CS317
Information Retrieval
Week 09

Muhammad Rafi
April 01, 2021

Web Search Basics

Chapter No. 19

Today's Agenda

- Web Search basic
- Background & History
- Web Characteristics
- The search user experience
- Economic Models
- Top Ten Search Engines (2014)
- Index Estimates
- Duplicate detection
- Conclusion

Web Search – Client Server

- Client
  - The client – generally a browser, an application within a graphical user environment
- Server
  - The server communicates with the client via a protocol HTTP
  - It is lightweight and simple, asynchronously carrying a variety of payloads (text, images and – over time – richer media such as audio and video files) encoded in a simple markup language called HTML (for hypertext markup language)

Web Search – Client Server

- Browser
  - The first browser was developed by Tim Berners-Lee in 1990- very limited functionality
  - Mosaic was first GUI based browser in 1993 by Marc Andreesen
  - Marc started Netscape in 1994 and launch Netscape Navigator
  - Microsoft started IE in 1995 for free. 95% market share in 2002
  - Marc started Mozilla foundation and started Firefox in 2004 reached 23% market share in 2011

Web Search – Client Server

- HTTP
  - HTTP is an application protocol for distributed, collaborative, and hypermedia information systems. {Domain Sharding, Blocking, multiplexing)}
  - HTTP/2, was standardized in 2015, and is now supported by major web servers and browsers.
  - HTTP Header contains a lot of fields for effective transfer of information. HTTP Header injection?
  - HTTP/2 was standardized in 2018
  - Persistence connection, Session State, Authentication mechanism, server push etc

Web Search – Client Server

- HTTP Status Code
  - Informational 1XX
  - Successful 2XX
  - Redirection 3XX
  - Client Error 4XX
  - Server Error 5XX

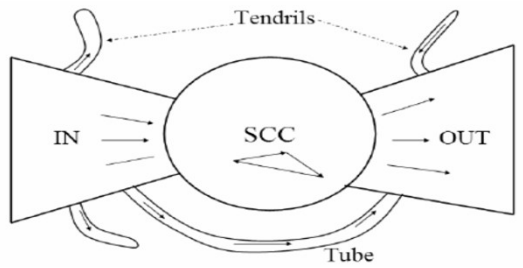
Web Search – Client Server

- HTML
  - HTML 2.0 -1995; HTML 3.0 1997; HTML 4.0 1997
  - HTML 5.0 2014; XHTML vs. XML
- Server Side Scripting
  - A number of server side scripting available.
- Client Side Scripting
  - Generally UI and interaction with local machine, mostly Java Script
- Cascading Style Sheet (CSS)
  - CSS is a language that describes the style of an HTML document.

Web as a Graph

- The web can be viewed as a graph of connected web-pages.
- Anchor-text and link to a web resource is used to link these pages.
- We refer to the hyperlinks into a page as in-links and those out of a page as out-links.
- Power law states that these in-link and outline are govern with a power functions. In-link proportional to $1/i^a$
- Bowtie Model

Web as a Graph



► **Figure 19.4** The bowtie structure of the Web. Here we show one tube and three tendrils.

Web as a Graph

- Spam, (in the context of web search) is the manipulation of web page content for the purpose of appearing high up in search results for selected keywords.
- A major concern of web 2.0 and 3.0
- Given that spamming is inherently an economically motivated activity, there has sprung around it an industry of Search Engine Optimizers, or SEOs to provide consultancy services for clients who seek to have their web pages rank highly on selected keywords.
- Adversarial Information Retrieval, is the area of study making a balance between SEO and SE.

Advertising as an Economic Model

- The ad-words are located in query and the advertisements are push to make some business value.
- Typically these advertisements are priced on a cost per mil (CPM) basis: the CPM cost to the company of having its banner advertisement displayed 1000 times.
- Cost Per Click (CPC), Cost per Users Action CPA, etc
- Understanding how search engines do the ranking and how to allocate marketing campaign budgets to different keywords and to different sponsored search engines for benefitting maximum, has become a profession known as Search Engine Marketing (SEM).
- Click Spam – a phenomena for advertising competition.

