The Web IR and Crawling

Vasudeva Varma
IIIT Hyderabad


Web 1.0
Web 2.0
Web 3.0

The web and its challenges

- Unusual and diverse documents
- Unusual and diverse users, queries, information needs
- Beyond terms, exploit ideas from social networks
  - link analysis, clickstreams ...


Surface Web
Deep Web
Dark Web

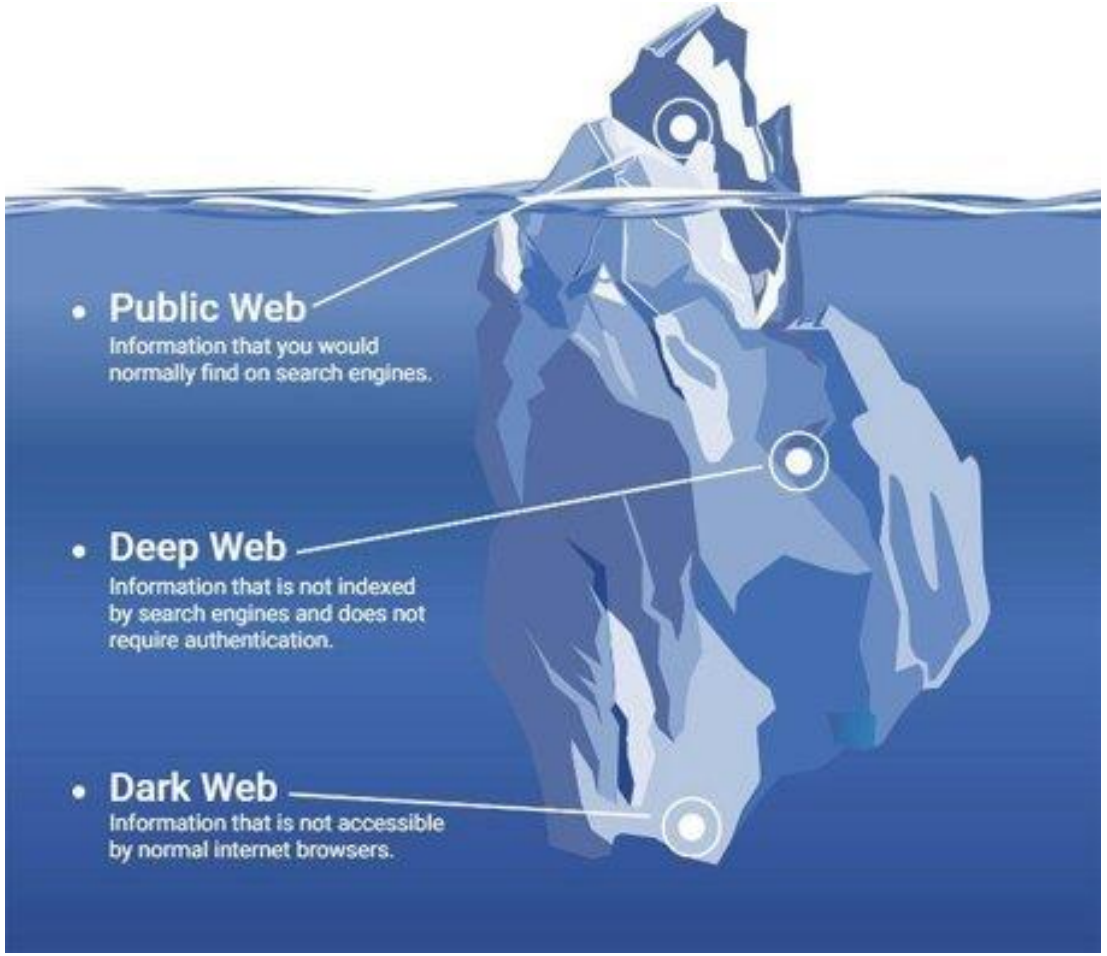

SURFACE WEB
Google
Bing
Wikipedia
DEEP WEB
Academic Information
Medical Records
Legal Documents
Scientific Reports
Subscription Information
Contains 90% of the information on the Internet, but is not accessible by Surface Web crawlers.
Social Media
Multilingual Databases
Financial Records
Government Resources
Competitor Websites
Organization-specific Repositories
(DARK WEB)
A part of the Deep Web accessible only through certain browsers such as Tor designed to ensure anonymity. Deep Web Technologies has zero involvement with the Dark Web.


