Web Crawling

Introduction to Information Retrieval
INF 141/ CS 121
Donald J. Patterson

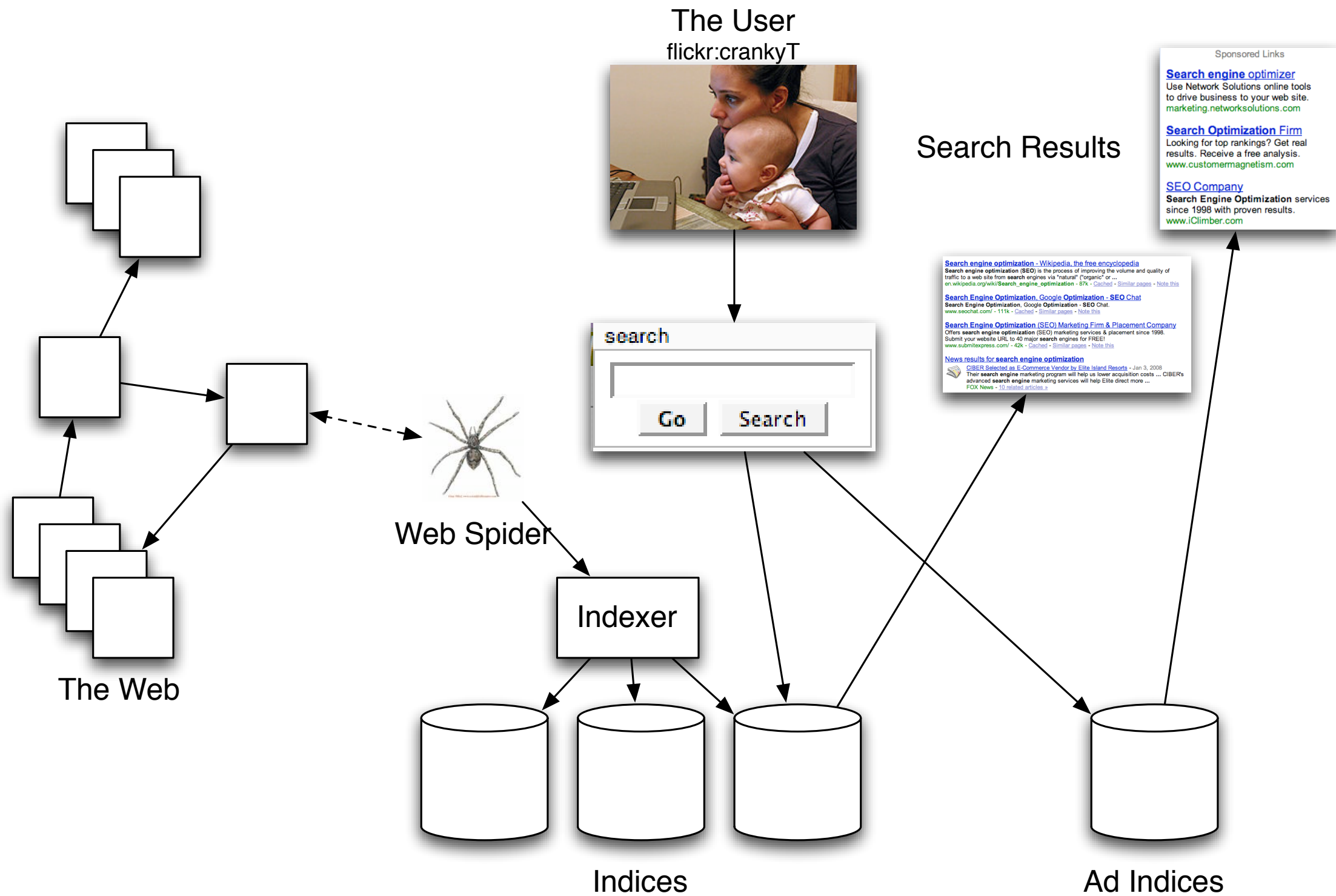Content adapted from Hinrich Schütze
http://www.informationretrieval.org