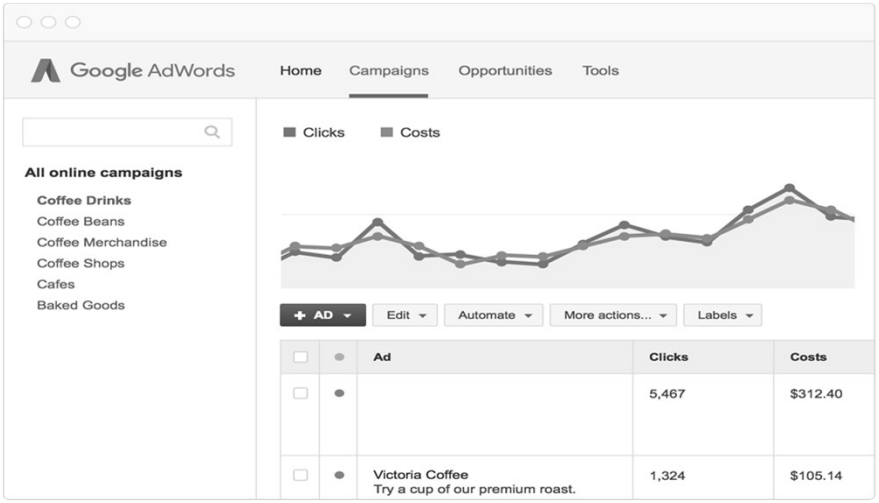
Advertising as an Economic Model

- Pure Vs. Sponsored search
- Combing Pure vs. sponsored search
- Local vs. Global Search vs. Social Media search
- Blog vs Products reviews

Web Economic Model

- Advertisement Model for Revenue
- Unit of Measurement
  - CPM, CPC, CPI, CPD,
- Complex Advertisement Models
  - AdWords
    - Ads
    - Search terms
    - Daily budget

AdWords



Sec. 19.4.1

User Needs

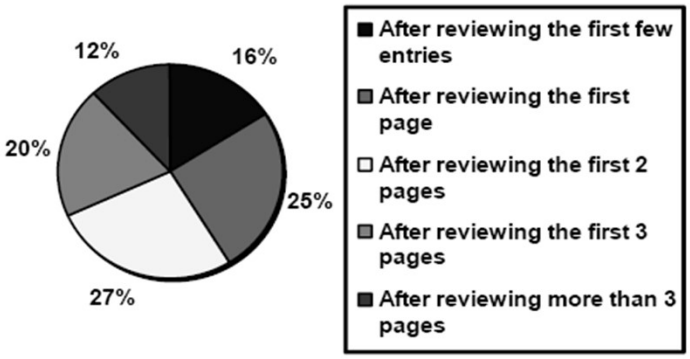
- Need [Brod02, RL04]
  - **Informational** – want to learn about something (~40% / 65%)
    `Low hemoglobin`
  - **Navigational** – want to go to that page (~25% / 15%)
    `United Airlines`
  - **Transactional** – want to do something (web-mediated) (~35% / 20%)
    - Access a service
    - Downloads
    - Shop

    `Mars surface images`
    `Canon S410`
    `Seattle weather`
  - **Gray areas**
    - Find a good hub
    - Exploratory search "see what's there"

    `Car rental Brazil`

How far do people look for results?

"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"

