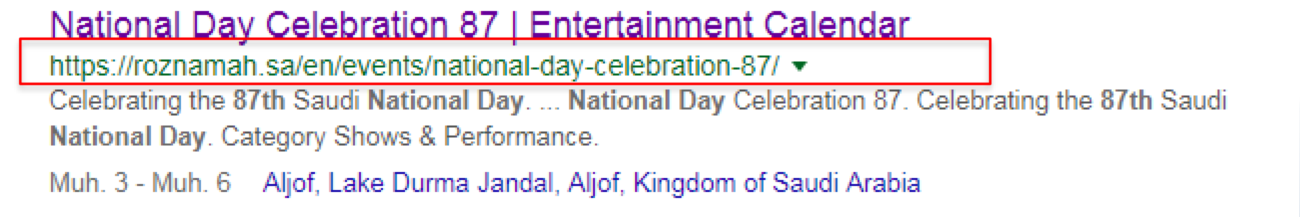
- ❑ The <title> tag in the <head> portion of your page is the single most important tag to optimize for search engines.
  - It defines the title of your web page, and is typically the text that appears as a blue link on search engine results pages
  - Several of the words in the title match the user's query- this causes them to be in bolded in the SERPs (Search Engine Results Page)
  - Use powerful CTR (Click-through Rate) keywords.

- ❑ CTR (Click-through Rate) is a measure of the percentage of clicks advertisers receive out of total ad impressions.
  - "Impressions" refers to the number of times your ad is viewed.
  - *Click Through Rate = (Total Clicks on Ad) / (Total Impressions)*
    - ➢ If your ad has a log of impressions but no clicks, you will have a low CTR, which generally reduces the effectiveness of your campaigns.

White-hat SEO..
URLs

- Search engines must download and save URLs since they identify the link to the resource. URLs can take a variety of forms, some of which are better for SEO purposes.
- In the example below, the landing page in the (roznamah.sa) example leverages sematic markup for breadcrumb navigation
  - As a result, the display URL provides valuable context and additional clickable navigation options

- Remember that: a website is not like a house – <u>people do not only come in through the front door.</u>
  - When it comes to huge websites, breadcrumbs is a great way to help users <u>identify where they are located</u>.
  - When developing URLs, pay attention to areas such as the folder, structure, word choice, query match and breadcrumb navigation.
  - Ensure that you website follows a well-organized folder structure and hierarchy that leverage standard navigational tools such as breadcrumbs.



National Day Celebration 87 | Entertainment Calendar
https://roznamah.sa/en/events/national-day-celebration-87/ ▾
Celebrating the **87th** Saudi **National Day**. ... **National Day** Celebration 87. Celebrating the **87th** Saudi **National Day**. Category Shows & Performance.
Muh. 3 - Muh. 6 Aljof, Lake Durma Jandal, Aljof, Kingdom of Saudi Arabia

White-hat SEO..
Meta-Tags

- <meta> tags can be used to define meta information, robots directives, HTTP redirects, and more.
- Early search engines made significant use of meta tags, since indexing meta tag was less data-intensive than trying to index entire pages.

- The *keywords meta tag* allowed a site to summarize its own keywords, which search engines could then use in their primitive indexes.
  - Unfortunately, since the tags are *not visible* to users, the content of the meta tags *might not reflect the actual content* of the pages.
  - *Keywords* are mostly ignored nowadays, since search engines build their own indexes for your site
    - When using keywords → always choose the most significant and unique words

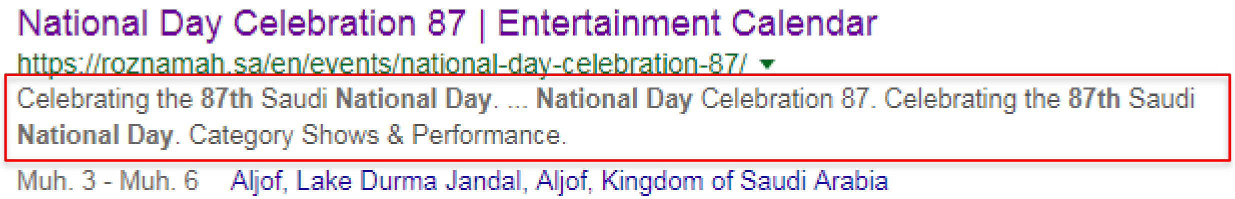
      >> Example: CCSE, College of Computer Science and Engineering, ccse, etc.

- Other meta tags are still widely used, and used by search engines.

White-hat SEO..