Overview

- Introduction

- URL Frontier
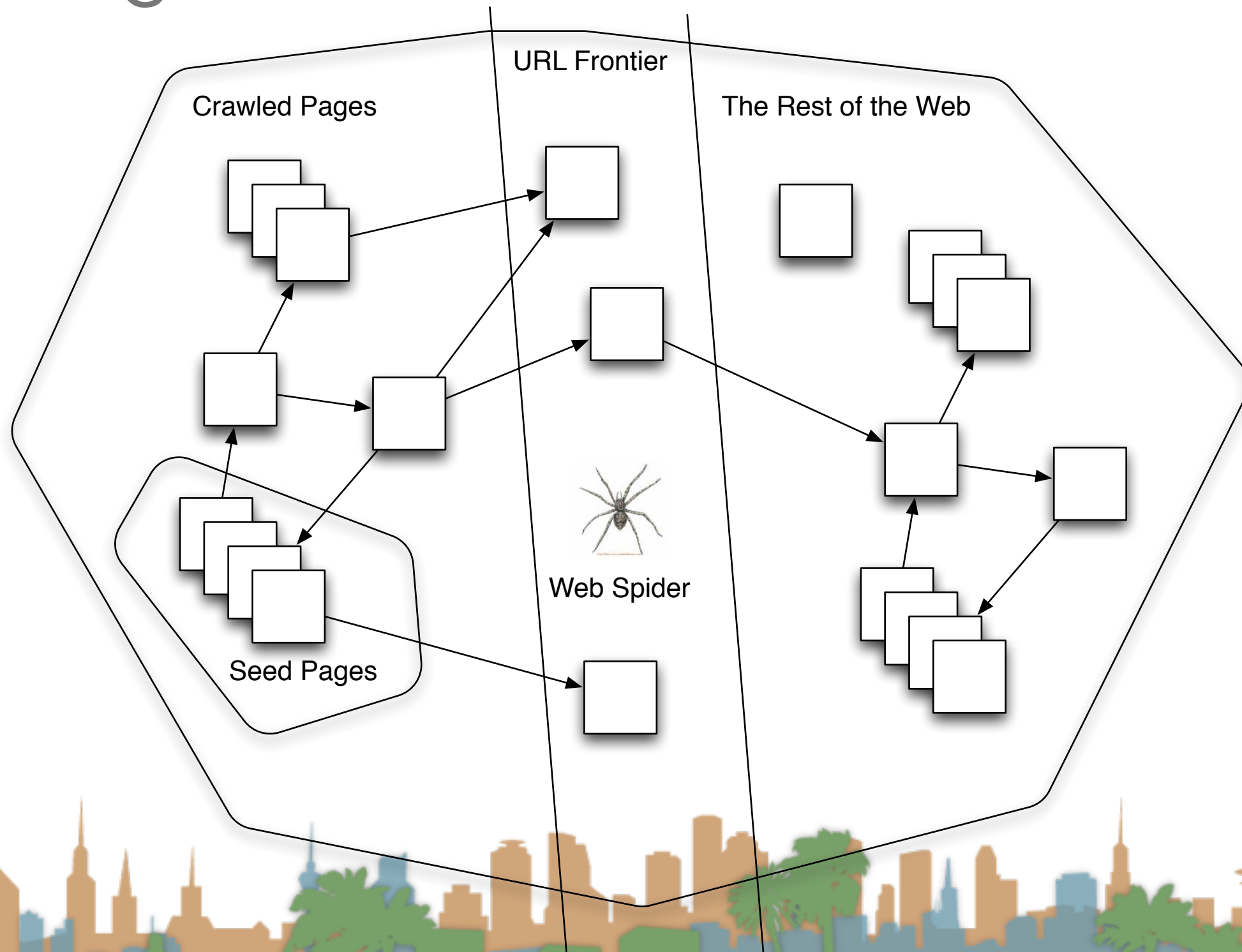
- Robust Crawling

  - DNS

The basic crawl algorithm

- Initialize a queue of URLs ("seed" URLs)

- Repeat

  - Remove a URL from the queue

  - Fetch associated page

  - Parse and analyze page

  - Store representation of page

  - Extract URLs from page and add to queue

Crawling the web

URL Frontier

Crawled Pages

The Rest of the Web

Web Spider

Seed Pages

Basic Algorithm is not reality...

