https://vvtesh.sarahah.com/

Information Retrieval

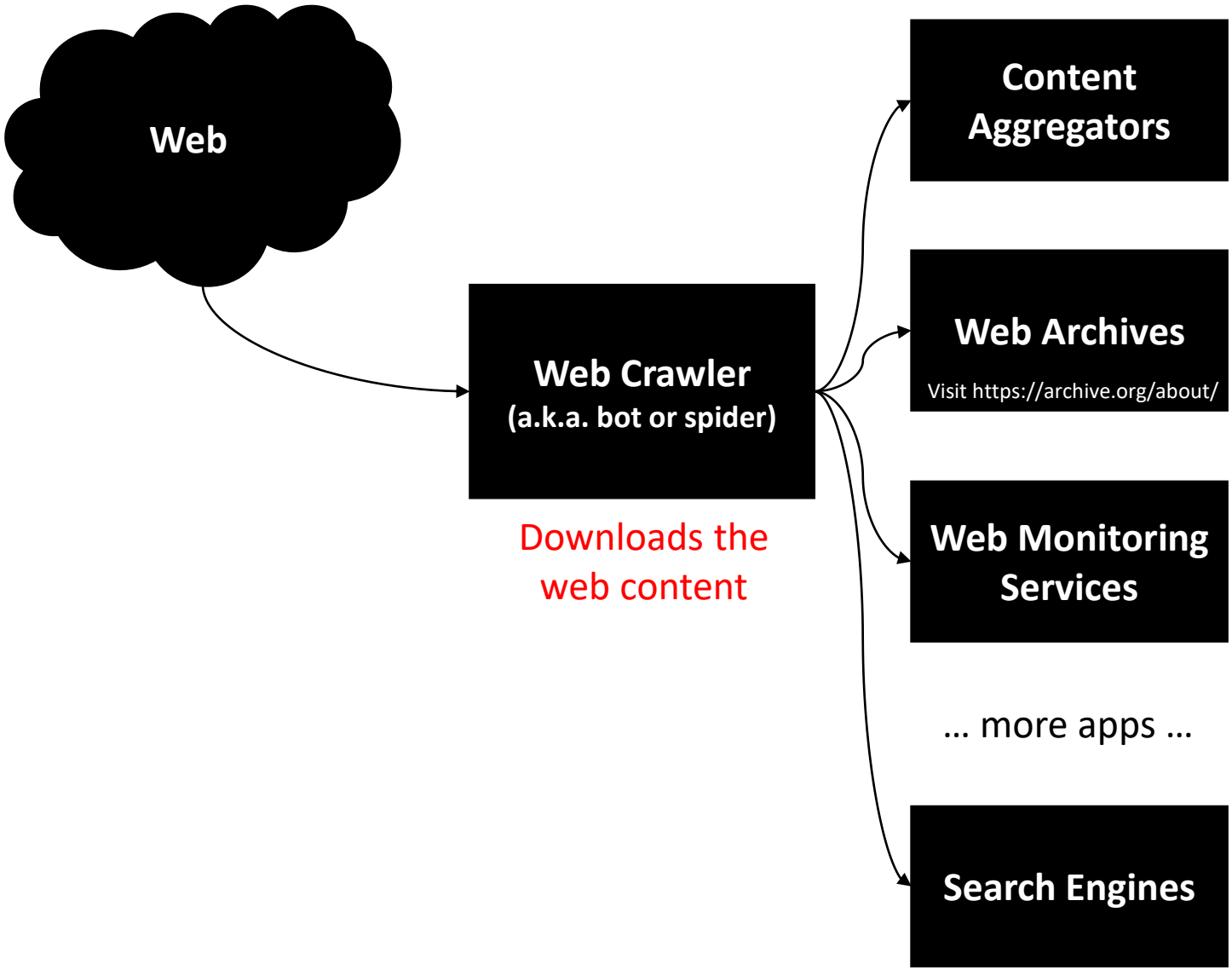
Venkatesh Vinayakarao

Term: Aug – Sep, 2019
Chennai Mathematical Institute

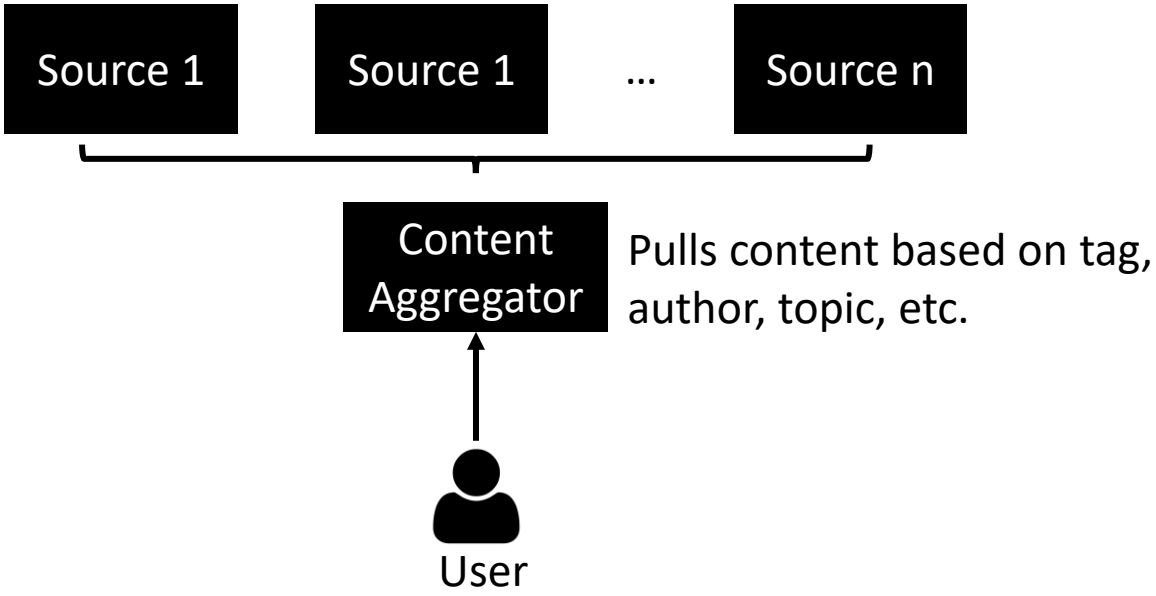
While at first glance web crawling may appear to be merely an application of breadth-first-search, the truth is that there are many challenges ranging from systems concerns such as managing very large data structures to theoretical questions such as how often to revisit evolving content sources.
-Christopher Olston and Marc Najork

An Introduction to Web Crawling

40% of web traffic is due to web crawlers!


Web
Web Crawler
(a.k.a. bot or spider)
Downloads the web content
Content Aggregators
Web Archives
Visit https://archive.org/about/
Web Monitoring Services
… more apps …
Search Engines

The Role of Content Aggregators



There are many content aggregation websites… Some have curated content, and some not.

Rotten Tomatoes POPURLS msn yahoo! finance

Web Archives

Common Crawl

Us

BIG PICTURE THE DATA ABOUT BLOG CONNECT

We build and maintain an open repository of **web crawl data** that can be **accessed and analyzed by anyone**.

About the Internet Archive

The Internet Archive, a 501(c)(3) non-profit, is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, the print disabled, and the general public. Our mission is to provide Universal Access to All Knowledge.