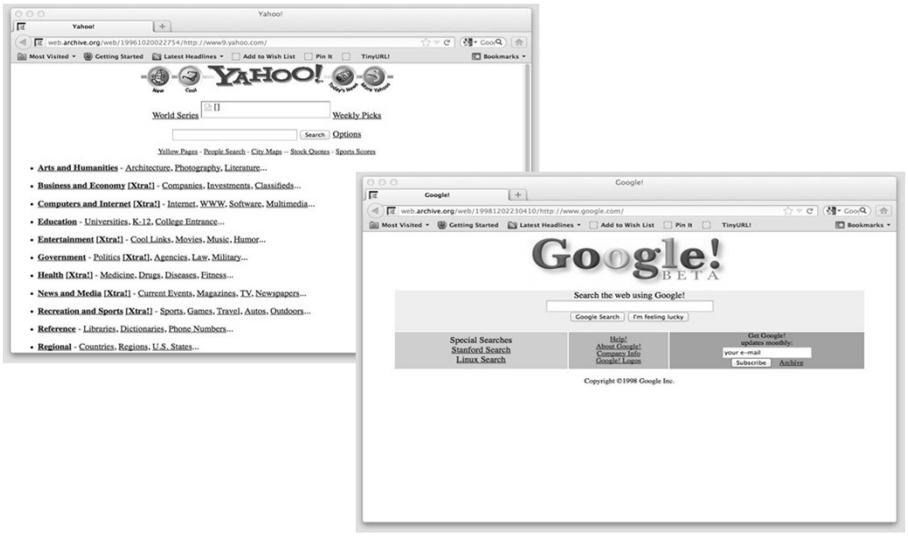


(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Web Information Discovery

- Directories
  - Taxonomies populated with web pages in categories, such as Yahoo!
  - The user to browse through a hierarchical tree of category labels.
- Search Engines
  - Full-text index search engines such as Altavista, Excite and Infoseek
  - The user with a keyword search interface supported by inverted indexes and ranking mechanisms.

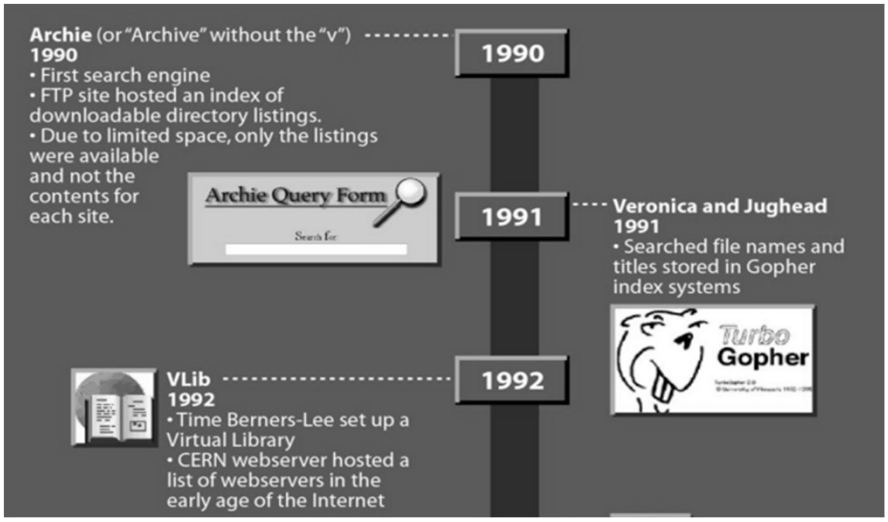
Web Information Discovery



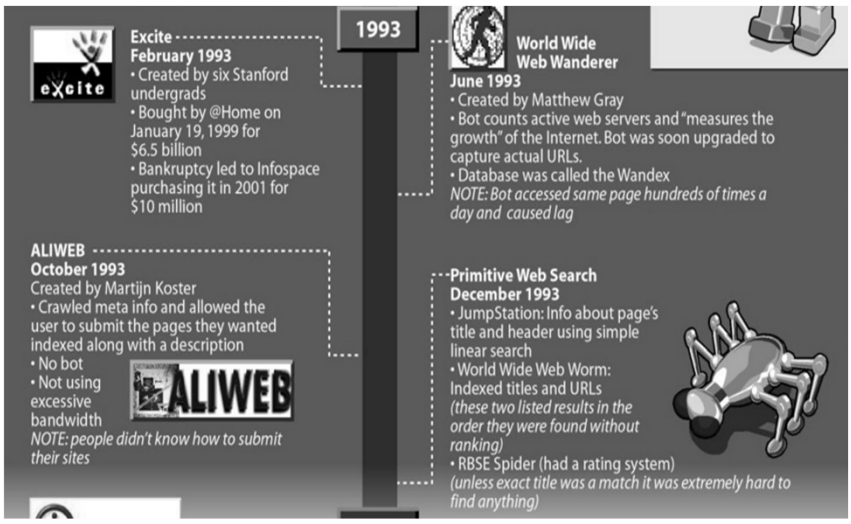
Directories Vs. Search Engines

- **A directory allows you to explore and get what you want eventually.**
- **Use a directory to find cooking-related websites.**
- **Use a directory to find travel guides in a country.**
- **A search engine brings you to the exact page on the words or phrases you are looking for.**
- **Use a search engine to find a specific recipe, by providing the name of the ingredients.**
- **Use a search engine to find the transport trains schedule in Germany**

