Master of Technology in Knowledge Engineering

KE5106
Semester II 2018

Lecture 6a: Data Sourcing

Charles Pang
Institute of Systems Science
National University of Singapore
Email: charlespang@nus.edu.sg

Part B Contents

KE5106 Part B	Dates	Contents
Lesson 6	29/7/2018 (Sat) 30/7/2018 (Mon)	Data Sourcing Webscraping 1
Lesson 7	4/8/2018 (Sat) 6/8/2018 (Mon)	NoSQL Document DB
Lesson 8	11/8/2018 (Sat) 13/8/2018 (Mon)	Webscraping2
Lesson 9	18/8/2018 (Sat) 20/8/2018 (Mon)	Column DB, Key-value DB Graph DB
Lesson 10	25/8/2018 (Sat) 27/8/2018 (Mon)	CA Presentation

Exam: Monday, 19th November, 6:30 – 10pm

Outline

- Why Data Sourcing?
- Sources of Data
- What is Web Crawling?
- Why do we want to Crawl?
- Crawling Strategy
- Robot Exclusion
- Crawling Policies
- Summary

Why Data Sourcing?

Sources of Data

1. Existing Data
   a. In-house data
   b. Open datasets
   c. Third-party data
2. Creating Your Own Data
3. Sourcing Data from the Internet

1a. In-House Data

All data that is living inside your company

- Databases
- Spreadsheets, Reports
- Intranet, webpages
- local computers

- Usually in-house data is propriety and you may not share freely.
- Just because your company owns the data, does not mean that it is freely available to you. Oftentimes you need permission to extract and use them.

1b. Open Datasets

Widely and easily available; across Industry segments

- Government/Non-profit released
  - data.gov.sg
  - World bank
  - WHO
- Public (general)
  - cloud.google.com
  - aws.amazon.com
  - nytimes.com
  - Kaggle.com
- Scientific Community
  - UCI machine learning
  - deeplearning.net
  - nature.com

1c. Third-Party Datasets

You could also buy data from data providers (Brokers)

- Fair Isaac Corporation (credit ratings)
- Nielsen (audience)
- Thomson Reuters (market)
- Twitter/Gnip (Tweets)
- Datasift (Social Media)

Why do you need to buy?

- You need personal data
- You need private data
- You need a layer of predictive information
- Open data not immediately usable

2. Creating You Own Data

What if you cannot find the data for your analysis?

- For example: You have a great idea for the next killer app; You want to prove/disprove a theory; You want to launch a new service, etc.
- Creating your own data becomes the only option. There are several ways to do this:
  - Conduct Interviews/focus groups
  - Run Surveys (like what you did in ISBA)
  - Conduct Experiments (wearables to collect vital data)

3. Sourcing Data from the Internet

The Internet is a HUGE data repository! There are several ways you can “download” data for your use

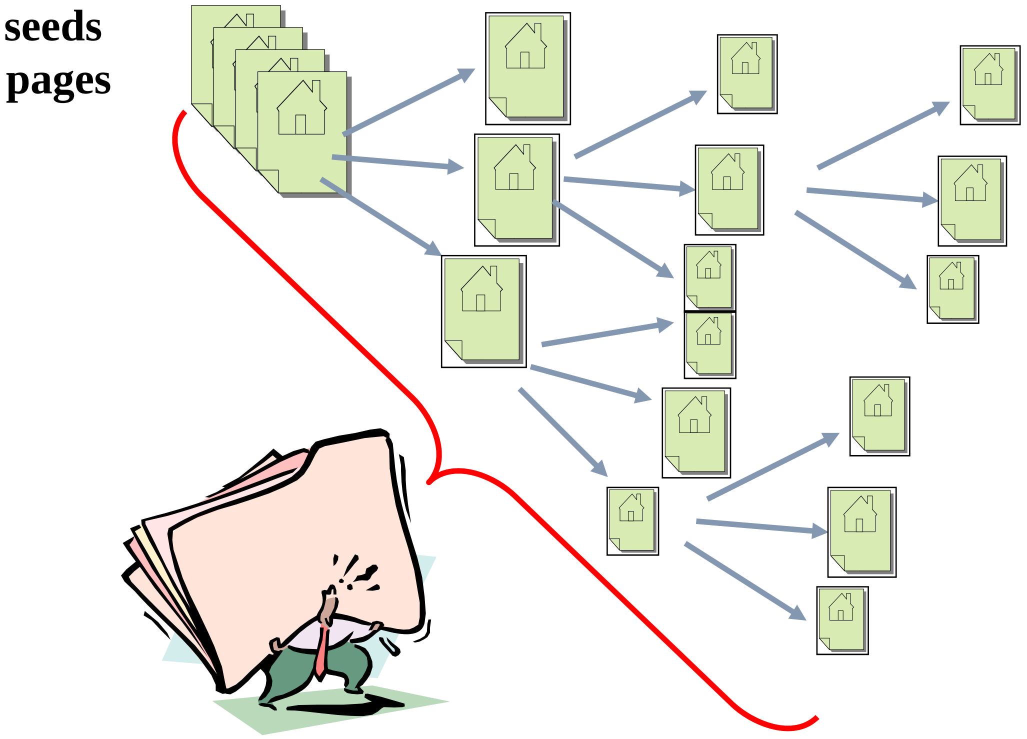
- **Ctrl-c Ctrl-v**
  - Easiest ☺
  - Doesn’t work all the time ☹
- **Application Programming Interfaces (APIs)**
  - Great way to access live streams ☺
  - But not every website offers them ☹
- **PDFs**
  - Right-click->”Save as…” ☺
  - Difficult to work with ☹
- **Web scraping/crawling**
  - Get anything you want! ☺
  - Need to write programs ☹

What is Web Crawling?