We began in 1996 by archiving the Internet itself, a medium that was just beginning to grow in use. Like newspapers, the content published on the web was ephemeral - but unlike newspapers, no one was saving it. Today we have 20+ years of web history accessible through the Wayback Machine and we work with 625+ library and other partners through our Archive-It program to identify important web pages.

As our web archive grew, so did our commitment to providing digital versions of other published works. Today our archive contains:

- 330 billion web pages
- 20 million books and texts
- 4.5 million audio recordings (including 180,000 live concerts)
- 4 million videos (including 1.6 million Television News programs)

We work with thousands of partners globally to save c

Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.

Search GO

Advanced Search

Cloudflare now populating and using the Internet Archive's Wayback Machine in its content distribution network application

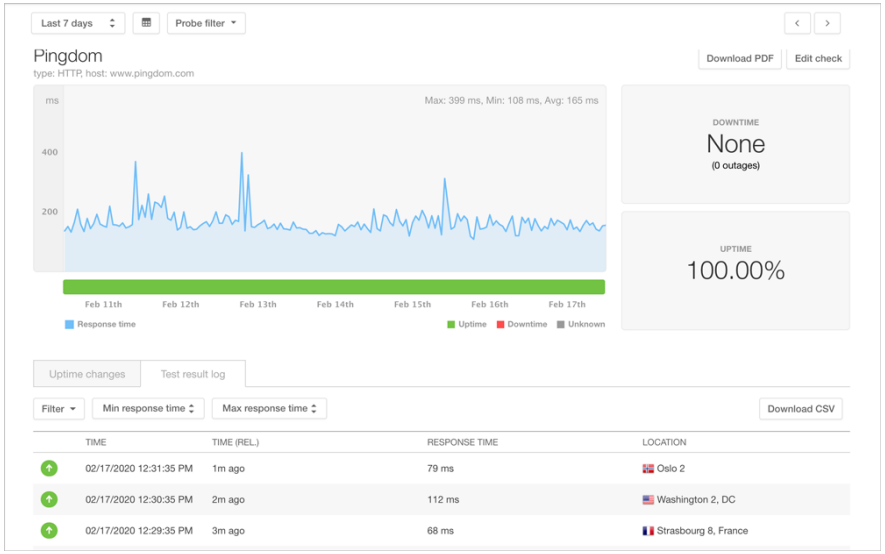
Cloudflare and the Internet Archive are now working together to help make the web more reliable. Websites that enable Cloudflare's Always Online service will now have their content automatically archived, and if by chance the original host is not available to Cloudflare, then the Internet Archive will step in to make sure the pages get through to users.

