Web Crawling
Introduction to Information Retrieval
Informatics 141 / CS 121
Donald J. Patterson

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

- Introduction
- URL Frontier
- Robust Crawling
  - DNS
Introduction
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

Introduction

Seed Pages

Web Spider

Crawled Pages

URL Frontier

The Rest of the Web

The Rest of the Web
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
Introduction

Basic Algorithm is not reality...

- Adversarial Web Pages
- Spam Pages
- Spider Traps
Introduction

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 SilverLight, then....
Updated Crawling picture

Seed Pages
Spider Thread
Crawled Pages
URL Frontier
"Priority Queue"
The Rest of the Web
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 specify:
    • Location of pages (URL islands)
    • 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.
Politeness?

URL Frontier

**Monthly history**

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