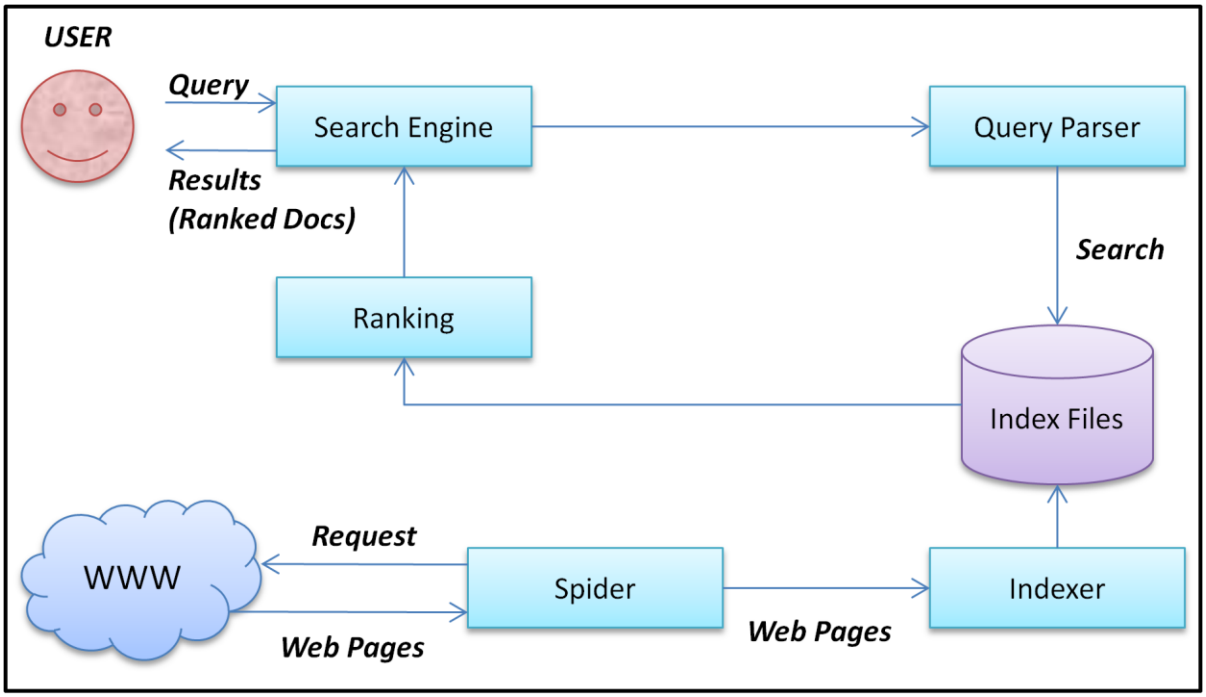
SEARCH ENGINE

- A program that searches for documents for specified keyword and returns a list of the documents where the keywords are found.
- Typically, a search engine works by sending out a spider to fetch as many as documents.
- Another program indexer then reads these documents and creates index based on the word contained in each document such that only meaningful results are retrieved to query.
- A web search engine is designed to search the information on WWW.



ARCHITECTURE OF SEARCH ENGINE

HOW DOES SEARCH ENGINE WORK?

- A search engine operates in the following order:
  1. Web crawling
  2. Indexing
  3. Searching
- Web search engine works by storing information about many web pages which they retrieve.

- These pages are retrieved by a web crawler (sometimes also called spiders), i.e. automated web browser which follows every links on the site.
- Exclusions can be made by the use of robots.txt.
- The contexts of each page can be analyzed to determine how it should be indexed.
- When a user enters a query into a search engine, the engine examines and provides the listing of best matching web pages with ranking.

WEB CRAWLING

- Web crawling is the process by which we gather pages form the web, in order to index them and support a search engine.
- The feature of a crawler must provide; (1) *Robustness* [detect the spider trap] and (2) *Politeness* [follow the restrictions to spider (robots.txt)].