THE WEB SEA
Google
PORN
DEEP WEB
HACKERS
WIKILEAKS
GOVERNMENT
UNDETECTED
ILLEGAL PORN

The web: size

- What is being measured?
  - Number of hosts
  - Number of (static) html pages
    - Volume of data
- Number of hosts – netcraft survey
  - http://news.netcraft.com/archives/web_server_survey.html
  - Gives monthly report on how many web servers are out there
- Number of pages – numerous estimates
  - For a Web engine: how big its index is
- https://www.internetlivestats.com/

The Web: Summary



Surface web is the public web

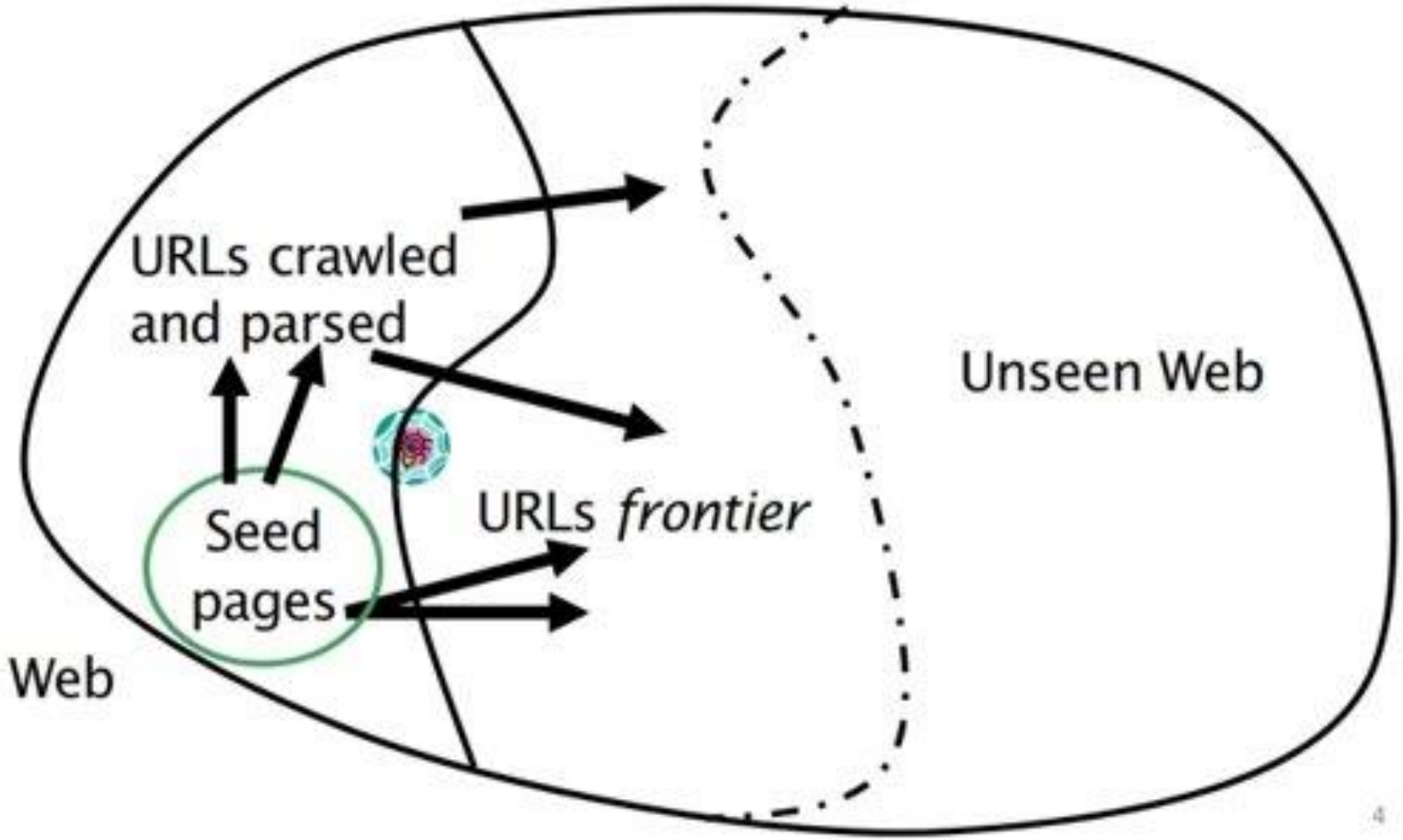
Deep Web
- includes mail servers, private databases
- can be crawled with explicit permissions