See https://archive.org/about/
https://commoncrawl.org/

Web Monitoring Services

Does your web host provide 99.99% uptime? Really?

Many services are available over the web to check.



StatusCake

Montastic
The Mighty Website Monitoring Service

solarwinds pingdom

When was the page last updated?

Does the site owner want this page “not” to be searchable?

Are there more related pages on the site?

What meta-information would a crawler like to know about a page?

Importance of the page

How frequently does the page get updated?

Some of these are readily available in the page “HEAD”er, or on the sitemap.xml.

Sitemaps


```
▼<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  ▼<url>
     <loc>https://www.icicibank.com/</loc>
     <changefreq>daily</changefreq>
     <priority>0.1</priority>
   </url>
  ▼<url>
     <loc>https://www.icicibank.com/Personal-Banking/account-deposit/fixed-deposit/index.page</loc>
     <changefreq>daily</changefreq>
     <priority>0.9</priority>
   </url>
  ▼<url>
     <loc>https://www.icicibank.com/Personal-Banking/account-deposit/iwish/index.page</loc>
     <changefreq>daily</changefreq>
     <priority>0.9</priority>
   </url>
  ▼<url>
     <loc>https://www.icicibank.com/Personal-Banking/account-deposit/recurring-deposits/index.page</loc>
     <changefreq>daily</changefreq>
     <priority>0.9</priority>
   </url>
  ▼<url>
     <loc>https://www.icicibank.com/Personal-Banking/account-deposit/Tax-Saver-Fixed-Deposit/index.page</loc>
     <changefreq>daily</changefreq>
     <priority>0.9</priority>
   </url>
  ▼<url>
     <loc>https://www.icicibank.com/Personal-Banking/cards/Consumer-Cards/Credit-Card/all-cards.page</loc>
     <changefreq>daily</changefreq>
     <priority>0.9</priority>
```

View HTTP Request and Response Headers

For more information on HTTP see RFC 2616

HTTP(S)-URL: vvtesh.co.in Submit (IDN supported)

Request type: HEAD HTTP version: HTTP/1.1 User agent: Web-Sniffer

Share This Query

Share this query on a website or a forum by copying the following link:
https://websniffer.cc/?url=http://vvtesh.co.in/

Do you have a **feature request** or **found a bug**? Please, **let us know** and we'll try our best to fix or implement it just within **1 week**!

We have some interesting information about the domain **vvtesh.co.in** and the IP address **103.235.106.21**. Make sure to check it out:

- **vvtesh.co.in**
- **103.235.106.21**

HTTP Request Header

Connect to **103.235.106.21** on port **80** ... ok

```
HEAD / HTTP/1.1
User-Agent: WebSniffer/1.0 (+http://websniffer.cc/)
Host: vvtesh.co.in
Accept: */*
Referer: https://websniffer.cc/
Connection: Close
```

HTTP Response Header

Name	Value
HTTP/1.1 200 OK	
Date:	Sat, 19 Sep 2020 08:54:21 GMT
Server:	Apache
Last-Modified:	Sat, 15 Aug 2020 05:33:22 GMT
Accept-Ranges:	bytes
Content-Length:	18494
Cache-Control:	max-age=0, public
Expires:	Sat, 19 Sep 2020 08:54:21 GMT
Vary:	Accept-Encoding,User-Agent
Connection:	close
Content-Type:	text/html

Robots.txt

identifies a crawler. * refers to all crawlers.

Do not crawl these pages

```
User-agent: *
Disallow: /yoursite/temp/
```

“searchengine” crawler may crawl everything!

```
User-agent: searchengine
Disallow:
```

Site owner may add a robots.txt file to request the bots “not” to crawl certain pages.

How Many Bots Exist?

Robots Database

The Robots Database lists robot software implementations and operators.

Robots listed here have been submitted by their owners, or by web site owners who have been visited by the robots. A listing here does not mean that a robot is endorsed in any way.

For a list of User-Agents (including bots) in the wild, see www.botsvsbrowsers.com

This robots database is currently undergoing re-engineering. Due to poular demand we have restored the existing data, but addition/modification are disabled.