- Real web crawling requires multiple machines

  - All steps distributed on different computers

- Even Non-Adversarial pages pose problems

  - Latency and bandwidth to remote servers vary

  - Webmasters have opinions about crawling their turf

    - How "deep" in a URL should you go?

  - Site mirrors and duplicate pages

- Politeness

  - Don't hit a server too often

Basic Algorithm is not reality...

- Adversarial Web Pages

  - Spam Pages

  - Spider Traps (http://www.devin.com/sugarplum/)

    - Find crawlers that don't respect robots.txt files

      - spammers

    - Griefing

   - Naturally occurring

      - Calendars

Minimum Characteristics for a Web Crawler

- Be Polite:

  - Respect implicit and explicit terms on website

  - Crawl pages you're allowed to

  - Respect "robots.txt" (more on this coming up)

- Be Robust

  - Handle traps and spam gracefully

Desired Characteristics for a Web Crawler

- Be a distributed systems

  - Run on multiple machines

- Be scalable

  - Adding more machines allows you to crawl faster

- Be Efficient

  - Fully utilize available processing and bandwidth

- Focus on "Quality" Pages

 - Crawl good information first
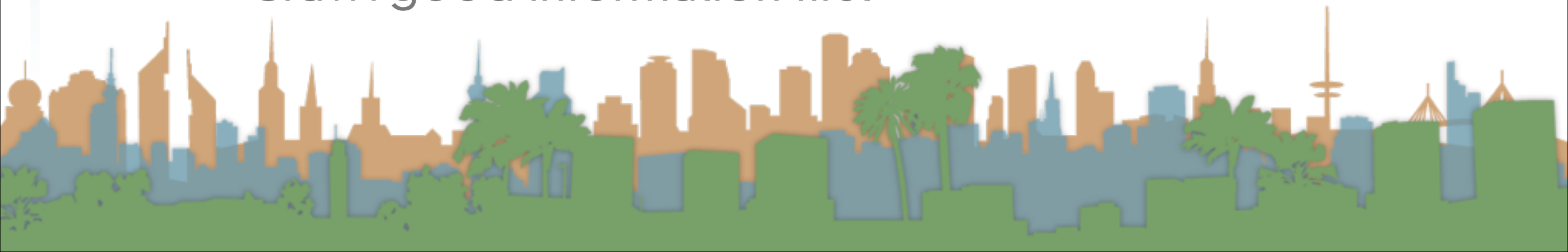
Desired Characteristics for a Web Crawler

- Support Continuous Operation

  - Fetch fresh copies of previously crawled pages

- Be Extensible

  - Be able to adapt to new data formats, protocols, etc.

  - Today it's AJAX, tomorrow it's HTML6, then....

Updated Crawling picture

URL Frontier
"Priority Queue"