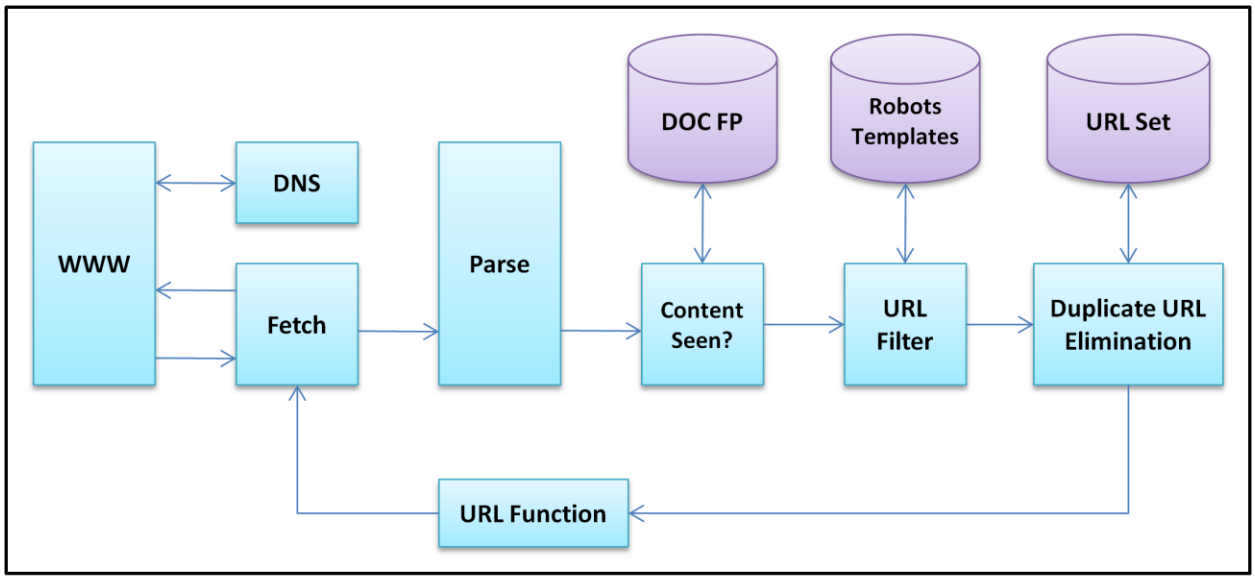
FEATURES A CRAWLER SHOULD PROVIDE

1. Distributed
2. Performance and Efficiency
3. Quality
4. Freshness
5. Extensible

CRAWLING OPERATION

- The crawler begins with one or more URLs that constitute a seed set.
- It picks a URL from this seed set, and then fetches the web page at that URL.
- The fetched page is then parsed to extract both the text and the links form the page.
- The extracted text is fed to a text indexer.
- The extracted links (URLs) are then added to a URL frontier which at all time consists of URLs whose corresponding pages have yet to be fetched by crawler.
- Initially URL frontier contains the seed set.

CRAWLER ARCHITECTURE



THE BASIC CRAWLER ARCHITECTURE

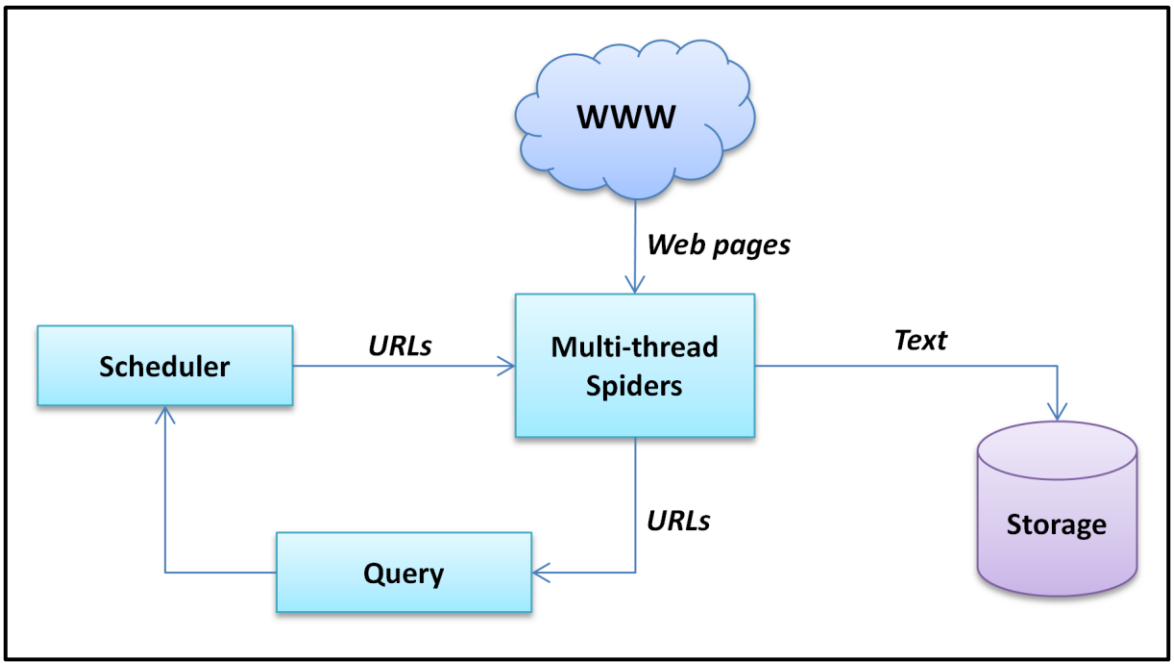
- URL Frontier: Contains URL's yet to be fetched.
- DNS Resolution Module: Determines the web server from which to fetch the page specified by the URL.
- Fetch Module: Retrieve web page at URL.
- Parsing Module: Extracts the text and set of links from a fetched web page.
- Duplicate Elimination Module: Determines whether an extracted link is already in the URL frontier.
- URL Filter: Determine whether the extracted link should be excluded from the URL frontier (for example: Robots Exclusion Protocol).
- Document Finger Print Module: Checks whether a web page with the same content has been already seen at another URL.