1. ABCdatos BotLink
2. Acme.Spider
3. Ahoy! The Homepage Finder
4. Alkaline
5. Anthill
6. Walhello appie
7. Arachnophilia
8. Arale
9. Araneo
10. AraybOt
11. ArchitextSpider
12. Aretha
13. ARIADNE
14. arks
15. AskJeeves
16. ASpider (Associative Spider)
17. ATN Worldwide
18. Atomz.com Search Robot
19. AURESYS
20. BackRub
21. Bay Spider

History

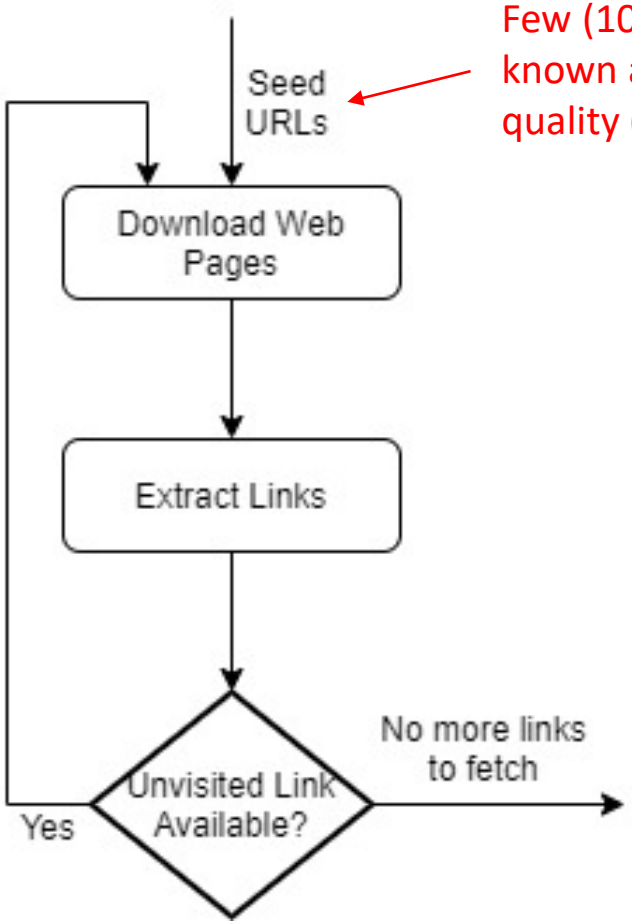
- First Generation Crawlers
  - WWW Wanderer – Matthew Gray – 1993
    - Written in Perl.
    - Worked out of a single machine.
    - Fed the index, the Wandex, thus contributing to the first search engine of the world.
  - MOMSpider
    - First polite crawler (rate of requests limited per domain).
    - Introduced “black list” to avoid crawling few sites.
  - Several followed: RBSESpider, WebCrawler, Lycos Crawler, Infoseek, Excite, AltaVista, and HotBot.
  - Brin and Page’s Google Crawler – 1998
    - Implemented with Python, asynchronous I/O, 300 downloads in parallel, 100 pages per second.

https://www.robotstxt.org/db/momspider.html
A Robots DB is here (https://www.robotstxt.org/db.html)

History

- Second Generation Crawlers (Scalable Versions)
  - Mercator - 2001
    - 891 Million pages in 17 days
  - Polybot
    - Introduced URL-Frontier (idea of seen-URLs set)
  - IBM WebFountain
    - Multi-threaded processes called Ants to crawl.
    - Applied Near-duplicate detection to reject webpages.
    - Central controller for scheduling tasks to Ants.
    - C++ and MPI (Message Passing Interface) based. Used 48 machines to crawl.
  - Several followed: UbiCrawler, IRLbot.
  - Open Source Crawlers
    - Heritrix
    - Nutch.

A Basic Crawl Algorithm



Few (10 or 100) web pages known apriori to be high-quality (popular)

Source: Web Crawling, Christopher Olston and Marc Najork, Foundations and Trends in Information Retrieval, Vol. 4, No. 3 (2010) 175–246

Challenges

Scale

Can I get high value content quickly?