Web **crawlers**, also known as **spiders** or **robots**, are programs that automatically download Web pages. Since information on the Web is scattered among billions of pages served by millions of servers around the globe, users who browse the Web can follow hyperlinks to access information, virtually moving from one page to the next. A crawler can visit many sites to collect information that can be analyzed and mined in a central location, either online (as it is downloaded) or off-line (after it is stored).

Web Data Mining 2nd Edition, Bing Liu, Springer 2006

Web Crawling

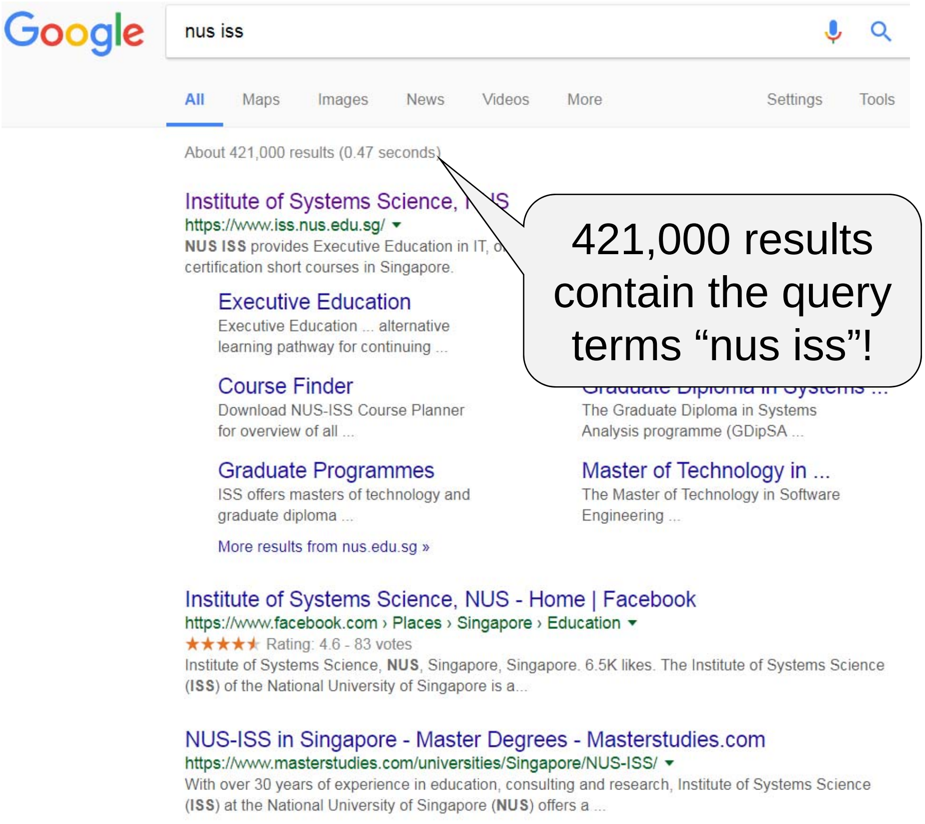


History of web crawling

- Web crawlers were written as early as 1993: World Wide Web Wanderer, Jump Station, World Wide Web Worm, and RBSE spider. These four spiders mainly collected information and statistic about the web using a set of seed URLs. Early web crawlers iteratively downloaded URLs and updated their repository of URLs through the downloaded web pages.
- By 1995, a few commercial web crawlers became available: Lycos, Infoseek, Excite, AltaVista and HotBot.
- In 1998, Brin and Page tried to address the issue of scalability by introducing a large scale web crawler called Google. Google addressed the problem of scalability in several ways: Firstly, it reduce disk access time through techniques such as compression and indexing. Secondly, it calculated the probability of a user visiting a page (*PageRank*) –it will then crawl a page as often as a user does.

Adapted from “A brief history of Web Crawlers” IBM Canada Ltd, 2013

Google Search



About 421,000 results (0.47 seconds)

Institute of Systems Science, NUS
https://www.iss.nus.edu.sg/
NUS ISS provides Executive Education in IT, o
certification short courses in Singapore.

Executive Education
Executive Education ... alternative learning pathway for continuing ...

Course Finder
Download NUS-ISS Course Planner for overview of all ...

Graduate Diploma in Systems ...
The Graduate Diploma in Systems Analysis programme (GDipSA ...

Graduate Programmes
ISS offers masters of technology and graduate diploma ...

Master of Technology in ...
The Master of Technology in Software Engineering ...

More results from nus.edu.sg »

Institute of Systems Science, NUS - Home | Facebook
https://www.facebook.com › Places › Singapore › Education
★★★★★ Rating: 4.6 - 83 votes
Institute of Systems Science, **NUS**, Singapore, Singapore. 6.5K likes. The Institute of Systems Science (**ISS**) of the National University of Singapore is a ...

