



Collection Building on the Web

CS 510 Winter 2007

1



Overview

- Collection architecture
- Collection goals
- What to collect
- How to collect it
- How often to re-collect it

CS 510 Winter 2007

2



Basic Algorithm

Initialize URL stack

While more URLs

 If URL is not a duplicate

 Get document with URL

 Add to database

 Extract URLs, add to stack

CS 510 Winter 2007

3



Collection Goals

- Completeness
- Topicality
- Quality
- Freshness
- Low duplication

CS 510 Winter 2007

4



Collection decisions

- Which pages to download
 - Interest-driven
 - Popularity-driven
 - Location-driven
- How should the crawler refresh pages
 - What are we trying to maximize
- How do we avoid over-taxing a site when it's crawled?

CS 510 Winter 2007

5



What to collect

- What are your users trying to find?
 - Pages
 - Sites
- What are the best pages?
 - Least popular/most popular
 - Most common/most unique
 - Most links/fewest links
 - Most commonly accessed/least commonly accessed

CS 510 Winter 2007

6



Starting Set

- Specific set of starting points
 - Enables some topical collection building
 - Comparable to offline page access
- Re-use site roots
 - Facilitates structural changes in the site
 - May lose “back door” access
- Re-use all links
 - May have many misses on sites with structural changes

CS 510 Winter 2007

7



Duplicate Detection

- How do you keep from retrieving the same link twice?
- Examples
 - <http://www.amazon.com/page.html> and <http://amazon.com/page.html>
 - <http://www.intel.com/page.html> and <http://www.intel.com:80/page.html>
 - <http://ww1.ibm.com/page.html> and <http://ww2.ibm.com/page.html>
 - <http://www.transarc.com/afs/tr/page.html> and <file://localhost/afs/tr/page.html>
- Approaches
 - Don't look for duplicates
 - Create canonical URL representation
 - Compare contents (e.g. with checksum)

CS 510 Winter 2007

8



Error Handling

- What happens when the referenced page isn't there?
- Approaches
 - Grab it from Alexa
 - Ignore it and remove it from the collection
 - Ignore it and keep the previously cached copy

CS 510 Winter 2007

9



Database Access

- How do you collect pages automatically generated from a database?
 - E.g. Amazon books
- Basic approach
 - Iterative query generation
 - Use domain knowledge to fill out forms
 - Ideal is “largest bounding” queries

CS 510 Winter 2007

10



How Often to Re-collect

- Assume
 - p1 changes 9 times a day
 - p2 changes 1 time a day
 - We can update one page per day
- What policy ensures the best collection?

For more information see: "Synchronizing a database to Improve Freshness" by Cho and Garcia-Molina

CS 510 Winter 2007

11



Definitions

- Freshness: the fraction of the local collection that is up-to-date
- Age: the cumulative time that the collection has been out-of-date
- For example
 - P1 changed 2 days ago
 - P2 changed 5 days ago
 - The collection was built 3 days ago
 - The freshness is $\frac{1}{2}$ (P2 is up-to-date)
 - The age is 2 (P1 is 2 days out-of-date)

CS 510 Winter 2007

12



Policies

- Frequency is the rate of change
- Uniform policy—all pages retrieved at the same rate
- Proportional policy—pages that change more frequently are retrieved more frequently

CS 510 Winter 2007

13



Refresh Rates

row	$f_1 + f_2$	f_1	f_2	benefit	best
(a)	1	1	0	$\frac{1}{2} \times \frac{1}{18} = \frac{1}{36}$	0 1
(b)		0	1	$\frac{1}{2} \times \frac{1}{2} = \frac{9}{36}$	
(c)	2	2	0	$\frac{1}{2} \times \frac{1}{18} + \frac{1}{2} \times \frac{1}{18} = \frac{2}{36}$	0 2
(d)		1	1	$\frac{1}{2} \times \frac{1}{18} + \frac{1}{2} \times \frac{1}{2} = \frac{10}{36}$	
(e)		0	2	$\frac{1}{3} \times \frac{2}{3} + \frac{1}{3} \times \frac{1}{3} = \frac{12}{36}$	
(f)	5	3	2	$\frac{3}{36} + \frac{12}{36} = \frac{30}{72}$	2 3
(g)		2	3	$\frac{2}{36} + \frac{6}{16} = \frac{31}{72}$	
(h)	10	9	1	$\frac{9}{36} + \frac{1}{4} = \frac{36}{72}$	7 3
(i)		7	3	$\frac{7}{36} + \frac{6}{16} = \frac{41}{72}$	
(j)		5	5	$\frac{5}{36} + \frac{15}{36} = \frac{40}{72}$	