SEARCH ENGINE SPIDER

- A spider is a program that a search engine uses to seek out information on WWW as well as to index the information that it finds so that actual search results appear when a search query for a keyword is entered.

- The spider reads the text on the web page and records any hyperlinks found.
- The search engine spider then follows these URL's, spider those pages, collects all the data by saving copies of the web pages into the index or the search engine for use by visitors.
- Search engine spiders are always working, sometimes to index new web pages and sometimes to update ones that change frequently.
- Goal of search engine spiders is to supply up-to-date materials to search engine.
- There are four distinct styles of behavior of search engine spider. They are:
  1. Selection: decide which page needs to be downloaded.
  2. Re-visitation: to check for changes in pages that has already been indexed.
  3. Politeness: obey the restrictions to spider.
  4. Parallelization: spiders are working on parallelization to co-ordinate with other spiders.

STRUCTURE OF A SPIDER



SPIDERING ALGORITHM

- Initialize queue (Q) with initial set of known URL's.
- Until Q is empty or page or time limit exhausted,

  - Pop URL, L from front of Q.
  - If L is not to an HTML page (.gif, .jpeg, .ps, .pdf, .ppt, etc) then continue loop.
  - If L is already visited then continue loop.
  - Download page P for L.
  - If cannot download P (For example: 404 error, robot excluded) then continue loop.
  - Index P.
  - Parse P to obtain list of new links N.
  - Append N to the end of Q.
- Loop

MULTI-THREADED SPIDERING

- Bottleneck is network delay in downloading individual pages.
- Best to have multiple threads running in parallel, each requesting a page from different host.
- Distribute URL's to thread to guarantee equitable distribution of requests across different hosts to maximize throughput.
- Early Google spider had multiple co-ordinate crawlers with about 300 threads each. Together able to download 100 pages per second.

DIRECTED / FOCUSED SPIDERING

- Sort queue to explore more interesting pages first.
- Two styles of focus: (1) *Topic directed* and (2) *Link directed*.

1. Topic Directed Spidering
   - Assume desired description or sample pages of interest are given.
   - Sort queue of links by similarity like using cosine similarity of their source pages and/or anchor text to this topic description.
   - Explores pages related to a specific topic.
2. Link Directed Spidering
   - Monitor links and keep track of in-degree and out-degree of each page encountered.

- Sort queue of preferred popular pages with many incoming links (authorities).
- Sort queue to preferred summary pages with many outgoing links (hubs).