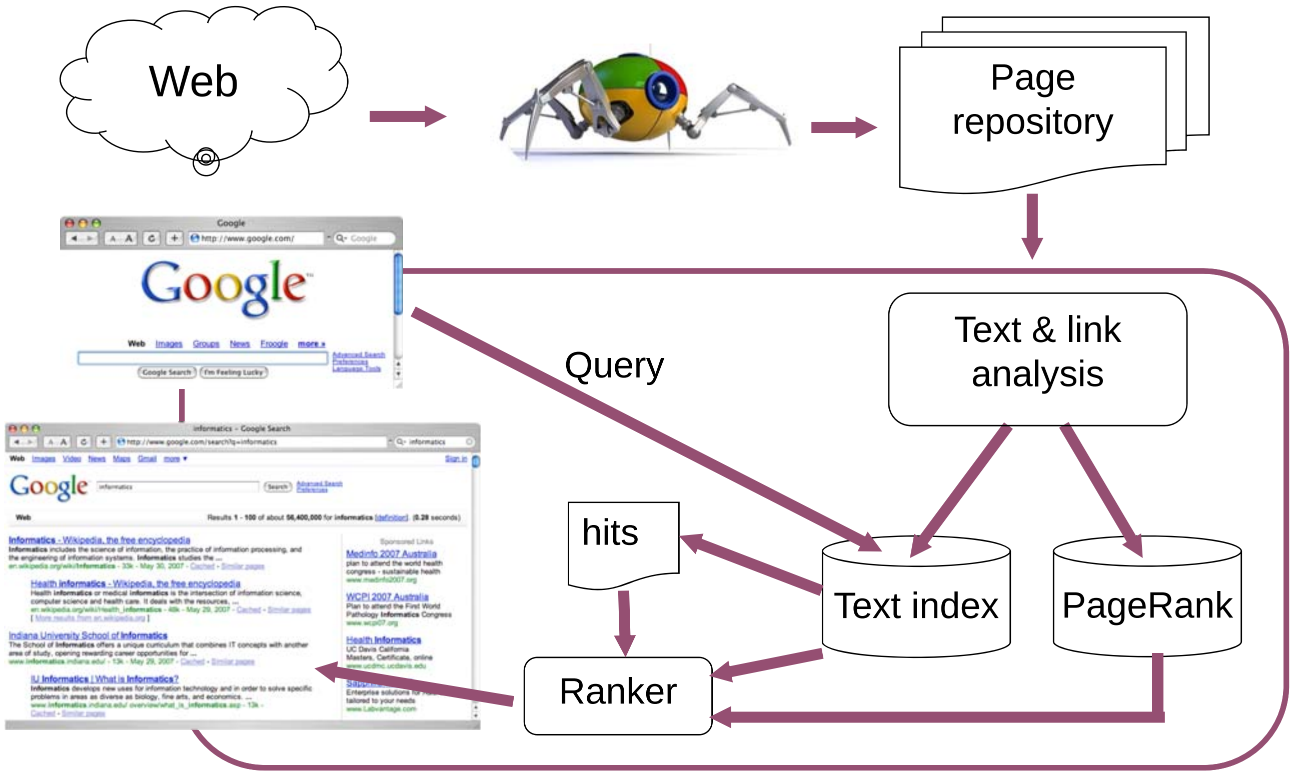
NUS-ISS in Singapore - Master Degrees - Masterstudies.com
https://www.masterstudies.com/universities/Singapore/NUS-ISS/
With over 30 years of experience in education, consulting and research, Institute of Systems Science (**ISS**) at the National University of Singapore (**NUS**) offers a ...

421,000 results contain the query terms “nus iss”!

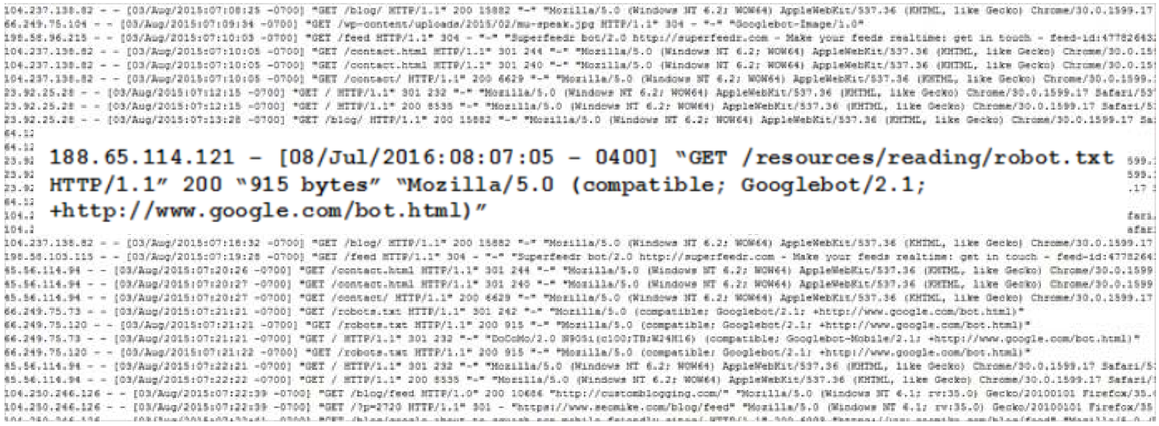
Googlebot

- Googlebot is Google's web Crawler
- Google uses a large number of computers to tirelessly crawl billions of web pages – WWW currently has 4.47 billion indexed web pages !
- Google uses an special algorithm to decide which sites to crawl, how often, and how many pages to fetch from each site.
- Googlebot's crawls each website, detects the links on each page and adds them to its list of pages to crawl. New sites, changes to existing sites, and dead links are noted and used to update the Google index.