Crawled Pages

The Rest of the Web

Seed Pages

Spider
Thread
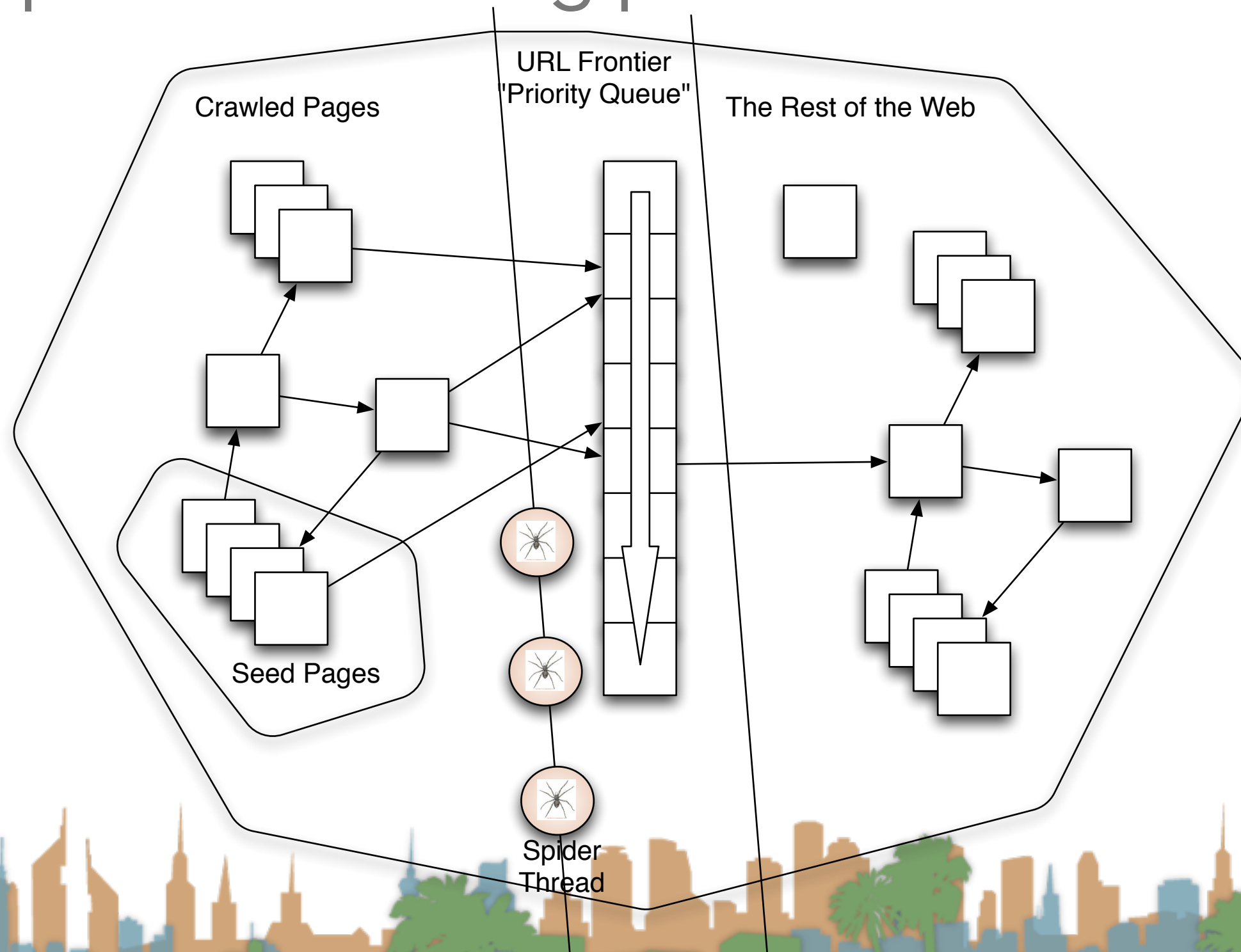
URL Frontier

- Frontier Queue might have multiple pages from the same host
  - These need to be load balanced ("politeness")
- All crawl threads should be kept busy

Politeness?

- It is easy enough for a website to block a crawler

- Explicit Politeness

  - "Robots Exclusion Standard"

    - Defined by a "robots.txt" file maintained by a webmaster

    - What portions of the site can be crawled.

      - Irrelevant, private or other data excluded.

    - Voluntary compliance by crawlers.

    - Based on regular expression matching

Politeness?

- Explicit Politeness

  - "Sitemaps"

    - Introduced by Google, but open standard

    - XML based

    - Allows webmasters to give hints to web crawlers:

      - Location of pages (URL islands)

      - Relative importance of pages

      - Update frequency of pages

    - Sitemap location listed in robots.txt

Politeness?

- Implicit Politeness

  - Even without specification avoid hitting any site too often

  - It costs bandwidth and computing resources for host.

URL Frontier

Politeness?