Description

- Meta descriptions of pages (when available) give users a clear idea of the URL's content before accessing the page.
  - Meta descriptions as free advertising

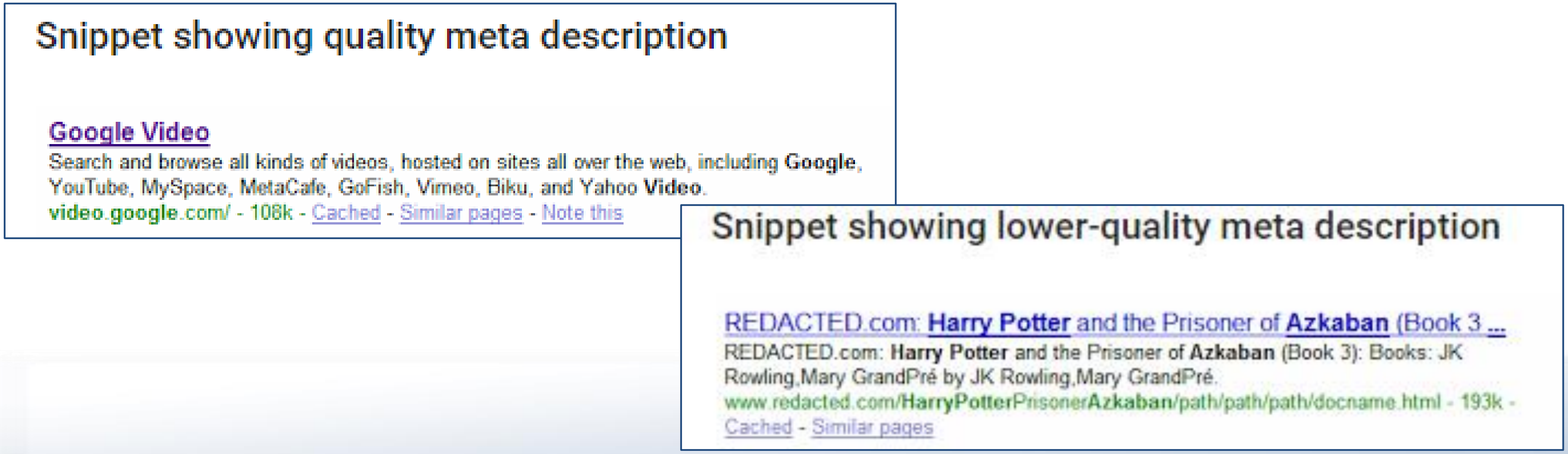


National Day Celebration 87 | Entertainment Calendar

https://roznamah.sa/en/events/national-day-celebration-87/ ▾

Celebrating the **87th** Saudi **National Day**. ... **National Day** Celebration 87. Celebrating the **87th** Saudi **National Day**. Category Shows & Performance.

Muh. 3 - Muh. 6   Aljof, Lake Durma Jandal, Aljof, Kingdom of Saudi Arabia



Snippet showing quality meta description

Google Video

Search and browse all kinds of videos, hosted on sites all over the web, including **Google**, YouTube, MySpace, MetaCafe, GoFish, Vimeo, Biku, and Yahoo **Video**.

video.google.com/ - 108k - Cached - Similar pages - Note this

Snippet showing lower-quality meta description

REDACTED.com: **Harry Potter** and the Prisoner of **Azkaban** (Book 3 ...

REDACTED.com: **Harry Potter** and the Prisoner of **Azkaban** (Book 3): Books: JK Rowling,Mary GrandPré by JK Rowling,Mary GrandPré.

www.redacted.com/**HarryPotter**Prisoner**Azkaban**/path/path/path/docname.html - 193k - Cached - Similar pages

White-hat SEO..
Site Design

- The design and layout of your site has a huge impact on your visibility to search engines.
    - Any sites that rely heavily on JavaScript or Flash for their *content* and *navigation* will suffer from poor indexing. >> because crawlers do not interpret scripts; they simply download and scrape HTML.

- Other aspects of site design that can impact your site's visibility include its internal link structure and navigation.