Architecture of a Search Engine



A visit by Googlebot



```
104.237.138.82 - - [03/Aug/2015:07:08:25 -0700] "GET /blog/ HTTP/1.1" 200 15882 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.17
66.249.75.104 - - [03/Aug/2015:07:09:34 -0700] "GET /wp-content/uploads/2015/02/mu-speak.jpg HTTP/1.1" 304 - "-" "Googlebot-Image/1.0"
198.58.96.215 - - [03/Aug/2015:07:10:03 -0700] "GET /feed HTTP/1.1" 304 - "-" "Superfeedr bot/2.0 http://superfeedr.com - Make your feeds realtime: get in touch - feed-id:47782643
104.237.138.82 - - [03/Aug/2015:07:10:05 -0700] "GET /contact.html HTTP/1.1" 301 244 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.15
104.237.138.82 - - [03/Aug/2015:07:10:05 -0700] "GET /contact.html HTTP/1.1" 301 240 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.15
104.237.138.82 - - [03/Aug/2015:07:10:05 -0700] "GET /contact/ HTTP/1.1" 200 6629 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.
23.92.25.28 - - [03/Aug/2015:07:12:15 -0700] "GET / HTTP/1.1" 301 232 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.17 Safari/53
23.92.25.28 - - [03/Aug/2015:07:12:15 -0700] "GET / HTTP/1.1" 200 8535 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.17 Safari/5
23.92.25.28 - - [03/Aug/2015:07:13:28 -0700] "GET /blog/ HTTP/1.1" 200 15882 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.17 Sa

188.65.114.121 - [08/Jul/2016:08:07:05 - 0400] "GET /resources/reading/robot.txt
HTTP/1.1" 200 "915 bytes" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"

104.237.138.82 - - [03/Aug/2015:07:18:32 -0700] "GET /blog/ HTTP/1.1" 200 15882 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.17
198.58.103.115 - - [03/Aug/2015:07:19:28 -0700] "GET /feed HTTP/1.1" 304 - "-" "Superfeedr bot/2.0 http://superfeedr.com - Make your feeds realtime: get in touch - feed-id:4778264
45.56.114.94 - - [03/Aug/2015:07:20:26 -0700] "GET /contact.html HTTP/1.1" 301 244 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599
45.56.114.94 - - [03/Aug/2015:07:20:27 -0700] "GET /contact.html HTTP/1.1" 301 240 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599
45.56.114.94 - - [03/Aug/2015:07:20:27 -0700] "GET /contact/ HTTP/1.1" 200 6629 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.17
66.249.75.73 - - [03/Aug/2015:07:21:21 -0700] "GET /robots.txt HTTP/1.1" 301 242 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
66.249.75.120 - - [03/Aug/2015:07:21:21 -0700] "GET /robots.txt HTTP/1.1" 200 915 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
66.249.75.73 - - [03/Aug/2015:07:21:21 -0700] "GET / HTTP/1.1" 301 232 "-" "DoCoMo/2.0 N905i(c100;TB;W24H16) (compatible; Googlebot-Mobile/2.1; +http://www.google.com/bot.html)"
66.249.75.120 - - [03/Aug/2015:07:21:22 -0700] "GET /robots.txt HTTP/1.1" 200 915 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
45.56.114.94 - - [03/Aug/2015:07:22:21 -0700] "GET / HTTP/1.1" 301 232 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.17 Safari/5
45.56.114.94 - - [03/Aug/2015:07:22:22 -0700] "GET / HTTP/1.1" 200 8535 "-" "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/30.0.1599.17 Safari/5
104.250.246.126 - - [03/Aug/2015:07:22:39 -0700] "GET /blog/feed HTTP/1.0" 200 10686 "http://customblogging.com/" "Mozilla/5.0 (Windows NT 6.1; rv:35.0) Gecko/20100101 Firefox/35.
104.250.246.126 - - [03/Aug/2015:07:22:39 -0700] "GET /?p=2720 HTTP/1.1" 301 - "https://www.seomike.com/blog/feed" "Mozilla/5.0 (Windows NT 6.1; rv:35.0) Gecko/20100101 Firefox/35
```

Host name: 188.65.114.121
Date and Time of access: [08/Jul/2016:08:07:05 -0400]
URL they have accessed: robot.txt
User Agent: Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
Method used: GET
HTTP status code: 200