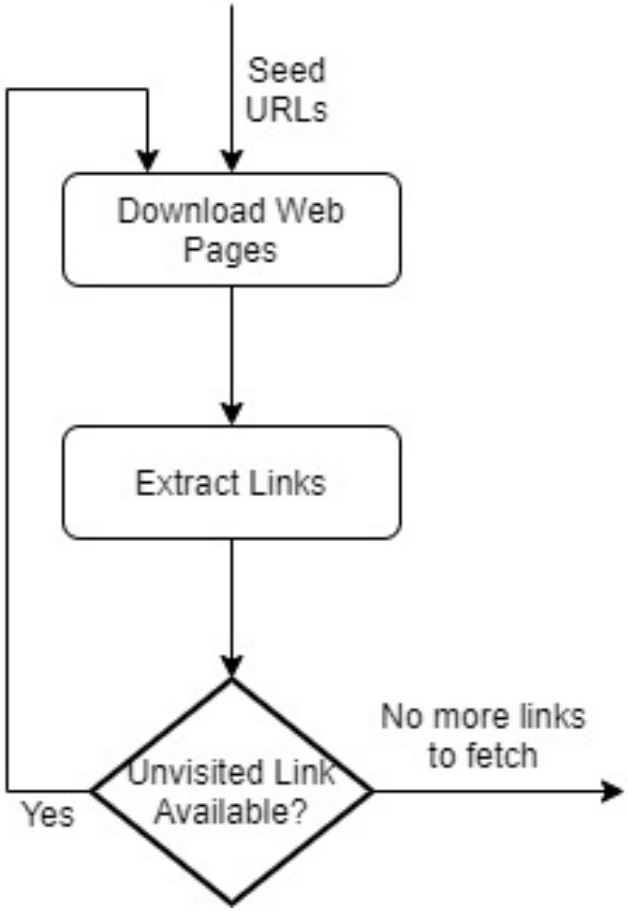
Coverage Vs. Freshness

Higher coverage => Higher Crawl Time => Lesser Freshness.

How to be fair?

Beware of adversaries

Fake Websites
Crawler Traps

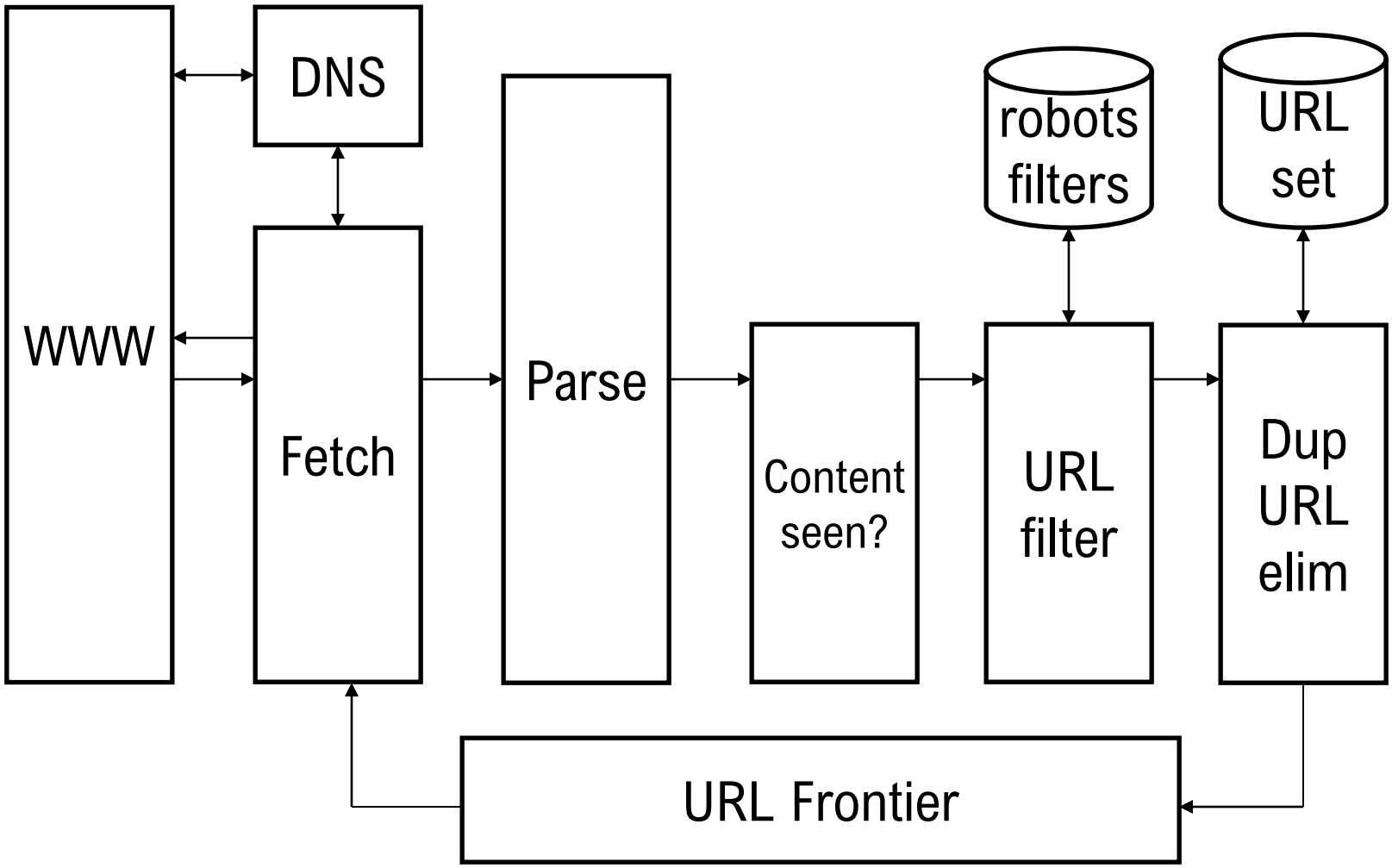


Source: Web Crawling, Christopher Olston and Marc Najork, Foundations and Trends in Information Retrieval, Vol. 4, No. 3 (2010) 175–246