LINK ANALYSIS

- Use of hyperlinks for ranking web search results
- Link analysis is one of many factors considered by web search engines in computing a score for a web page on any given query
- Two methods for link analysis
  - Page rank
  - HITS (hyperlinks induced topic search)

PAGE RANK

- Developed by Larry Page at Stanford University.
- Link analysis algorithm
- A hyperlink to a page counts as a vote of support
- A page that is linked to by many pages receives a high rank and if there is no links to a web page there is no support for that page.
- Assigns to every node in the web graph a numerical score between 0 and 1 to each element of hyperlinked set of documents.
- The rank value indicates the importance of a particular page.
- A page rank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with 0.5 page rank.
- Algorithm
  - Assume a small universe of four web pages A, B, C and D.
  - The initial approximation of Page Rank would be evenly divided between the four documents.
  - Hence each document would begin with an estimated Page Rank of 0.25.
  - If pages B, C and D each only link to A, they would each confer 0.25 page rank to A.

    i.e. PR(A) = PR(B) + PR(C) + PR(D) = 0.75

- Suppose that page B has link to page C as well as to page A, while pages d has links to all three pages.
- The value of link votes is divided among all the outbound links on the page.
- Thus B gives vote worth 0.125 to page A and a vote 0.125 to page C.
- Similarly, D's page rank is 0.083 (approximately)

  i.e. PR(A) = PR(B)/2 + PR(C)/1 + PR(D)/3
- In general, PR(A) = PR(B)/L(B) + PR(C)/L(C) + PR(D)/L(D)

  i.e. PR(m) = $\sum_{n \in B_m} \frac{PR(n)}{L(n)}$

  L(page) → normalized number of outbound links

  B_m → set of all pages link to page m

AUTHORITIES AND HUBS

- Jon Kleinberg developed an algorithm that made use of the link structure of the web in order to discover and rank pages relevant for particular topics.
- A page is called an authority for the query if it contains the valuable information on the subject.
- For example: for query "car" → www.bmw.com, www.mercedes-benz.com
- Authoritative pages are truly relevant to the given query.
- However there is a second category of pages relevant to the process of finding authoritative pages called hubs.
- Hubs contain useful links towards the authoritative pages, i.e. hubs point the search engine to the right direction.
- Jon Kleinberg's algorithm called HITS identifies good authorities and hubs for a topic by assigning two numbers on a page.
  - Authority weight
  - Hub weight
- The weights are defined recursively
- A higher authority weight occurs if the page is pointed to by pages with high hub weights.

- A higher hub weight occurs if the page points to many pages with high authority weights.
- For a web page (p), h(p) = $\sum_{p \to y} a(y)$, a(p) = $\sum_{y \to p} h(y)$; where m → n denotes the existence of hyperlink from m to n.

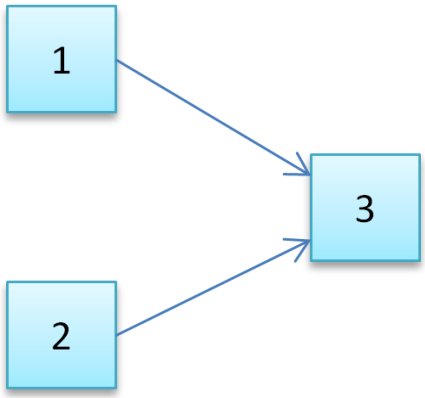
CALCULATION PROCESS

- Find adjacency matrix A, $A_{ij} = \begin{cases} 1, if\ there\ is\ a\ hyperlink\ from\ page\ i\ to\ j \\ 0, otherwise \end{cases}$
  - $a = A^T h$
  - $h = Aa$
  - In general, $a_i = A^T h_{i-1} = A^T A a_{i-1}$

    $h_i = A a_{i-1} = A A^T h_{i-1}$

- Example

$A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ $A^T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$



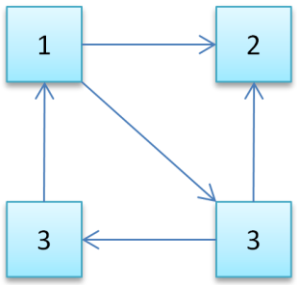
Assume that initially hub vector $h = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