Last Update: 14 Jan 2008 - 02:59

Reported period: - Year - ▾ 2007 ▾ OK

speakeasy

Back to main page

Robots/Spiders visitors

30 different robots	Hits	Bandwidth	Last visit
Googlebot	1393868+104	5.11 GB	31 Dec 2007 - 23:50
Inktomi Slurp	36668+221	554.25 MB	31 Dec 2007 - 23:55
MSNBot	19522+2	699.90 MB	28 Dec 2007 - 08:01
Unknown robot (identified by 'crawl')	15949+13	89.34 MB	31 Dec 2007 - 22:24
AskJeeves	7016+1	106.29 MB	31 Dec 2007 - 23:49
Google AdSense	2701	100.26 MB	31 Dec 2007 - 22:10
psbot	2268+1	80.48 MB	31 Dec 2007 - 09:59
Unknown robot (identified by 'robot')	930+1	19.10 MB	31 Dec 2007 - 09:34
Turn It In	350+1	6.32 MB	03 Sep 2007 - 15:44
BaiDuSpider	300	10.22 MB	26 Nov 2007 - 07:32
GigaBot	243	5.27 MB	30 Dec 2007 - 05:06
Scooter	90+3	288.75 KB	27 Nov 2007 - 14:30
PhpDig	91	2.28 MB	21 Oct 2007 - 09:51
WISENutbot	76	1.94 MB	13 Jan 2007 - 14:04
Magpie	25	43.48 KB	24 Dec 2007 - 00:51
Unknown robot (identified by hit on 'robots.txt')	0+16	4.38 KB	14 Nov 2007 - 03:43
EchO!	14	287.09 KB	27 Dec 2007 - 13:56
Internet Shinchakubin	13	385.03 KB	27 Nov 2007 - 15:23
BBot	10	146.35 KB	13 Jun 2007 - 15:17
arks	8	142.24 KB	27 Nov 2007 - 12:25
MSIECrawler	8	263.02 KB	26 Dec 2007 - 11:16

Politeness?

Monthly history
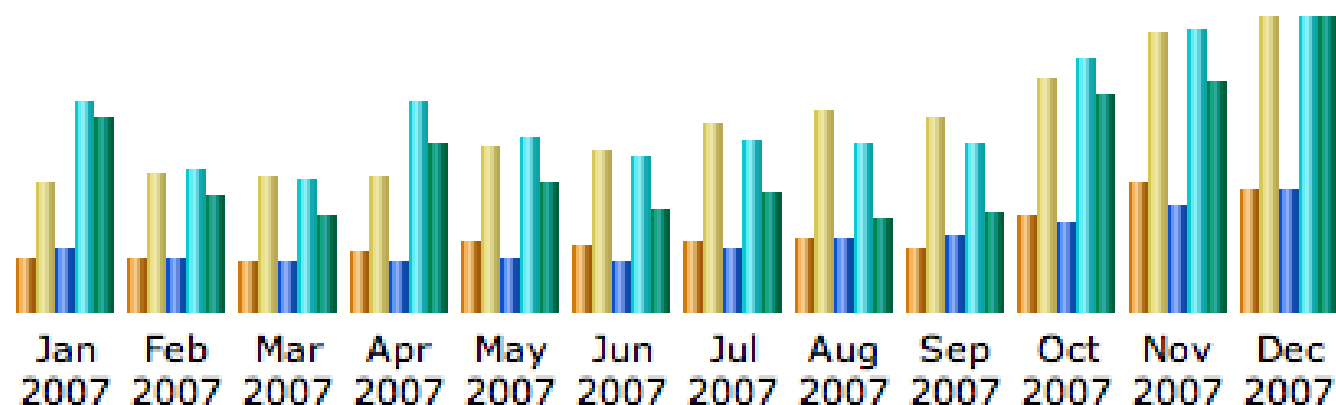
Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Jan 2007	1221	2946	8938	30536	699.28 MB
Feb 2007	1179	3099	7852	20475	415.75 MB
Mar 2007	1120	3063	7099	18978	350.88 MB
Apr 2007	1362	3067	7175	30320	599.91 MB
May 2007	1612	3746	7584	25114	469.32 MB
Jun 2007	1474	3662	7138	22292	370.11 MB
Jul 2007	1592	4210	9165	24766	430.61 MB
Aug 2007	1658	4567	10600	24142	336.08 MB
Sep 2007	1458	4403	11149	24414	356.60 MB
Oct 2007	2148	5299	12877	36427	783.78 MB
Nov 2007	2890	6317	15300	40487	833.75 MB
Dec 2007	2748	6631	17553	42281	1.03 GB
Total	20462	51010	122430	340232	6.55 GB

Politeness?