Scaling to Web

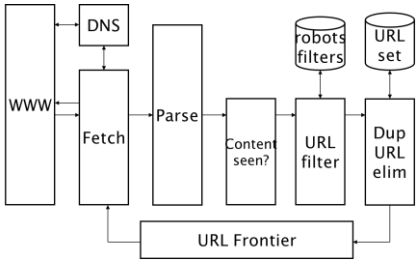
- Caching
  - Cache IP addresses to avoid repeated DNS lookups.
  - Cache robots.txt files.
- Avoid Fetching Duplicate Pages
  - Remember fetched URLs
- Prioritize
  - For Freshness

A Scalable Crawl Architecture

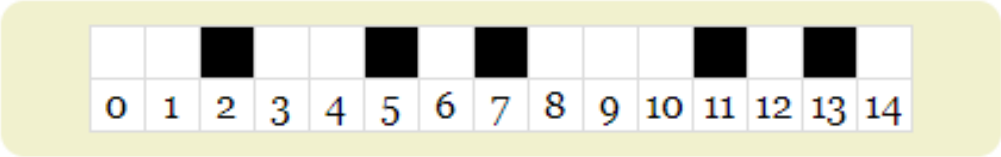


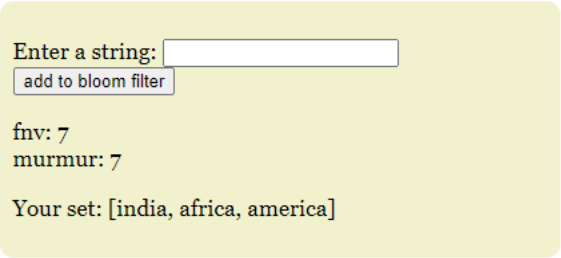
Data Structures

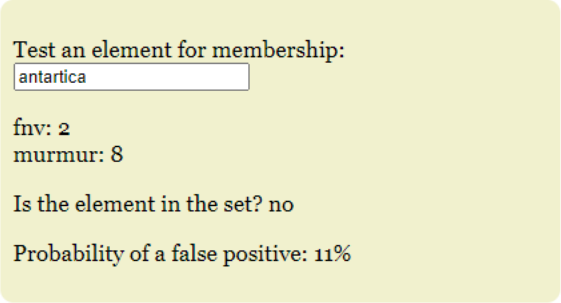
- A queue of URLs per web site.
  - Allows throttling the access per site.
- Dequeue an URL → Download the page → Extract URLs → Add them to queue → Iterate.
- A bloom filter to avoid revisiting same URL.



A Bloom Filter







https://llimllib.github.io/bloomfilter-tutorial

http://vvtesh.co.in
http://www.vvtesh.co.in
http://vvtesh.co.in/index.html
http://www.vvtesh.co.in/index.html
http://vvtesh.co.in/index.html?a=1
http://vvtesh.co.in/index.html?a=1&b=2
/index.html
teaching/../index.html
…

Same page on the web can have multiple URLs!

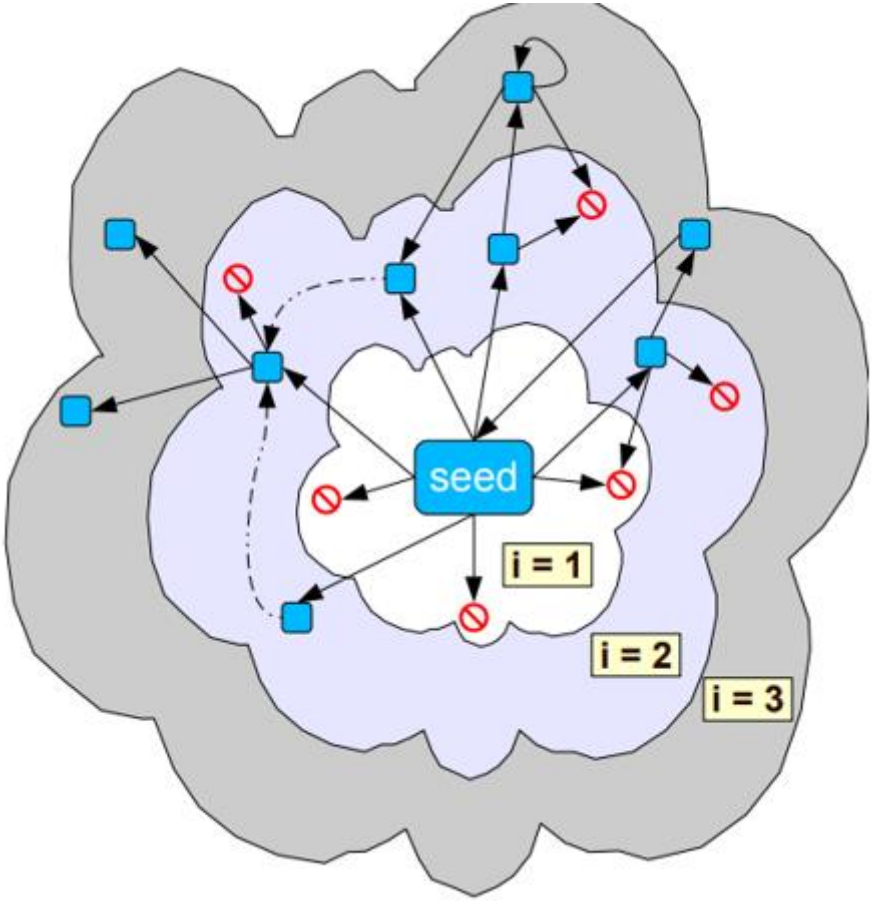
So, crawlers need to canonicalize the URLs.