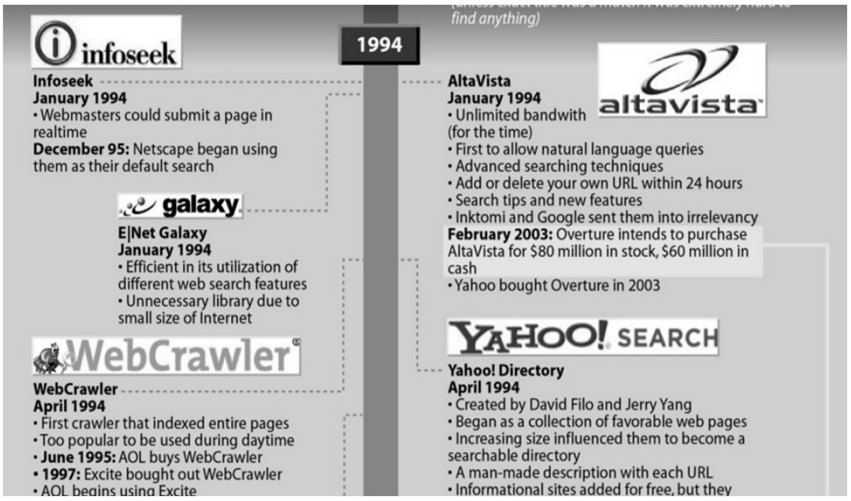
Search Engine History



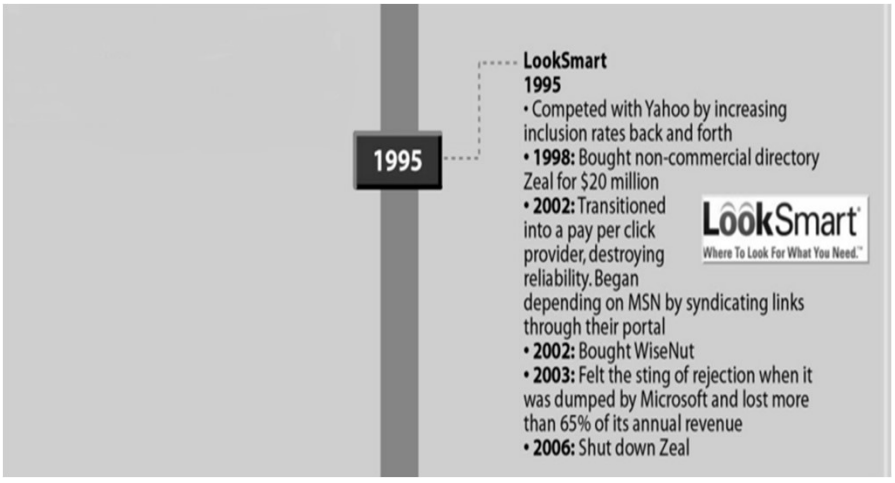
Search Engine History



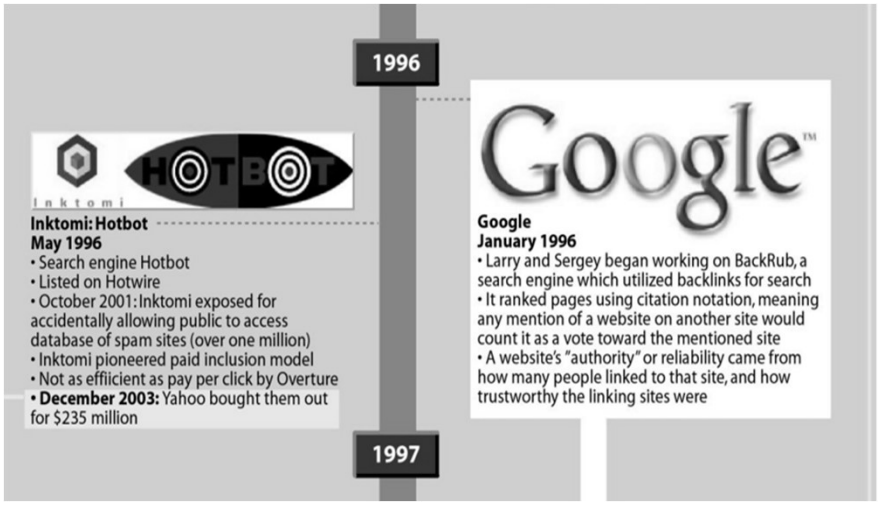
Search Engine History



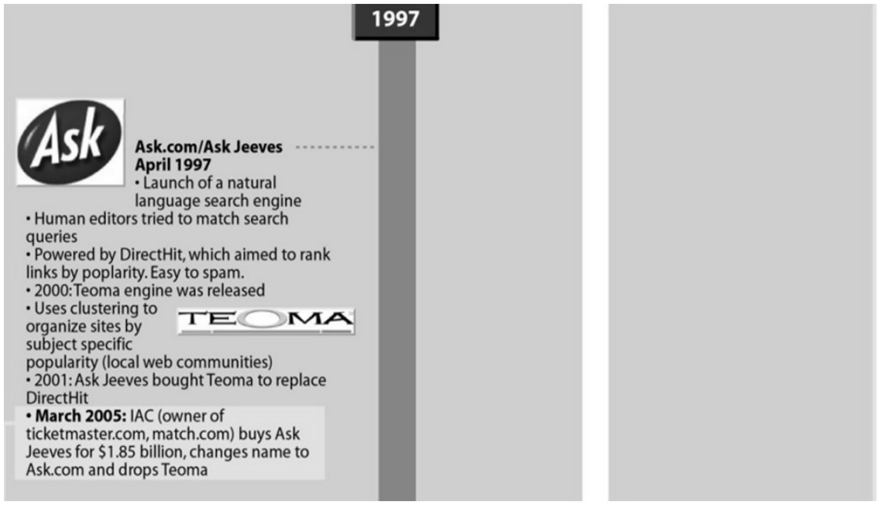
Search Engine History



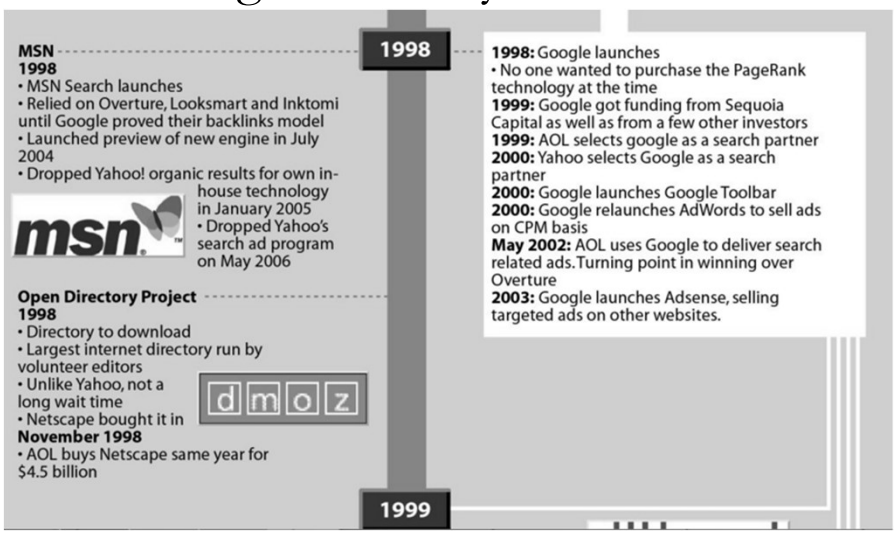
Search Engine History