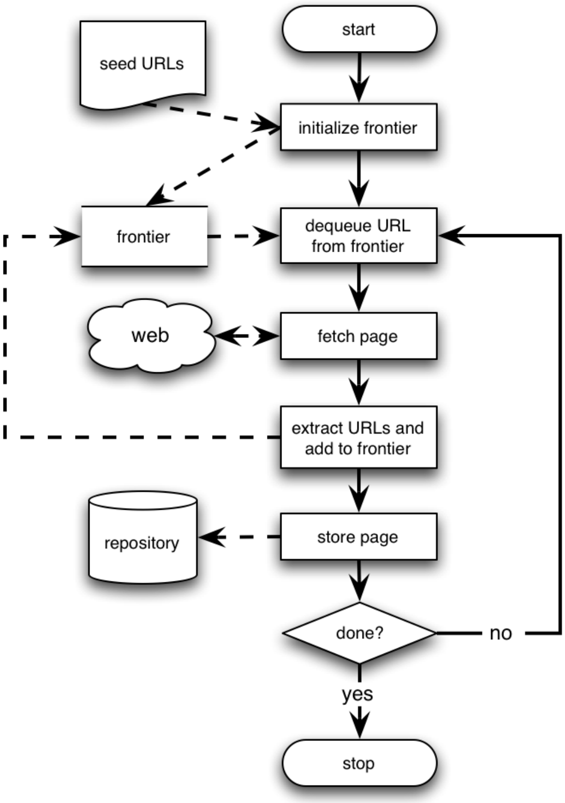
Reasons to Webcrawl

- Support universal search engines (Google, Yahoo, Bing, Apple, etc.)
- Vertical (specialized) search engines, e.g. news, shopping, papers, recipes, reviews, etc.
- Business intelligence: collect information about potential competitors, partners
- Monitor Web sites of interest: a user of community can be notified when new information appears
- Evil desires: harvest emails for spamming, personal information for phishing
- Automated testing of web applications
- Others …

A Crawling Algorithm

```
Initialize queue (Q) with initial set of known URL’s.
Until Q empty or page or time limit exhausted:
      Pop URL, L, from front of Q.
      If L is not a HTML page (.jpeg, .ps, .pdf, .ppt …)
              exit loop.
      If already visited L,
             continue loop (get next url).
      Download page, P, for L.
      If cannot download P (404 error, robot excluded)
              exit loop.
      Index P (e.g. add to index or store cached copy).
      Parse P to obtain list of new links N.
      Append N to the end of Q.
```

Crawling Overview

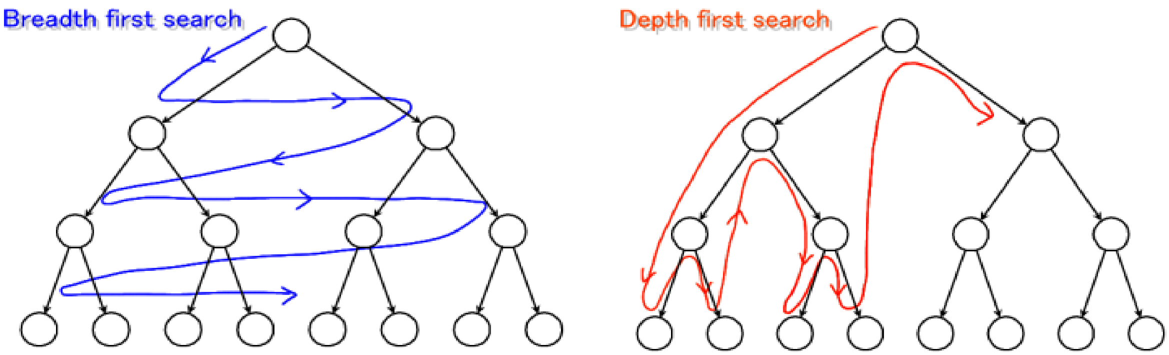


- This is a sequential crawler
- Crawling starts with Seed URLS
- As the crawler visits a URL, it identifies all hyperlinks on that page and adds them to a frontier (data structure)
- URLs in the frontier are recursively visited according to a set of policies
- Crawler Stops when:
  - All useful pages are crawled
  - Empty frontier
  - Break in crawl

Crawling Strategies

New links can be added to the queue in 2 ways:

1. Append new links to the end of the Q
   - First-in-First-Out (FIFO) approach
   - Results in a Bread-First-Search Strategy
2. Add new links to the front of the Q (LIFO)
   - Last-in-First-Out (LIFO) approach
   - Results in a Depth-First-Search Strategy



BFS or DFS

- For most general purposes, the BFS is used. Most of the time, you are targeting a website because it has most (or even all) the information that you want. Hence there is no need to go deeper.
- However, if you are looking for deep information, DFS is what you would want to use. You can also perform a depth-limiting search so that your crawler wont wonder off too far.
- Crawling doesn’t have to always stick to BFS or DFS. There are variants such as Greedy Search, Focused Search, etc.

Focused crawling

- We can sort the Q to prioritise and explore more interesting pages first.
- Two methods of prioritizing:
- Topic Focus:
  - Prioritise the Q based on similarity measure (e.g. cosine) between the desired topic and the anchor-texts in the Q
- Link Focus:
  - Prioritise the Q based on in-coming links ("authoritative")
  - Prioritise the Q based on out-going links ("aggregator")
- Another approach uses the relevance of a page after downloading its content. Relevant pages are sent for content indexing and their contained URLs are added to the crawl frontier; pages that fall below a relevance threshold are discarded.

Crawling Best Practices

- Crawlers are one of the biggest consumers of internet bandwidth. Therefore webcrawling must be carried out responsibly.
- While websites want to be crawled to boost their marketing reach, they also want to ensure that regular customers are not negatively affected.
- Websites owners can give **robot exclusion instructions** to crawling robots on what they can or cannot do while crawling their site.
- At the same time, crawlers need observe **crawling policies**

Robot Exclusion Instructions

- There are two ways that website owners can give crawling robots behavioural instructions:
  1. Robots.txt which contains site-wide specification of excluded directories.
  2. Robots META Tag which contains individual document tags to exclude indexing or following links found inside the document page.