Don’t touch the dark Web

Crawling overview

- Types of crawlers
- Functionality
  - Start the crawl: Seed URLs
  - …
  - End the crawl: Halting Criteria
- Policies
- Architecture of a Web crawler

Web Crawling



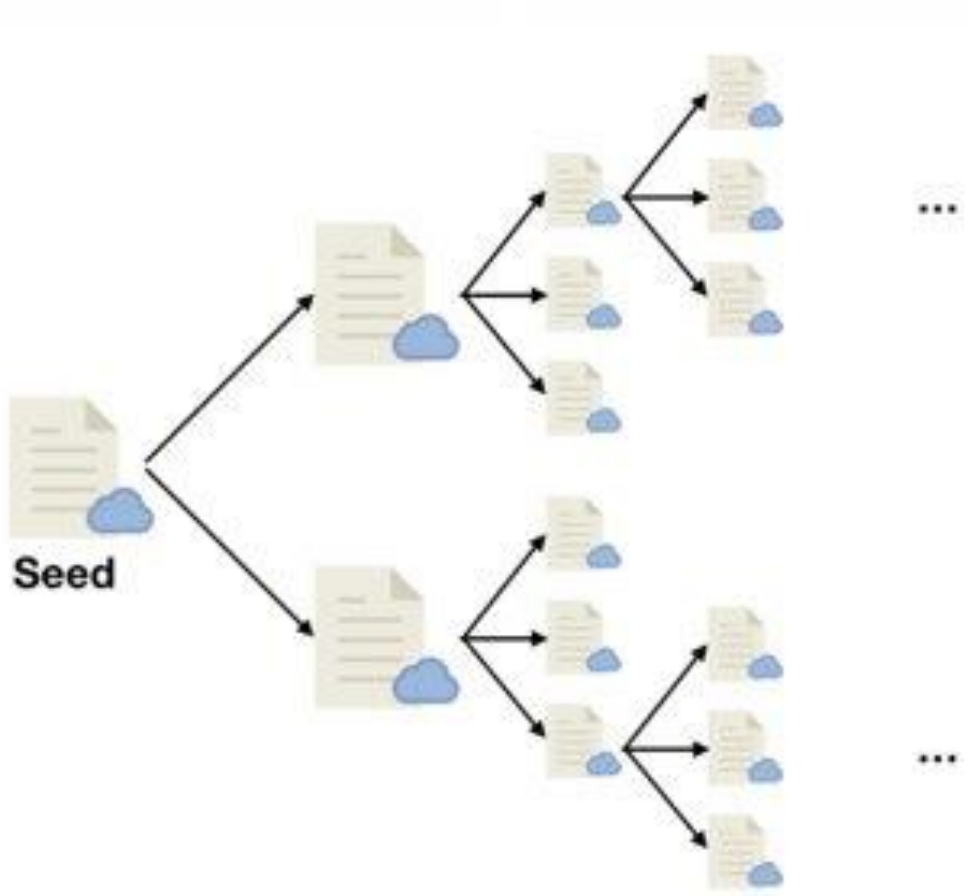
Popular Crawlers

- GoogleBot
- BingBot
- MSNbot
- Slug
- Yahoo!Slurp

❑ Nutch
❑ Scrapy
❑ DataParkSearch
❑ Grub
❑ Heritrix

Web Crawling

1. Begin with known *seed* URLs
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



Types of Crawlers

- Search engine crawlers
- Enterprise crawlers
- Monitoring crawlers
  - Copyright violation checkers
  - DRM crawlers
  - Malware detection
  - Web analytics
- Document feeds (RSS/Atom or commercial feeds)