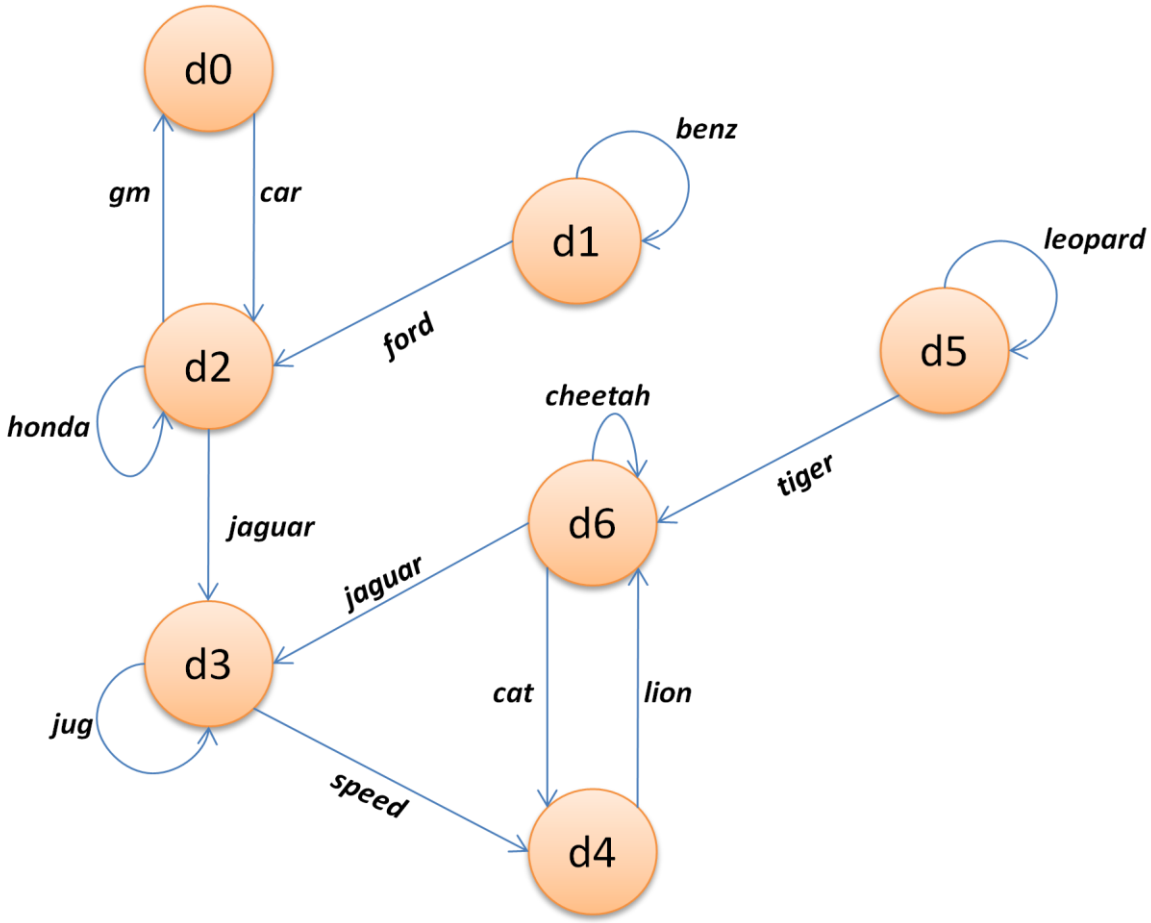
$a = A^T h = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$

Updated hub, $h = Aa = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$

i.e. node 3 is the most authority weight, since it is only one with incoming edges, and node 1 & 2 are equally important hubs.

Homework :





Assuming the query "jaguar" and double weighting of links whose anchors contain the query word

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 \end{bmatrix}$$

$$\vec{h} = \left(\frac{1}{\lambda h}\right) AA^T \vec{h} \; (\lambda h = AA^T)$$

$$\vec{a} = \left(\frac{1}{\lambda a}\right) AA^T \vec{a} \; (\lambda a = AA^T)$$

$\vec{h}$ = (0.03 0.04 0.33 0.18 0.04 0.35)

$\vec{a}$ = (0.10 0.01 0.12 0.47 0.16 0.01 0.13)

SHOPPING AGENTS

- Some people would be happy if they could find a product on the web at any price.
- Others are bargain shopping and want to find the best price available anywhere on the web.
- Software for comparison shopping are shopping agents, shopping bots, shop bots.
- Not only compare products but keep looking for them over time so that you can be notified as new items that suit your personal tasks becomes available.
- They may also be able to suggest other items that might substitute for or enhance the item you are looking for.
- Example:
  - shopping.yahoo.com → comparison shopping for broad range of products
  - shopping.com → shopping ideas with reviews
  - epinions.com → helps you decide what to buy and where to buy

INTERNET BOT (BOT, WEB ROBOTS)

- Internet bots are software applications that run automated tasks over the internet.
- Largest use of bots is in web spidering in which an automated script fetches and analyzes the information from WWW.
- Each server can have a file called robots.txt containing rules for the spidering of that server that bot is supplied to obey.
- Other examples are chat bot.