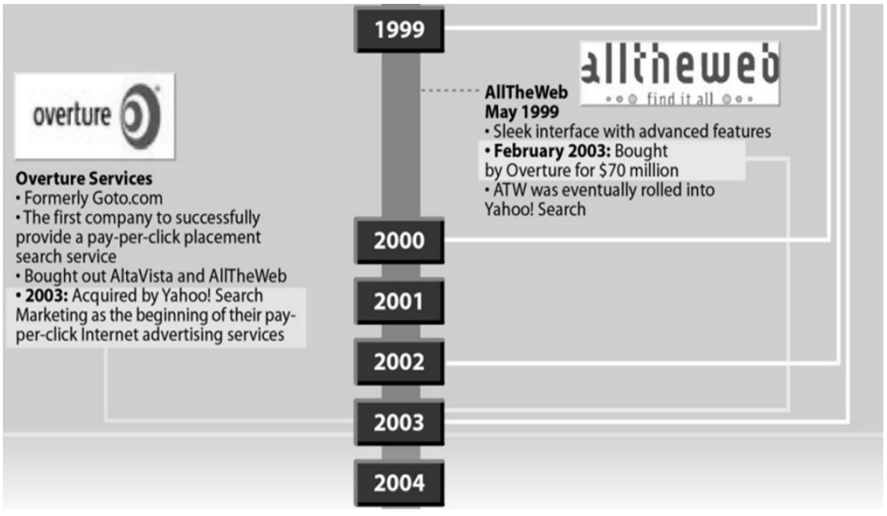
Search Engine History



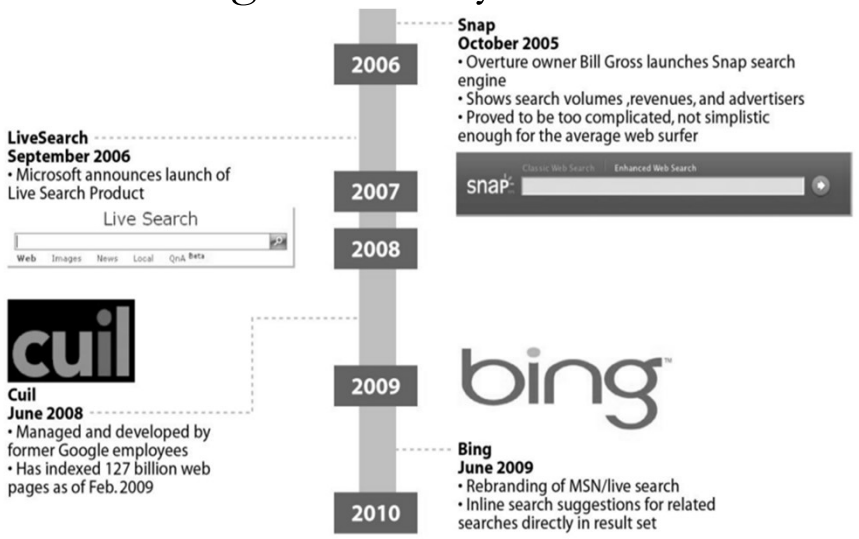
Search Engine History



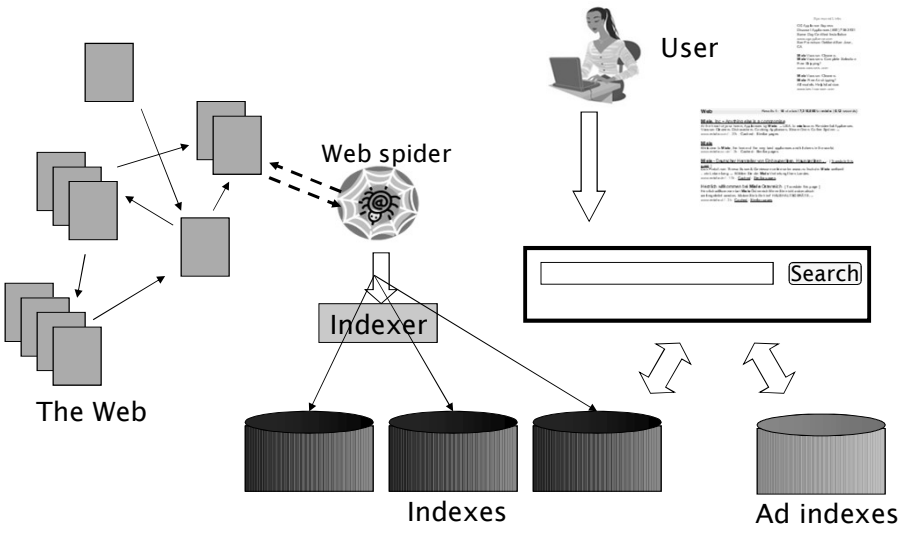
Search Engine History



Search Engine History



Web Search



Web Characteristics

- Web User Interaction
- Web as a Graph
- Web Spam

Top 10 Search Engines

- Google - Offering everything from image searches, map searches, news searches, etc. With impressive keyword relevancy and a continuously improving search algorithm, it's easy to see why Google is still the reigning champ.
- Mahalo - Mahalo is a unique 'human-powered' search engine that employs a group of editors to manually sift and organize thousands of pieces of content.