Functionality of the crawlers

- Start with seed URLs
  - Selection of seed URLs is important
    - Quality (avoid spam/objectionable/non-hub pages)
    - Importance (popularity/trustworthiness/reliability)
    - Potential yield documents
  - Web graph helps pick right seed URLs
- Survive
  - Avoid crawler traps
    - Causes infinite number of requests being made
    - Infinitely deep directory structures
  - Follow the rules and behave well (adhere to ***policies***)
- End when time comes
  - Some crawlers are designed to go on forever
  - Some stop when a particular criteria is met (after reaching depth K, after crawling N pages or time T, after Index reaches K Units)

Policies

The behavior of a Web crawler is the outcome of a combination of policies

- Selection policy: states which pages to download
  - Prioritization: predict high yield pages from web graph
- Revisit policy: states when to check for changes to the pages
  - goal: high avg Freshness and low avg age
  - Two policies: Uniform policy or proportional policy
- Politeness policy: states how to avoid overloading Web sites
  - Robots.txt
  - Sitemap – organize the site to control crawling its parts
  - Meta tag: <META NAME “ROBOTS” CONTENT=“NOINDEX, NOFOLLOW”>
- Parallelization policy: states how to coordinate distributed Web crawlers

Robots exclusion protocol

- robot wants to visit http://www.example.com/welcome.html
  - checks for http://www.example.com/robots.txt
- User-agent: *
  - Disallow: /
- robots can ignore /robots.txt
  - malware robots that scan the web for security vulnerabilities
  - email address harvesters used by spammers
- /robots.txt file is publicly available
  - not meant for information hiding

~ Use of sitemaps

Survival Guide to Crawlers

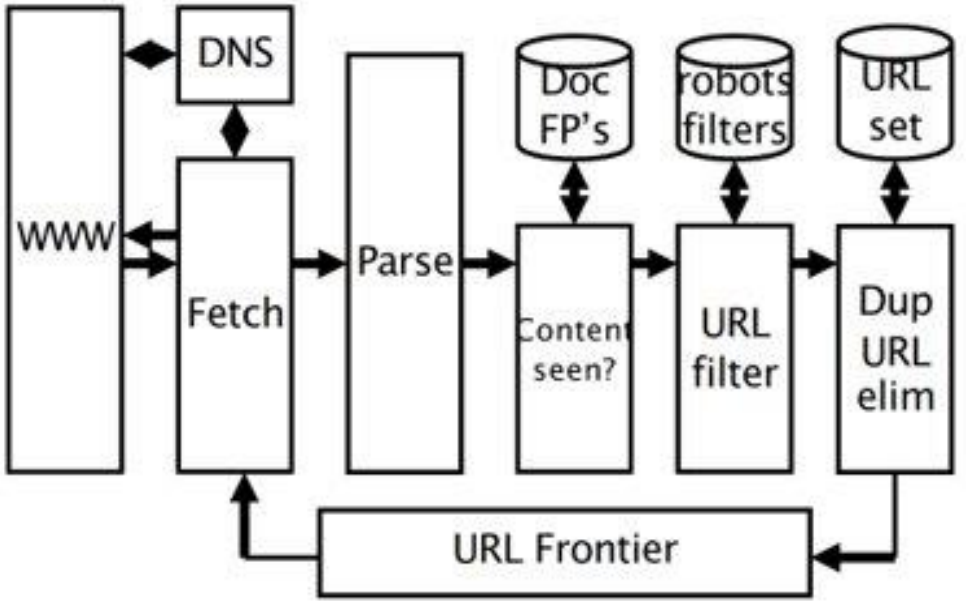
- Robustness: Be immune to spider traps and other malicious behavior from web servers
- Politeness: Respect implicit and explicit politeness considerations
  - Explicit politeness: specifications from webmasters on what portions of site can be crawled
    - adherence to policies
      - use of robots.txt
  - Implicit politeness: even with no specification, avoid hitting any site too often

Characteristics of Web Crawlers

- Capable of distributed operation
- Scalable
  - increase the crawl rate by adding more machines
- Performance and Efficiency
  - maximize usage of available processing power, network resources
- Fetch pages of higher quality first
- Continuous operation
  - fresh copies of a previously fetched page
- Extensible and Adaptable
  - to new data formats, protocols

Crawling Steps

1. Pick a URL from the frontier
2. Fetch the document at the URL
3. Parse URL
   a. Extract links
4. Check if URL has content already seen
   a. If not, add to index
5. For each extracted URL
   a. Ensure that it passes certain filter tests
   b. Check if it is already in the crawl frontier (duplicate URL elimination)