You can help the crawler by identifying the canonical URL

```
<html>
<head>
<link rel="canonical" href="[canonical URL]">
</head>
</html>
```

Frontier Expansion

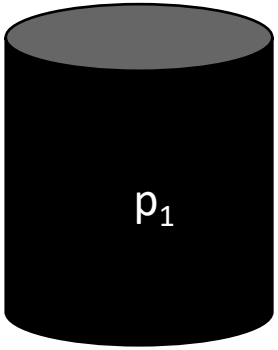
- Should we do Breadth-First or Depth-First Crawl?



How Frequently to Crawl?

Crawling the whole web every minute is not feasible.

Metrics and Terminology



A Crawl
(refers to the pages p_i collected from one crawl over the web)

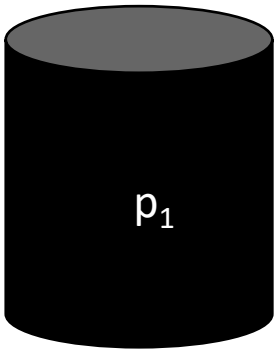
The page p_1
- is fresh if it hasn't changed after we crawled.
- is stale if it changed after we crawled.

$$\text{Freshness} = \frac{\#fresh}{\#crawled}$$

Fast changing websites bring freshness of our crawl down!

Can we do better?

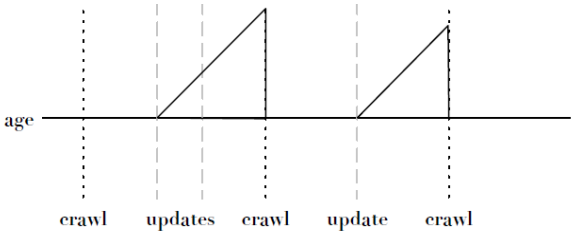
Metrics and Terminology



A Crawl

(refers to the pages p_i collected from one crawl over the web)

The page p_1
- has age 0 till it is changed.
- then its age grows until the page is crawled again.

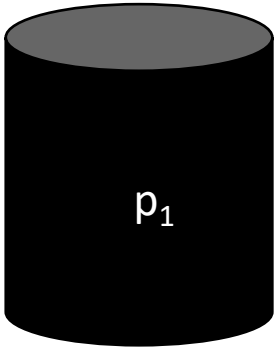


Suppose p_1 changes λ times per day.

Expected age of p_1 after t days from last crawl is:

$$\text{Age}(\lambda,t) = \int_0^t P(page\ changed\ at\ time\ x)(t-x)dx$$

Estimating the Age



A Crawl
(refers to the pages p_i collected from one crawl over the web)

Studies show that, on average, page updates follow Poisson Distribution.

Expected age of p_1 after t days from last crawl is:

$$\text{Age}(\lambda,t) = \int_0^t (\lambda e^{-\lambda x})(t-x)\,dx$$

Cho & Garcia-Molina, 2003

Crawler Traps

- Websites can generate possibly infinite URLs!
  - Often setup by spammers
  - E.g., Dynamically redirect to infinitely deep directory structures like http://example.com/bar/foo/bar/foo/bar/foo/bar/...