Last Update: 20 Jan 2009 - 03:14

Reported period: - Year - 2008 OK

speakeasy

Back to main page

Robots/Spiders visitors			
30 different robots	Hits	Bandwidth	Last visit
Googlebot	110194+158	1.95 GB	31 Dec 2008 - 23:43
Inktomi Slurp	13689+156	244.14 MB	31 Dec 2008 - 23:38
AskJeeves	6344+3	113.74 MB	30 Dec 2008 - 02:55
MSNBot	5014+11	146.12 MB	31 Dec 2008 - 22:36
Unknown robot (identified by 'crawl')	4302+5	47.72 MB	31 Dec 2008 - 03:04
Google AdSense	3005	112.90 MB	31 Dec 2008 - 21:14
Unknown robot (identified by 'robot')	2518+138	89.74 MB	31 Dec 2008 - 15:15
psbot	1263	27.87 MB	17 Dec 2008 - 07:56
GigaBot	261	7.14 MB	30 Dec 2008 - 18:05
Unknown robot (identified by 'spider')	201	9.23 MB	29 Dec 2008 - 03:30
Turn It In	178+1	3.34 MB	24 Nov 2008 - 13:57
BBot	84	1.65 MB	24 Nov 2008 - 19:42
The Python Robot	74+3	3.12 MB	31 Aug 2008 - 11:58
Unknown robot (identified by hit on 'robots.txt')	0+54	15.04 KB	28 Dec 2008 - 19:38
EchO!	48	1.14 MB	28 Nov 2008 - 23:02
Magpie	43	21.74 KB	17 Nov 2008 - 11:16
MSIECrawler	42	561.59 KB	01 Dec 2008 - 23:41
arks	39	438.47 KB	21 Dec 2008 - 04:30
SurveyBot	21	14.07 KB	29 Dec 2008 - 23:18
Internet Shinchakubin	17	267.64 KB	26 Aug 2008 - 06:55
Scooter	8	16 Bytes	24 Apr 2008 - 00:14
BaiDuSpider	7	204.01 KB	25 Apr 2008 - 19:46
Fish search	6	34.53 KB	08 Apr 2008 - 15:46

ly history

Apr 2008 May 2008 Jun 2008 Jul 2008 Aug 2008 Sep 2008 Oct 2008 Nov 2008 Dec 2008
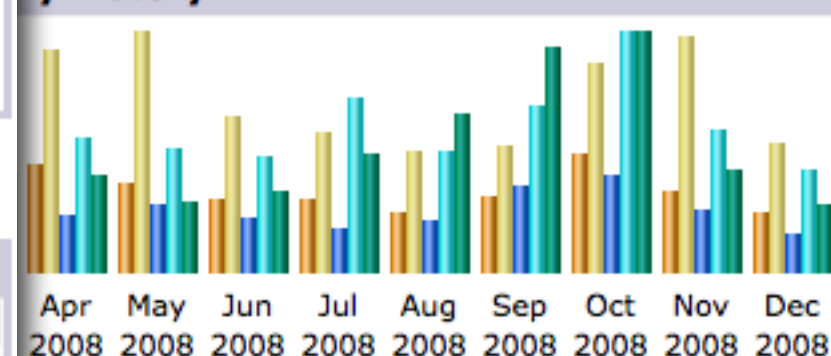
Number of visits	Pages	Hits	Bandwidth
7047	17437	40911	1.03 GB
8477	18878	47929	1.08 GB
9001	22892	57030	1.32 GB
11510	26871	61373	1.29 GB
12405	31227	57028	947.39 MB
8088	25365	53173	1.08 GB
7275	20348	79872	1.56 GB
6212	24236	55827	2.11 GB
6593	40161	76103	3.00 GB
10887	45152	110285	3.19 GB
12143	29037	65081	1.36 GB
6724	17716	47438	924.81 MB
106362	319320	752050	18.84 GB