1.Robots.txt

- It is common for website owners to deposit a "robots.txt" file at the root of the host's web directory.
  - http://www.asiaone.com/robots.txt
  - http://www.hardwarezone.com.sg/robots.txt
  - http://www.nus.edu.sg/robots.txt
- Robots.txt contains a list of excluded directories (no-go zones) for a given robot (user-agent).
- For example, to exclude all robots from the entire site:

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

Robot Exclusion Protocol Examples

- Exclude specific directories:

```
User-agent: *
Disallow: /pages/mobile_home/
Disallow: /videos/mobile_index/
Disallow: /price-guide/0/
```

- Exclude a specific robot:

```
User-agent: mtechkeBot
Disallow: /
```

- Allow a specific robot:

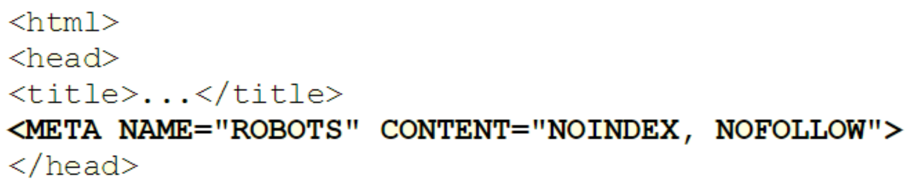
```
User-agent: GoogleBot
Disallow:
```

- No robots allowed:

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

2.Robots META Tag

- Website owners can use a special HTML <META> tag to tell robots not to index the content of a page, and/or not scan it for links to follow.



```
<html>
<head>
<title>...</title>
<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">
</head>
```

- Content value pair :
  - FOLLOW: to follow the links in that webpage
  - NOFOLLOW: do not follow the links in that webpage
  - INDEX: to index that webpage
  - NOINDEX: do not index that webpage
- FOLLOW and INDEX are actually redundant since it is the job of crawlers to do so!
  - It is more for documentation

Summary of Robot Exclusion

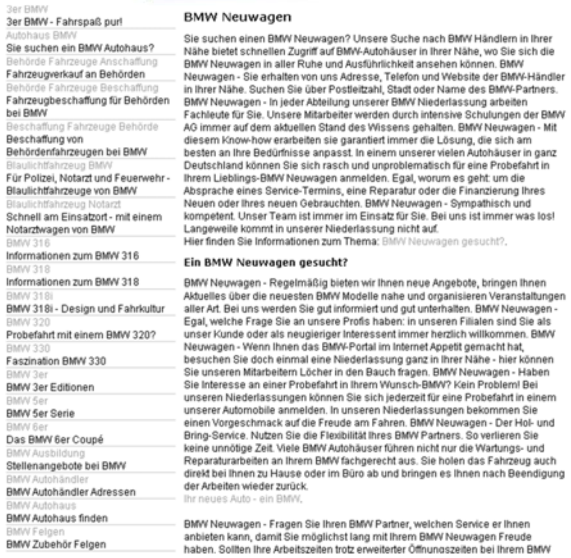
- robots.txt is well-established, understood & accepted, and any "good" robot will respect its T&C
- META tag is a de-facto standard (1999) but less well-adopted. It is also described in the HTML 4.01 specification, Appendix B.4.1.
- "*The instructions in robots.txt files cannot enforce crawler behavior to your site; instead, these instructions act as directives to the crawlers accessing your site…*" Google
- Some crawlers disguise themselves
  - Using false User-Agent
  - Randomizing access frequency to look like a human/browser- click fraud for ads

Servers can also deceive robots

Servers can disguise themselves, too

- Cloaking: Server will present different content based on User-Agent
  - E.g. push keywords, tags that will boost your SERP
- During a user search, the website will appear high on the SERP.
- When the user click, they will be show an entirely different page from the one fetched by the crawler
- This act is called "spamdexing" – classic case of bmw.de made the news

Spamdexing



3er BMW
3er BMW - Fahrspaß pur!
Autohaus BMW
Sie suchen ein BMW Autohaus?
Behörde Fahrzeuge Anschaffung
Fahrzeugverkauf an Behörden
Behörde Fahrzeuge Beschaffung
Fahrzeugbeschaffung für Behörden bei BMW
Beschaffung Fahrzeuge Behörde
Beschaffung von Behördenfahrzeugen bei BMW
Blaulichtfahrzeug BMW
Für Polizei, Notarzt und Feuerwehr - Blaulichtfahrzeuge von BMW
Blaulichtfahrzeug Notarzt
Schnell am Einsatzort - mit einem Notarztwagen von BMW
BMW 316
Informationen zum BMW 316
BMW 318
Informationen zum BMW 318
BMW 318i
BMW 318i - Design und Fahrkultur
BMW 320
Probefahrt mit einem BMW 320?
BMW 330
Faszination BMW 330
BMW 3er
BMW 3er Editionen
BMW 5er
BMW 5er Serie
BMW 6er
Das BMW 6er Coupé
BMW Ausbildung
Stellenangebote bei BMW
BMW Autohändler
BMW Autohändler Adressen
BMW Autohaus
BMW Autohaus finden
BMW Felgen
BMW Zubehör Felgen