URL normalization

- When a fetched document is parsed
  - some extracted links are relative URLs

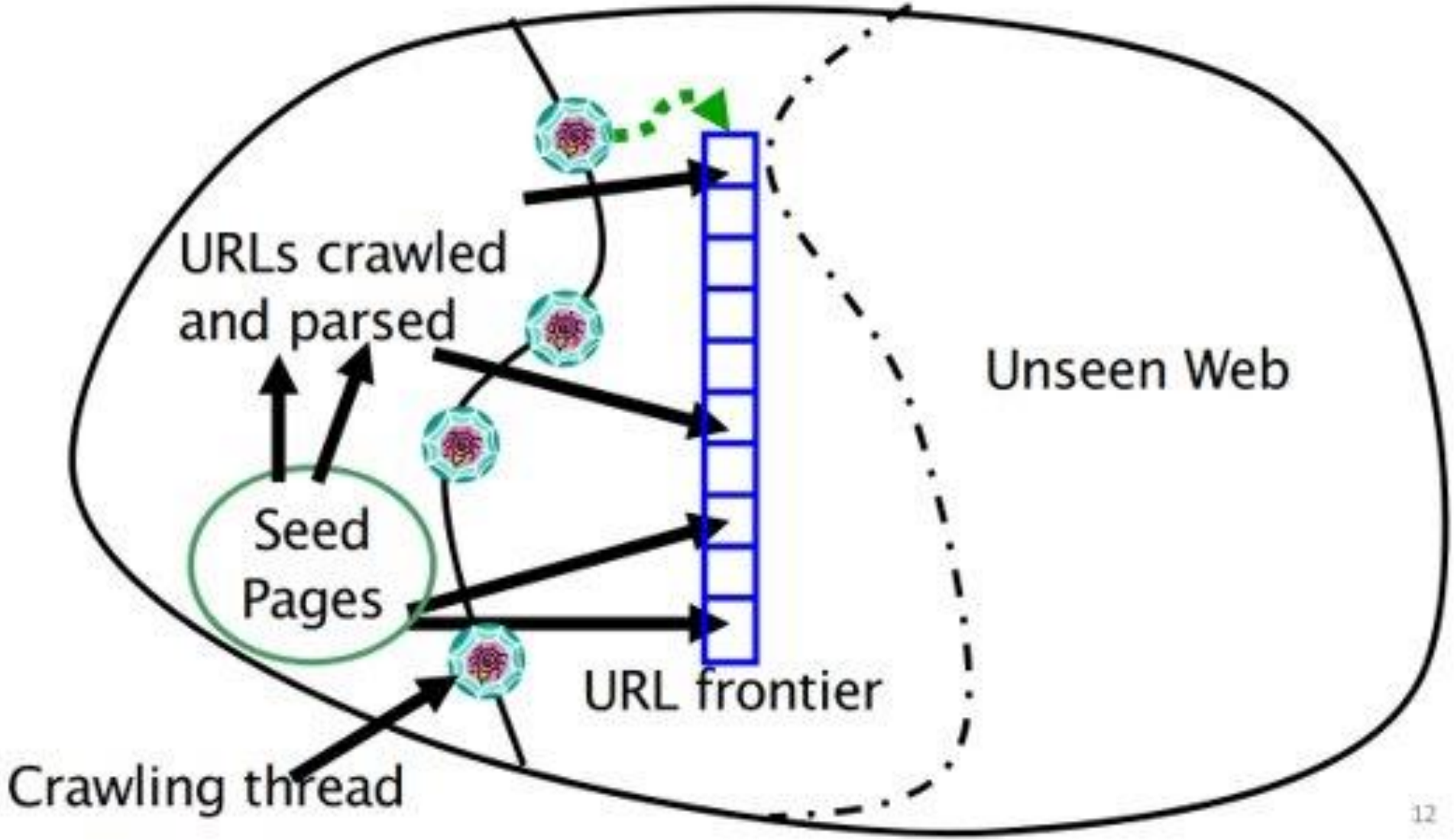
- http://en.wikipedia.org/wiki/Main_Page
  - /wiki/Wikipedia:General_disclaimer

- corresponding absolute URL
  - http://en.wikipedia.org/wiki/Wikipedia:General_disclaimer
- Normalize URLs while parsing

Content seen?