Top 10 Search Engines

- Yahoo - While Yahoo has been suffering as of late, it's still a classic and a popular search engine.
- Bing - The Microsoft powered search engine prides itself on being a "decision engine" by offering search suggestions on the side column and providing extra search options.

Top 10 Search Engines

- Ask - Clean layout and handy results grouping.
- AOL Search - AOL continues to be used, primarily by people who still use AOL. They're out there somewhere.
- Blekko - Blekko's clean, minimalist layout is easy to navigate, and /tags allow for grouping searches.

Top 10 Search Engines

- DogPile - the once alternative to Google is getting a comeback and is a great alternative to bigger search engines.
- Duck Duck Go - Doesn't track your search history and is avoids spammy sites.
- The Internet Archive - This search engine lets users travel back in time to see how web pages looked in years gone by. A very fun search engine to play around with.

Index Size & Estimate

- Capture / Recapture Method
  - Suppose that we could pick a random page from the index of E1 and test whether it is in E2's index and symmetrically, test whether a random page from E2 is in E1.
  - These experiments give us fractions x and y such that our estimate is that a fraction x of the pages in E1 are in E2, while a fraction y of the pages in E2 are in E1.
  - Then, letting |Ei| denote the size of the index of search engine Ei, we have $x|E1| \approx y|E2|$,from which we have the form we will use $|E1|/|E2| \approx y/x$

Index Size & Estimate

- Sampling Methods
  - Random Searches
  - Random IP addresses
  - Random Walks
  - Random Queries
- Actual Estimate is quite challenging

Duplicate / Near Duplicate Detection

- Web pages are mirrored for redundancy and high availability, hence while indexing for web search engine we may come up for duplicate (identical copy). Checksum is a common method to detect a duplicate.
- Near Duplicate – not identical, but a portion is common, based on pre-set threshold we can filter out the near duplicates.
- Shingling - Given a positive integer k and a sequence of terms in a document d, define the k-shingles of d to be the set of all consecutive sequences of k terms in d.

Shingling

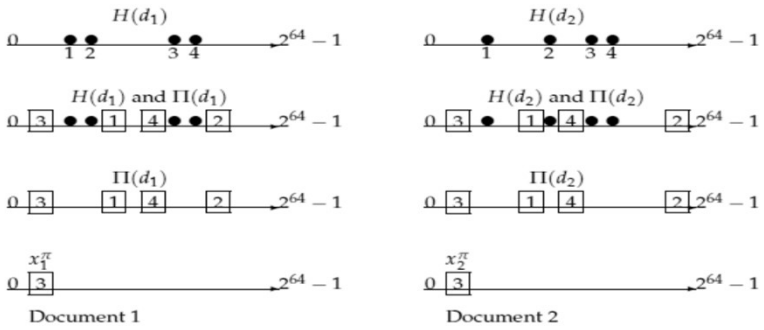
- To find a near duplicate, a shingling approach is used. If there are many common shingling for some k in a pair of documents, its contents will be the same.
- Consider a sentence below

  a rose is a rose is a rose.
- Its shingling set Z = {a rose is a ; rose is a rose ; is a rose is ; a rose is a ; rose is a rose } , which has |Z|=5
- Overlap, by Jaccard = 2/5

Near-Duplicate Scaled Approach

- A pair-wise approach seems unavoidable for using shingling overlap to detect near duplicate.
- We can perform better, by using a large integer Hash Function and doing Hashing for shingling patterns.

Near-Duplicate Scaled Approach



► **Figure 19.8** Illustration of shingle sketches. We see two documents going through four stages of shingle sketch computation. In the first step (top row), we apply a 64-bit hash to each shingle from each document to obtain $H(d_1)$ and $H(d_2)$ (circles). Next, we apply a random permutation Π to permute $H(d_1)$ and $H(d_2)$, obtaining $\Pi(d_1)$ and $\Pi(d_2)$ (squares). The third row shows only $\Pi(d_1)$ and $\Pi(d_2)$, while the bottom row shows the minimum values x_1^π and x_2^π for each document.