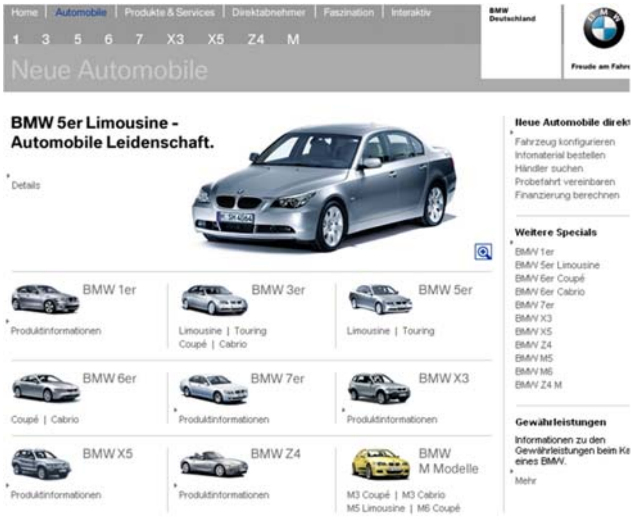
BMW Neuwagen

Sie suchen einen BMW Neuwagen? Unsere Suche nach BMW Händlern in Ihrer Nähe bietet schnellen Zugriff auf BMW-Autohäuser in Ihrer Nähe, wo Sie sich die BMW Neuwagen in aller Ruhe und Ausführlichkeit ansehen können. BMW Neuwagen - Sie erhalten von uns Adresse, Telefon und Website der BMW-Händler in Ihrer Nähe. Suchen Sie über Postleitzahl, Stadt oder Name des BMW-Partners. BMW Neuwagen - In jeder Abteilung unserer BMW Niederlassung arbeiten Fachleute für Sie. Unsere Mitarbeiter werden durch intensive Schulungen der BMW AG immer auf dem aktuellen Stand des Wissens gehalten. BMW Neuwagen - Mit diesem Know-how erarbeiten sie garantiert immer die Lösung, die sich am besten an Ihre Bedürfnisse anpasst. In einem unserer vielen Autohäuser in ganz Deutschland können Sie sich rasch und unproblematisch für eine Probefahrt in Ihrem Lieblings-BMW Neuwagen anmelden. Egal, worum es geht: um die Absprache eines Service-Termins, eine Reparatur oder die Finanzierung Ihres Neuen oder Ihres neuen Gebrauchten. BMW Neuwagen - Sympathisch und kompetent. Unser Team ist immer im Einsatz für Sie. Bei uns ist immer was los! Langeweile kommt in unserer Niederlassung nicht auf.
Hier finden Sie Informationen zum Thema: BMW Neuwagen gesucht?.

Ein BMW Neuwagen gesucht?

BMW Neuwagen - Regelmäßig bieten wir Ihnen neue Angebote, bringen Ihnen Aktuelles über die neuesten BMW Modelle nahe und organisieren Veranstaltungen aller Art. Bei uns werden Sie gut informiert und gut unterhalten. BMW Neuwagen - Egal, welche Frage Sie an unsere Profis haben: in unseren Filialen sind Sie als unser Kunde oder als neugieriger Interessent immer herzlich willkommen. BMW Neuwagen - Wenn Ihnen das BMW-Portal im Internet Appetit gemacht hat, besuchen Sie doch einmal eine Niederlassung ganz in Ihrer Nähe - hier können Sie unseren Mitarbeitern Löcher in den Bauch fragen. BMW Neuwagen - Haben Sie Interesse an einer Probefahrt in Ihrem Wunsch-BMW? Kein Problem! Bei unseren Niederlassungen können Sie sich jederzeit für eine Probefahrt in einem unserer Automobile anmelden. In unseren Niederlassungen bekommen Sie einen Vorgeschmack auf die Freude am Fahren. BMW Neuwagen - Der Hol- und Bring-Service. Nutzen Sie die Flexibilität Ihres BMW Partners. So verlieren Sie keine unnötige Zeit. Viele BMW Autohäuser führen nicht nur die Wartungs- und Reparaturarbeiten an Ihrem BMW fachgerecht aus. Sie holen das Fahrzeug auch direkt bei Ihnen zu Hause oder im Büro ab und bringen es Ihnen nach Beendigung der Arbeiten wieder zurück.
Ihr neues Auto - ein BMW.

BMW Neuwagen - Fragen Sie Ihren BMW Partner, welchen Service er Ihnen anbieten kann, damit Sie möglichst lang mit Ihrem BMW Neuwagen Freude haben. Sollten Ihre Arbeitszeiten trotz erweiterter Öffnungszeiten bei Ihrem BMW



Home | Automobile | Produkte & Services | Direktabnehmer | Faszination | Interaktiv

1 3 5 6 7 X3 X5 Z4 M

Neue Automobile

BMW Deutschland

Freude am Fahr

BMW 5er Limousine - Automobile Leidenschaft.

Details

BMW 1er	BMW 3er	BMW 5er
Produktinformationen	Limousine | Touring Coupé | Cabrio	Limousine | Touring
BMW 6er	**BMW 7er**	**BMW X3**
Coupé | Cabrio	Produktinformationen	Produktinformationen
BMW X5	**BMW Z4**	**BMW M Modelle**
Produktinformationen	Produktinformationen	M3 Coupé | M3 Cabrio M5 Limousine | M6 Coupé

Neue Automobile direkt

Fahrzeug konfigurieren
Infomaterial bestellen
Händler suchen
Probefahrt vereinbaren
Finanzierung berechnen

Weitere Specials

BMW 1er
BMW 5er Limousine
BMW 6er Coupé
BMW 6er Cabrio
BMW 7er
BMW X3
BMW X5
BMW Z4
BMW M5
BMW M6
BMW Z4 M

Gewährleistungen

Informationen zu den Gewährleistungen beim Ka eines BMW.