Robots.txt - Exclusion

- Protocol for giving spiders ("robots") limited access to a website

  - Source: http://www.robotstxt.org/wc/norobots.html

- Website announces what is okay and not okay to crawl:

  - Located at http://www.myurl.com/robots.txt

  - This file holds the restrictions

Robots.txt Example

- http://www.ics.uci.edu/robots.txt

```
User-agent: MOMspider              # The Multi-Owner Maintenance Spider
Disallow: /cgi-bin/                #      Script files
Disallow: /Admin/MOM/              #      Local MOMspider output
Disallow: /~fielding/MOM/          #      Local MOMspider output
Disallow: /TR/                     #      Dienst Technical Report Server
Disallow: /Server/                 #      Dienst Technical Report Server
Disallow: /Document/               #      Dienst Technical Report Server
Disallow: /MetaServer/             #      Dienst Technical Report Server
Disallow: /~eppstein/pubs/cites/        #      Eppstein Database
Disallow: /~fiorello/pvt/          #      Private pages

User-agent: *                      # All other spiders should avoid
Disallow: /cgi-bin/                #      Script files
Disallow: /Test/                   #      The test area for web experimentation
Disallow: /Admin/                  #      Huge server statistic logs
Disallow: /TR/                     #      Dienst Technical Report Server
Disallow: /Server/                 #      Dienst Technical Report Server
Disallow: /Document/               #      Dienst Technical Report Server
Disallow: /MetaServer/             #      Dienst Technical Report Server
Disallow: /~fielding/MOM/          #      Local MOMspider output
Disallow: /~kanderso/hidden        #      Ken Anderson's stuff
Disallow: /~eppstein/pubs/cites/        #      Eppstein Database
Disallow: /~fiorello/pvt/          #      Private pages
Disallow: /~dean/
Disallow: /~wwwoffic/
Disallow: /~ucounsel/
Disallow: /~sao/
Disallow: /~support/
Disallow: /~icsdb/
Disallow: /bin/
```
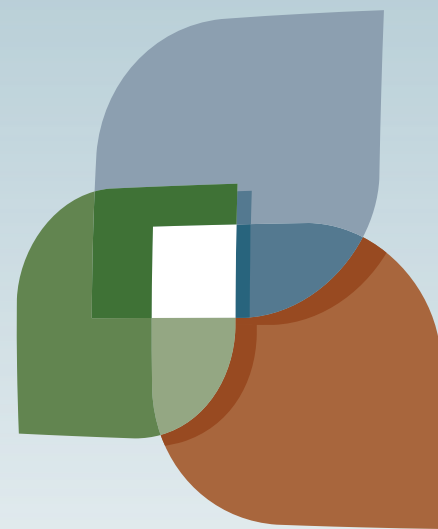
Sitemaps - Inclusion

- https://www.google.com/webmasters/tools/docs/en/protocol.html#sitemapXMLExample

```xml
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

    <url>
        <loc>http://www.example.com/</loc>
        <lastmod>2005-01-01</lastmod>
        <changefreq>monthly</changefreq>
        <priority>0.8</priority>
    </url>
    <url>
        <loc>http://www.example.com/catalog?item=12&amp;desc=vacation_hawaii</loc>
        <changefreq>weekly</changefreq>
    </url>
    <url>
        <loc>http://www.example.com/catalog?item=73&amp;desc=vacation_new_zealand</loc>
        <lastmod>2004-12-23</lastmod>
        <changefreq>weekly</changefreq>
    </url>
    <url>
        <loc>http://www.example.com/catalog?item=74&amp;desc=vacation_newfoundland</loc>
        <lastmod>2004-12-23T18:00:15+00:00</lastmod>
        <priority>0.3</priority>
    </url>
    <url>
        <loc>http://www.example.com/catalog?item=83&amp;desc=vacation_usa</loc>
        <lastmod>2004-11-23</lastmod>
    </url>
</urlset>
```