Re-Crawling and characteristics of change for Web pages

1

Crawling **large** number pages

- indexing is **not** dynamic and continuous
  - Index all pages collected at certain time (end of crawl?)
  - Provide search half of engine with new index
- crawling is continuous
  - start over
    - in some sense

2

Re-crawling

- When re-crawl what pages?
  - finish crawl and start over
    - finish = have enough?
  - re-crawl high priority pages in middle of crawl
  - how determine priority?
- How integrate re-crawl of high priority pages?
  - One choice – separate cycle for crawl of high priority pages

3

How re-crawl the Web?

- want to understand how Web grows and changes
- Many studies since late 1990's
- Hard to compare
  - different techniques for sampling
  - different times
    - 2000 vs 2008?

4

Sample of Results

studies from 2002 - 2006

What do we expect to change from 2002 to now?

5

Set-up Ntoulas et. al.

- Experiments Oct. 2002- Oct. 2003
- 154 Web sites
  - 5 top-ranked by PageRank from subset Google Directory
- weekly breadth-first crawl for 12 months
  - up to 200,000 pages per site
    - only 4 sites contained > 200,000 pages
  - average 4.4 million pages per week
- Published in *WWW* 2004

6

Set-up Kim et. al.

- Experiments Jan.-Mar. 2004
- 34,000 Korean sites
- 1.8 million URLs initially
- downloaded every 2 days for 100 days
- 2 million URLs total after all 50 “crawls”
- published in *ICCS* 2007

7

Set-up Fetterly et.al.

- Experiments Nov. 2002 – Jan. 2003
- first crawl from Yahoo.com giving 151million HTML pages
- try download pages 10 more times over next 10 weeks
- Published in *Software- Practice and Experience* 2004

8

Set-up Dontcheva et.al.

- Experiments June- Nov. 2006
- 100 Web pages from 24 “popular” Web sites by hand
- downloaded daily for 5 months
- U. Washington CSE Technical Report 2007

9

Set-up Olston et.al.

- Experiments approx 100 days in 2006
- Random sample:
  - 10,000 URLS from Yahoo crawled collection
  - download every two days
  - 50 snapshots total
- High-quality sample:
  - random sample 10,000 URLS from OpenDirectory
  - download every two days
  - 30 snapshots total
- Published in *WWW* 2008

10

Dynamics of Web pages

- average Web page size
  - 12KB [Ntoulas et al]
  - 66% between 4 and 32 KB [Fetterly et al]
- new pages per week
  - 8% [Ntoulas et al] ID by URL
- staying power of pages
  - Ntoulas ID pages by URL
    - 75% pages still exist after 1 months
    - 60% pages still exist after 6 months
    - 40% pages still exist after 12 months
    - page ½- life 9 months

11

Dynamics of Web links

- new links per week
  - 25% [Ntoulas et al]
- staying power of links
  - 24% initial links still seen after 12 months [Ntoulas et al]

12

Dynamics of Web page changes

- changes in pages
  - *any* change [Ntoulas et al] :
    - 50% unchanged in year
    - 15% of pages changed in week
  - *any* change [Kim et al]
    - 64% of "famous sites" unchanged in 100 days
    - 6.5% of "famous sites" change every download
    - 66% of "random sites" unchanged in 100 days
    - 3.25% of "random sites" change every download
    - gives detailed distribution

13

Dynamics of Web page changes

- changes in pages contents [Fetterly et. al.]
  - remove HTML mark-up
  - shingle content of pages
    - shingle size 5
  - make sketch from shingling
    - feature vector for each download of each page
  - conclude large documents change more often and more extensively
  - past change of a page good predictor for future
  - more detailed analysis

14

Dynamics of Web Page structure

[Dontschva et. al.]

- Look at structure of HTML DOM† tree
  - recall XML model
  - measure change in structure over consecutive days
- Results
  - many pages don't change much (agree Ntoulas et al)
  - little correlation between number of nodes in tree and amount of change
    - compare Fetterly et al size vs change
  - number of structural changes increased with
    - traffic volume
    - dynamic content
  - much more analysis

†Domain Object Model

15

Dynamics of Web content

- content across Web pages [Ntoulas et al]
  - remove HTML mark-up
  - shingle content of pages
    - shingle size 50 (size of paragraph)
    - 4.3 billion unique shingles
  - look at union of shingles of all pages
    call this the content
  - # of new shingles measures new content
  - new shingles per week average 5%
    (compare 8% new URLS)
  - 60% first week shingles still present after 12 months
    (compare 40% URLs)

16

Longevity of content [Olston et al]

- ephemeral content
  - changes *very* frequently
  - not worth indexing
  - Examples:
    - quote of day
    - advertisement
- versus persistent content
- look at shingles as fragments on page
- snapshots of pages over time (every 2 days)

17

Longevity of content cont.

- page change frequency: number of snapshots that differ from the previous snapshot
- information longevity: average lifetime of shingles on a page
  - average lifetime of a shingle = average number of contiguous (in time) snapshots in which shingle occurs
- Result: information longevity is not strongly correlated with change frequency
- much more analysis

18