Mehr

Some websites are just difficult to crawl

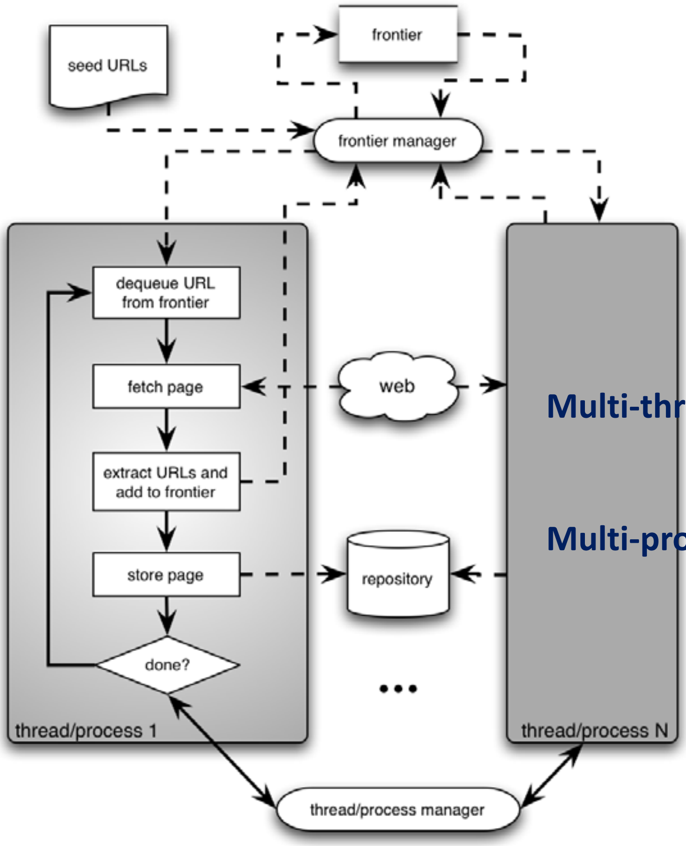
These sites are also called Deep or Hidden Web

- **Private Sites** that require login and password to limit access only to authorized people
- **Form Results** that require entering specific data before the data is served. Example Flight or ticket booking.
- **Scripted Pages** like JavaScript or Flash. Crawling gets slowed down because the scripts need to be executed before the data is served.

Concurrent Crawling

- A single crawler is not an efficient way to index the world wide web (1.9 billion websites!) or even a small portion of it.
- Crawlers inherently incur delays:
  - Resolving the host name in the URL to an IP address using DNS
  - Connecting a socket to the server and sending the request
  - Receiving the requested page in response
- The solution is to have **Concurrent** crawling to overlap the above delays by fetching many pages concurrently

Architecture of a Concurrent Crawler



Multi-threading

Multi-processing

Concurrent crawlers

- Can use multi-processing or multi-threading
- Each process or thread works like a sequential crawler, except they share data structures: frontier and repository
- Shared data structures must be synchronized (locked for concurrent writes)
- Speedup of factor of 5-10 are easy this way

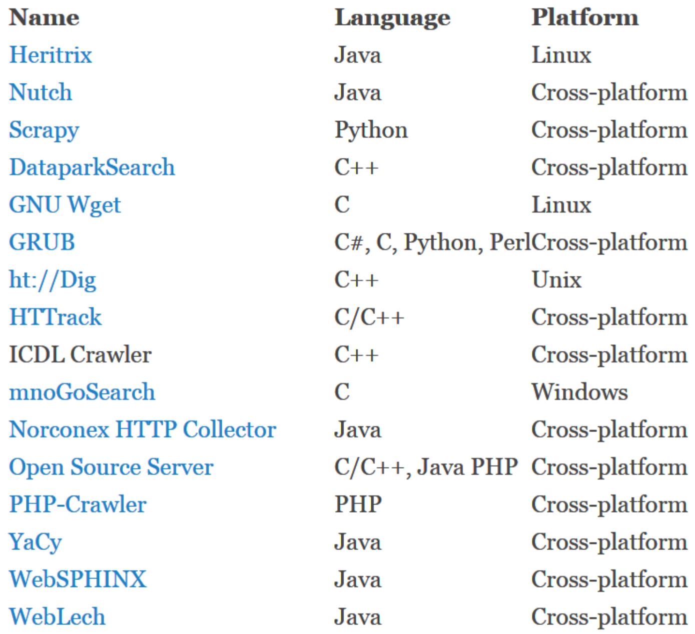
Industry best practice on Crawling

To prevent crawlers from being a threat to websites, companies practice the following crawling policies.

1. **Selection Policy:** download only “relevant” pages
2. **Re-visit Policy:** balance between Freshness and Age of pages in repository
3. **Politeness Policy:** avoid overloading Web sites by following robot exclusion instructions
4. **Parallelization Policy:** coordinate distributed Web crawlers to avoid duplicate downloads

Open Source Crawlers

- http://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining/



Name	Language	Platform
Heritrix	Java	Linux
Nutch	Java	Cross-platform
Scrapy	Python	Cross-platform
DataparkSearch	C++	Cross-platform
GNU Wget	C	Linux
GRUB	C#, C, Python, Perl	Cross-platform
ht://Dig	C++	Unix
HTTrack	C/C++	Cross-platform
ICDL Crawler	C++	Cross-platform
mnoGoSearch	C	Windows
Norconex HTTP Collector	Java	Cross-platform
Open Source Server	C/C++, Java PHP	Cross-platform
PHP-Crawler	PHP	Cross-platform
YaCy	Java	Cross-platform
WebSPHINX	Java	Cross-platform
WebLech	Java	Cross-platform

Summary

Last