- Page Duplication
  - widespread on the internet
- Document similarity
- If page just fetched is already in the index, do not process it further

Distributed crawling



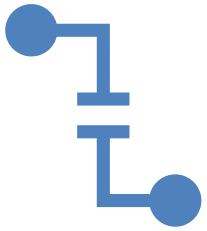
Considerations

Freshness

Crawl some pages more often than others

(News websites)

Conflicts with Politeness



Distributed crawling

one thread fetches from a single host

Insert time gap between successive fetches

Domain Specific Search

Vertical Search and Semantic Technology change Digital Advertising

Pay-Per-Click campaigns with a vertical search engine → higher click-through rates + higher conversion rates

Baidu

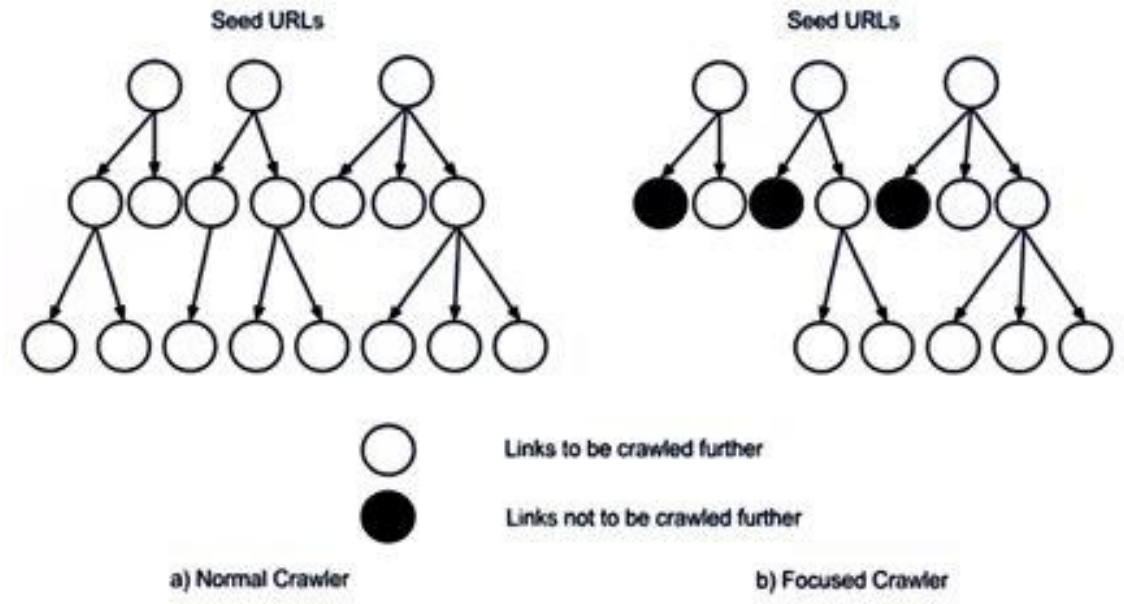
“Our deep understanding of Chinese language and culture is central to our success and this kind of knowledge allows us to tailor search technology for our users’ needs.”

— Robin Li, CEO of Baidu Inc.

Requirements

- Understanding a country’s languages and culture
- Region and domain specific tailoring of the technology
- 3 dimensions of “useful” search technology
  - Domain
  - Language
  - Region

Focused Crawling



Building seed sets for each domain

- Domain Crawling and Classification Models
- Domain Dependent Parsers

FC Metrics

- Precision: No. of relevant pages in crawl / Total number of pages crawled
- Recall: No. of relevant pages in crawl / Total number of relevant pages in the entire web
- Harvest ratio: Rate of change of precision per unit time.