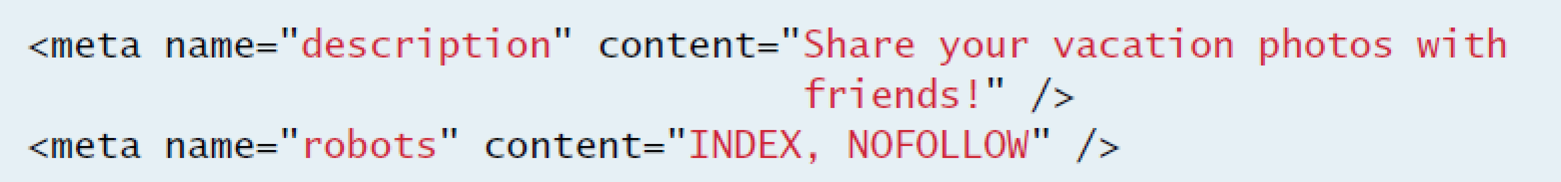
- Search engines can perform a sort of PageRank analysis of our site structure and determine which pages are more important.
    - Pages that are important are ones that contain many links, while less important pages will only have one or two links.
    - Links in a website can be categorized as: navigation, recurring, and ad hoc.

White-hat SEO..

Robots

- We can control some behavior of search engines through meta tags with the name attribute set to robots.

- The content for such tags are a comma-separated list of INDEX, NOINDEX, FOLLOW, NOFOLLOW.
  - Tags with a value of INDEX tell the search engine to index this page.
  - NOINDEX, advises the search robot to not index this page.
  - FOLLOW and NOFOLLOW values, which tell the search engine whether to scan your page for links and include them in calculating PageRank.



```
<meta name="description" content="Share your vacation photos with
                                  friends!" />
<meta name="robots" content="INDEX, NOFOLLOW" />
```

Black-hat SEO

- Black hat SEO (spamdexing) – techniques that are used to increase a page's rank by misleading search engines
  - These techniques are constantly evolving as people try to exploit weaknesses in the secret algorithms.
  - Black hat SEO is more frequently used by those who are looking for a quick financial return on their Web site, rather than a long-term investment on their Web site.

Example of black-hat optimization techniques:

- Content spamming is any technique that uses the content of a website to try and manipulate search engine results.
  - Keyword stuffing is a technique whereby you purposely add keywords into the site in a most unnatural way with the intention of increasing the affiliation between certain key terms and your URL.
    - Keyword stuffing can occur in the body of a page, in the navigation, in the URL, in the title, in meta tags, and even in the anchor text.
  - Hidden Content- rather than remove the unrelated keywords, some CSS tricks are used for moving and hiding the useless keywords.
  - Paid Links-

Black-hat SEO..

- ❑ Link Spam- since links, and backlinks in particular, are so important to PageRank>> many bad SEO techniques related to links
  - Hidden links are as straightforward as hidden content.
  - A link farm is a set of websites that all interlink each other. The intent of these farms is to share any incoming PageRank to any one site with all the sites that are members of the link farm.

- ❑ Google bombing is the technique attempts to trick the search engine to promote a certain page (creating large numbers of links, that cause a web page to have a high ranking for searches on unrelated or off topic keyword phrases)