Host	Docs	New Docs	Data	New Data	Blocked	Queued	Out of Scope
twitter.com	103,064	102,804	12.4 GB	12.4 GB	0	327	6,528,450
archive-it.org	114,039	47,283	4.5 GB	2.5 GB	0	2,755,615	0
www.archive-it.org	24,636	11,283	1.4 GB	937 MB	0	533,129	104
video.xx.fbcdn.net	14	4	134.6 MB	56.8 MB	0	0	0

- Several ideas to counter this has been suggested
  - E.g., "Budget Enforcement with Anti-Spam Tactics" (BEAST)

https://support.archive-it.org/hc/en-us/articles/208332943-Identify-and-avoid-crawler-traps-

Batch Vs. Incremental Crawling

- Incremental Crawling
  - Works with a base snapshot of the web.
  - Incrementally update the snapshot with new/ modified/ removed pages.
  - Works well for static web pages.
- Batch Crawling
  - Easier to implement.
  - Works well for dynamic web pages.
- Usually, we mix both.

Incremental Crawling, Kevin S. McCurley.

Distributed Crawling

- Can we use cloud computing techniques to distribute the crawling task?
- Yes! Modern search engines use several thousand computers to crawl the web.
- Challenges
  - We don’t like multiple nodes to download the same URL, do the same DNS look-ups or parse the same HTML pages.
- Solutions
  - Hash URLs to nodes.
  - Use central URL Frontier, caches and queues.

Read Cho and Garcia-Molina, Parallel Crawlers, WWW 2002.

Summary

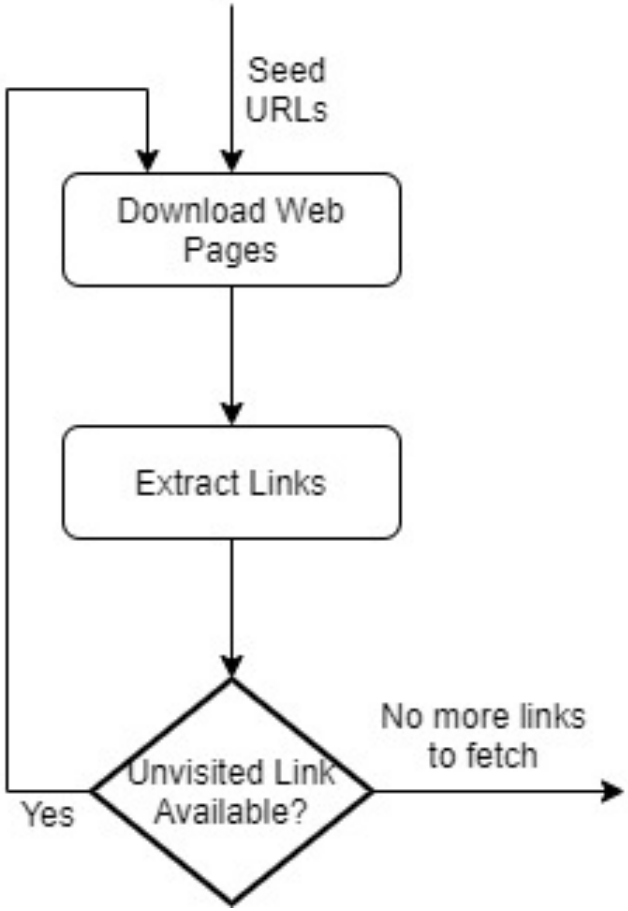
Scale

Can I get high value content quickly?

Coverage Vs. Freshness

How to be fair?

Beware of adversaries



An Experiment

- The Hardware
  - Intel Xeon E5 1630v3 4core 3.7 GHz
  - 64 GB of RAM DDR4 ECC 2133 MHz
  - 2x480GB RAID 0 SSD
  - Ubuntu 16.10 server
- Nutch
  - 11 Million URLs fetched in ~32 hours.
- StormCrawler
  - 38 Million URLs fetched in ~66 hours.

https://dzone.com/articles/the-battle-of-the-crawlers-apache-nutch-vs-stormcr

Apache Nutch

https://cwiki.apache.org/confluence/display/NUTCH/NutchTutorial

Using a Modern Crawler is Easy!

- How do we crawl with Nutch?
  - Give a name to your agent. Add seed urls to a file.
  - Initialize the Nutch crawl db
    - Nutch inject urls/
  - Generate more URLs
    - Nutch generate –topN 100
  - Fetch the pages for those URLs
    - Nutch fetch –all
  - Parse them
    - Nutch parse –all
  - Update the db and index in solr
    - Nutch updatedb –all
    - nutch solrindex <solr-url> -all

Caution: I have dropped dedup and link inversion steps for simplicity.

https://cwiki.apache.org/confluence/display/NUTCH/NutchTutorial

Readings/Playlists

- Berlin Buzzwords 2010 Talk on Nutch as a Web Mining Platform The Present & The Future
  - https://www.youtube.com/watch?v=fCtIHfQkUnY
- Nutch Tutorial
  - https://cwiki.apache.org/confluence/display/NUTCH/NutchTutorial
- Web Crawling, Christopher Olston and Marc Najork, Foundations and Trends in Information Retrieval, Vol. 4, No. 3 (2010) 175–246

Thank You