Benefits of FCs

+ Results are more relevant
+ Saves resources such as time/space/computational power/bandwidth
+ Fresh crawl

- Trade off is recall

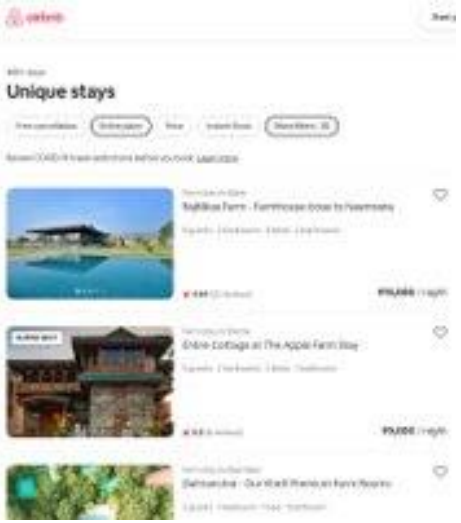
Seed Selection for Focused Crawling

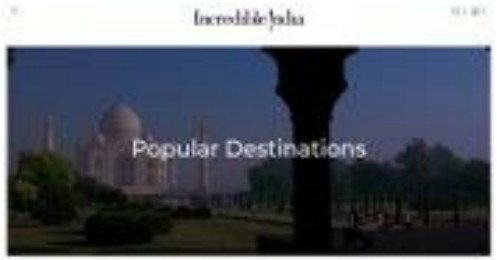
Whitelist approach

- start crawl from a list of high-quality seed URLs
- limit crawling scope to domains of such URLs
- URLs to be sorted based on rank indicators
  - top ranked URLs ~ highest crawl priority

Blacklist approach

- Names of hosts to be avoided
  - Manual creation
  - List grows





Crawl tourism related websites

Seed set

- focused on a particular domain, yet diverse within that domain

➔ Hotels
➔ Transportation
➔ Places of historic significance
➔ Famous Cuisines
➔ Shopping Malls
➔ Weather reports
➔ Tourism Industry
➔ Tourism Ministry
➔ Travel or VISA related information

Types of focused crawlers

- ➔ Classifier based focused crawlers
- ➔ Reinforcement learning based crawlers
- ➔ Context graph based focused crawlers
- ➔ Ontology based focused crawlers
- ➔ Page properties based focused crawlers
- ➔ Incrementally learning crawlers
- ➔ Adaptive Crawlers

Document Similarity

Duplicate documents

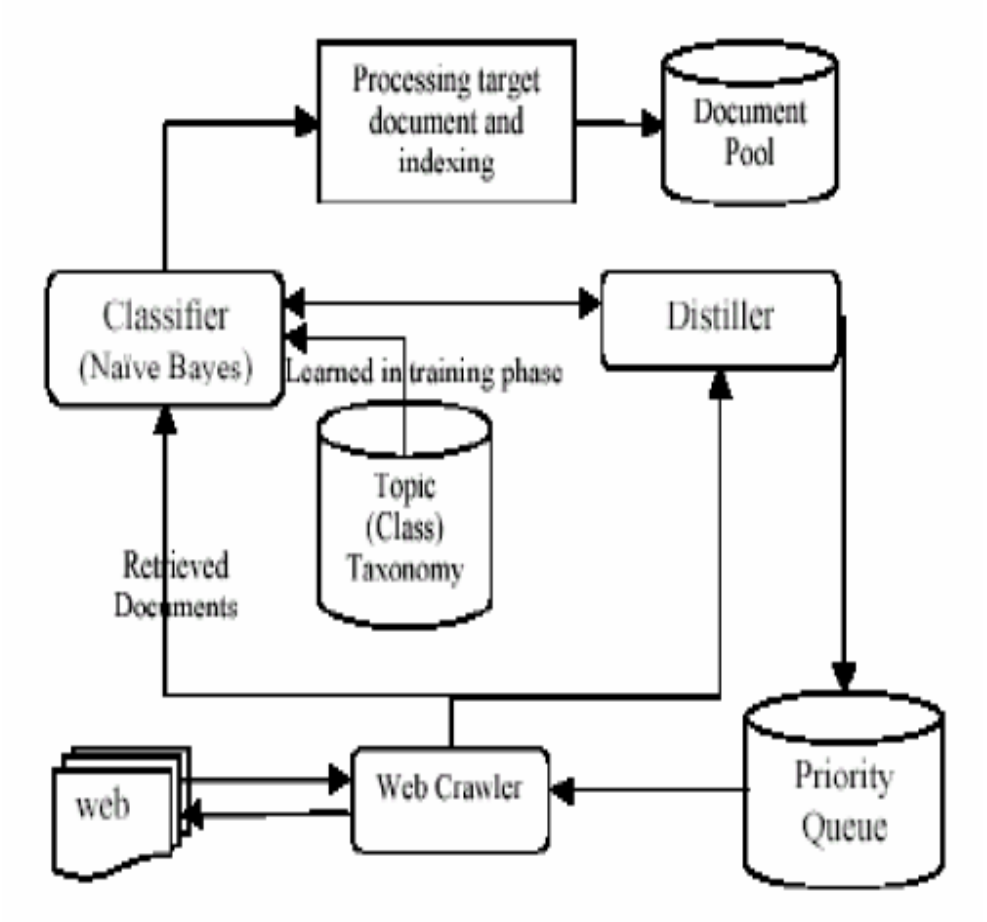
- Strict duplicate detection, exact match
  - Uncommon
  - Can be detected using fingerprints
- Near duplicates
  - Last modified date being the only difference