Invisible Web

Invisible Web

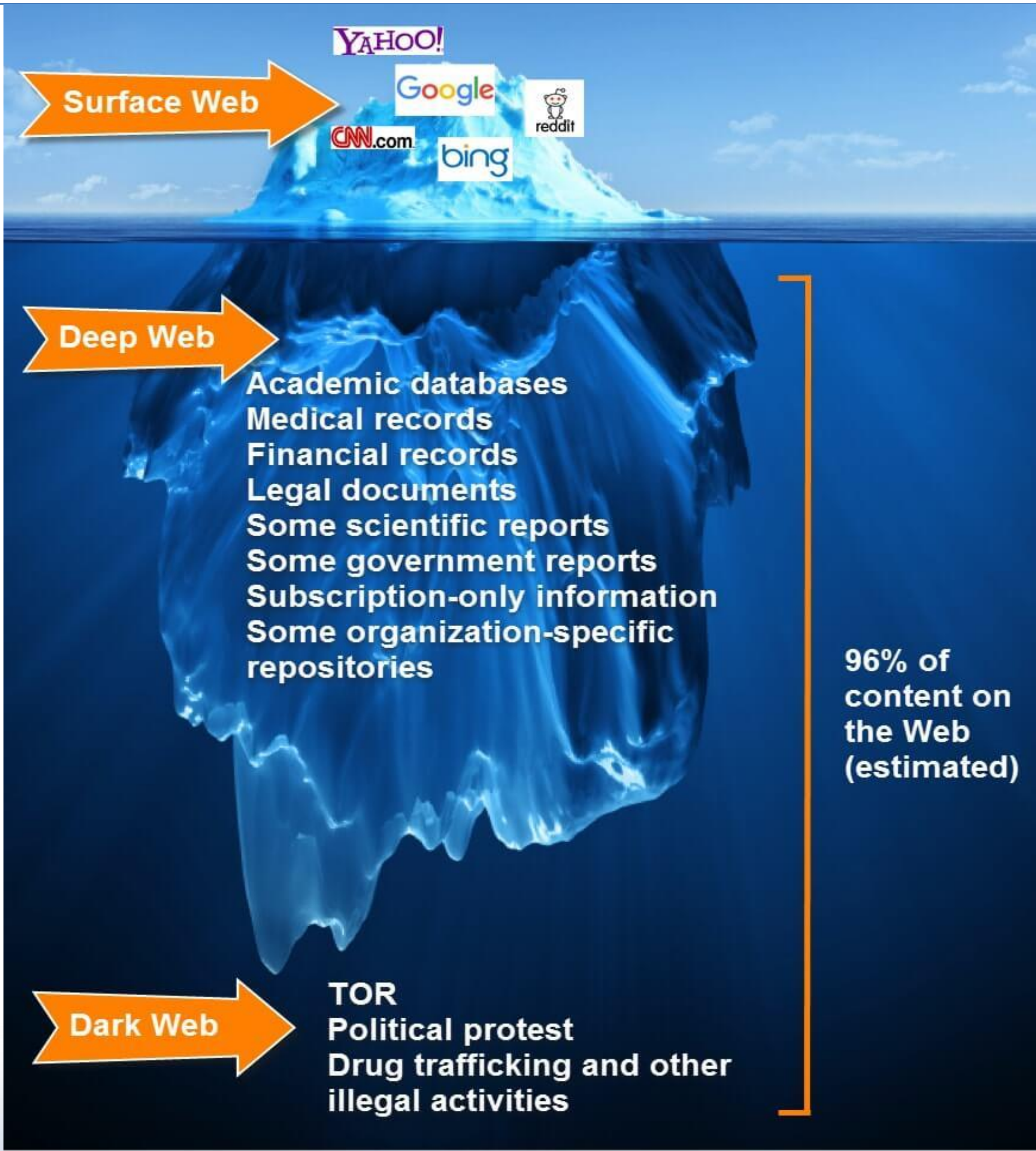
- Invisible Web : is the term used to describe all of the information available on the Web that cannot be found by using general-purpose search engines.
  - Also called deep Web, Deepnet, dark Web, hidden Web

- The Invisible Web is estimated to be X times larger than the Web content found with general search engine queries.

- Some Fast Facts About the Invisible Web
  - 2003 - Surface Web: 167 terabytes vs. Deep Web: 91,850 terabytes
  - 2006 - Academic Invisible Web between 20-100 billion documents
  - 2012- 11+ billion static pages are hidden from the public as well as 450+ billion database-driven are completely invisible to Google
  - 2013 - "Trillions + pages of information" that current search engines cannot find-

Invisible Web..

- ❑ The fact that search engines only search a very small portion of the web make the Invisible Web a very attractive resource.
  - There's a lot more information out there than we could ever imagine.



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)

TOR
Political protest
Drug trafficking and other illegal activities

Invisible Web..

- ❑ Why Is It Called "The Invisible Web"?
  - The “spiders” can record the address, but can't access anything about the information the page contains.
  - For instance, university library sites that require passwords to access their information will not be included in search engine results, as well as script-based pages that are not easily read by search engine spiders

- ❑ Why Is The Invisible Web Important?
  - Many users believe it could be easier to just stick with what can be found with Google or Yahoo.
  - However, it's not always easy to find what you're looking for with a search engine, especially if you're looking for something a bit complicated or obscure.
    - ➢ Example: student who is looking for a novel research idea!

Invisible web..
Characteristics of Invisible Web Content

Why Search Engines Can’t Find this Information:

- ❑ Content found in databases
  - Database content that is dynamically generated as the result of a query cannot be found by general-purpose search engines.
  - Example: ERIC database, Library catalogs.
- ❑ Subscription database content
  - Fee-based database content is only accessible to those who have subscribed.
  - Many libraries offer their members free access to subscription databases.
  - Examples: EBSCOhost databases, LexisNexis Academic.
- ❑ Real time content
  - Information about events currently taking place may not yet be indexed by general-purpose search engines.
- ❑ Sites requiring login authorization
  - These sites require users to login or identify themselves as having the right to access and use content. Examples: Blackboard, membership sites.

Invisible web..
Characteristics of Invisible Web Content

❑ Sites that are not linked to by other sites

- Search engines index websites by following links from one website to another, if there aren't any links to a site it might not be found or included.

❑ Dark Web

- Sometimes called Deep Web, Tor, or Hidden Wiki, this Dark Web is not to be confused with the Invisible Web as presented in this website.
- The Dark Web is usually used by those who seek anonymity for their web activities.

❑ Many others…