CS 510 Winter 2007

14

Observations

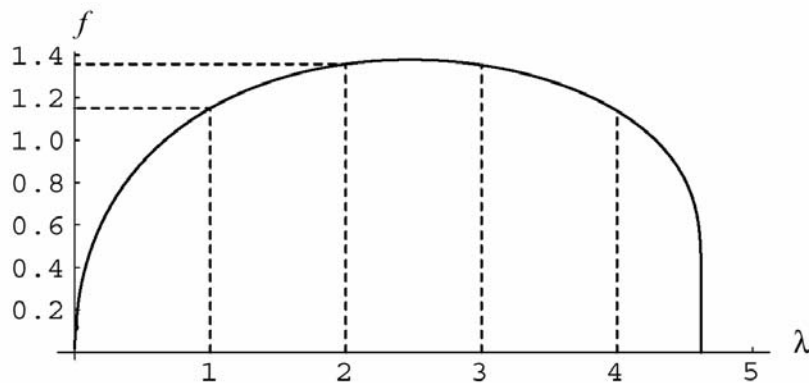
- Uniform method is ALWAYS better than proportional method for freshness
- Optimal refresh rate graph ALWAYS has the same shape

CS 510 Winter 2007

15

Freshness: Optimal Rate

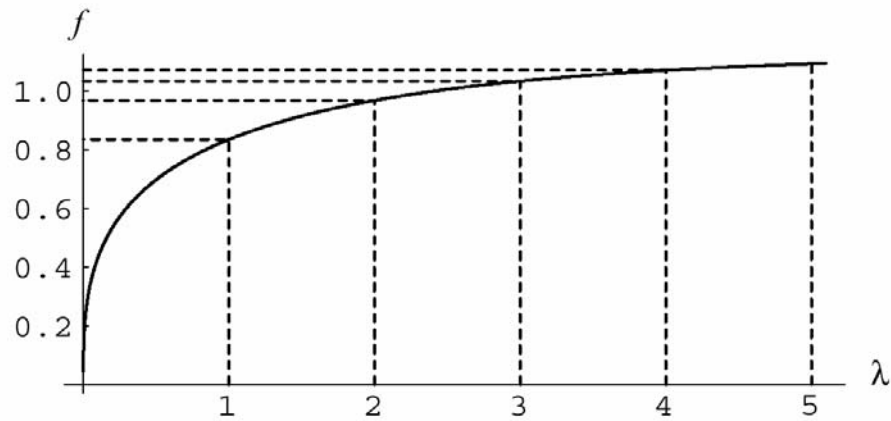
5 pages that change 1, 2, 3, 4, and 5 times a day



CS 510 Winter 2007

16

Age: Optimal Rate



CS 510 Winter 2007

17

Robot Exclusion File

<http://www.cat.pdx.edu/robots.txt>

```
User-agent: psu-gsa-crawler
Disallow: /
```

```
User-agent: *
Disallow: /administrator/
Disallow: /cache/
Disallow: /components/
Disallow: /editor/
Disallow: /help/
Disallow: /images/
Disallow: /includes/
Disallow: /language/
Disallow: /mambots/
Disallow: /media/
Disallow: /modules/
Disallow: /templates/
Disallow: /installation/
Disallow: /index.php?option=com_loudmouth
Disallow: /index.php?option=com_extcalendar
```

CS 510 Winter 2007

18