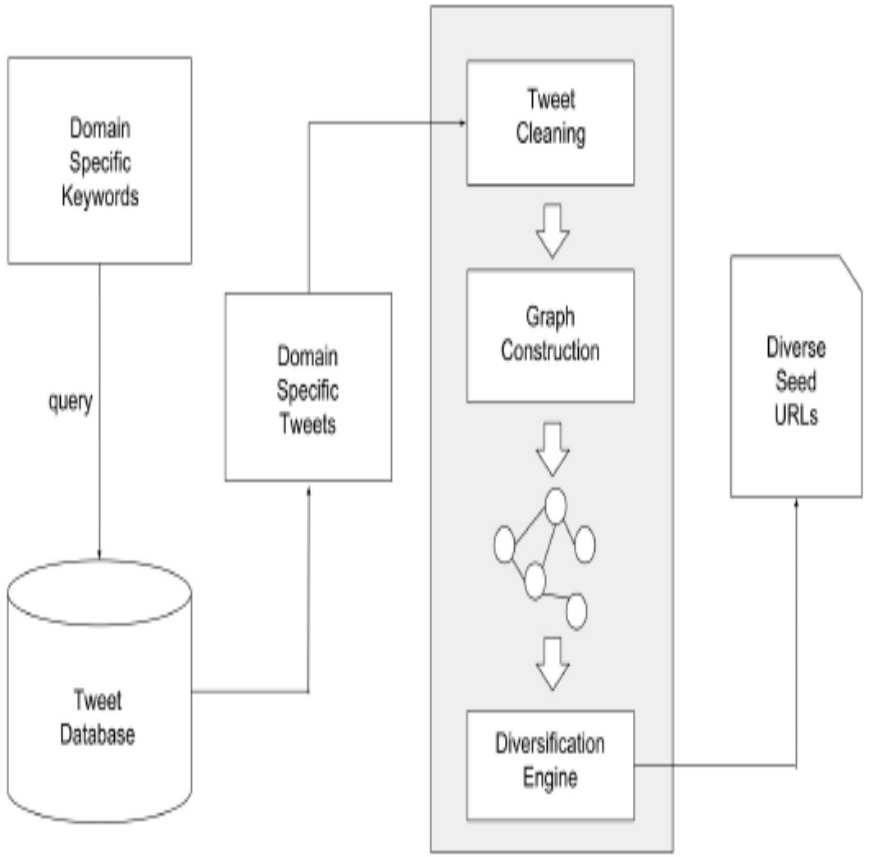
Near-Duplication detection

- Syntactic similarity with an edit distance measure
  - Similarity threshold for near-duplicates
- Segments of a document (natural or artificial breakpoints)
  - Shingles (word *N*-Grams)
- Similarity Measure between two docs (represented by their sets of shingles)
  - Jaccard coefficient [intersection / union]
- Advanced Shingling

Reference architecture for Domain Specific Crawler

- Work done by Nikhil et al at IREL





Selecting seeds for domain

Key messages

- It takes time to focus
- It takes effort and time to remove junk than to get the right pages
- Achieving high recall is the main challenge

Further reading

- Introductory article on deep web link
- Section on crawling the deep web, from this university's guide on deep web:
- Optional Reading: Paper on crawling the deep web
- Victor Lavrenko short videos on web crawling
- Implementation view