Characteristics of change for Web pages

1

Evidence for re-crawling strategies

- 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 2012?

2

What study?

- How long Web pages live?
- How often Web pages change?
  - any change
  - content change
- How often new Web pages?
- How long Web links live?
- How often new Web links?
- How do these properties vary for different kinds of Web pages?
  - what kinds?

3

Sample: 6 studies

•most studies sample same set of pages
•some studies truly crawl

•sample rate ranges by minute to daily
•sample size ranges from 100 to 150,000,000

links to papers posted on *Schedule and Assignments* page

4

Results from 3 most recent studies

Highlights and Summaries

5

Set-up Adar et.al.

- Experiments 5 weeks starting May 24, 2007
- Sample:
  - source: logs of opt-in users of Live Search bar
  - sample: 54788 of URLs visited by 612,000 users over 5 weeks in August 2006.
    - sample over page characteristics - ranges:
      - # unique visitors (4 "bins")
      - avg. revisit time (6 "bins")
      - avg. # revisits per user (5 "bins")
    - plus 5000 most visited
- download hourly

- Published in *WSDM* 2009

6

Set-up Adar et.al.: supplemental

- Additional temporally fine-grained revisits
- Sample: the ~42% of retrieved pages that change "nearly every hour"
- download at 0 (hourly crawl), 2, 16, 32 min.
  - equals one batch
- 8 batches over 3 days
  - shift batches by 4 hours for time-of-day
- For the ~47% of above that change at 2 min
- request simultaneously on 2 synchronized machines
  - capture change every reload

7

Results [Adar et. al.] Changes in Page Contents

- measure of change:
  - term-based Dice measure of change for a document from time t1 to time t2:

$$Dice(W_{t1}, W_{t2}) = 2*|W_{t1} \cap W_{t2}| / (|W_{t1}|+|W_{t2}|)$$

  where W_t is the set of words in document at time t
- findings
  - 34% pages do not change (compare Ntoulas et. al.)
  - avg change among rest: every 123 hours
  - avg Dice coefficient of changes = 0.794
  - popular pages (>39 visitors) change more frequently:
    - avg change every 102 hours
  - popular pages not change more - avg Dice .8123

8

more changes in page contents [Adar et. al.]

- findings for temporally fine-grained revisits
  - 6.5% all pages always change in simult. download
  - 9% all pages change at every 2 min. point
  - 19.3% all pages change at one or more 2 min. point
  - 11.8% all pages change at every 32 min. point
  - 23.99% all pages change at one or more 32 min. point
- many more results relating to types of pages
  - example: .gov and .edu change more slowly that .com, .net, .org
  - example: pages at URL depth ≥5 change more slowly but change more at each once

9

Dynamics of Web Page structure

- Look at structure of HTML DOM† tree
- Results [Adar et. al.]
  - measure change in structure over consecutive hours
  - many many pages don't change much
    at 2 hrs: avg. of 99.3% DOM elements still in page
    at 5 wks: avg. of 84.3% DOM elements still in page
    at 5 wks: median of 99.8% DOM ele.s still in page!?

†Domain Object Model

10

More dynamics of Web page structure

- Results [Dontschva et. al.]
  - measure change in structure over consecutive days
  - many pages don't change much
  - 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

11

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

12

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

13

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)

14

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

15

Longevity of content: Adar et.al.

- Adar studies content by hash of text in each DOM element (i.e. text between html tags)
- can see position changes
  - where in DOM tree content changes
- no quantitative results

16

Earlier results

17

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

18

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

19

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]

20

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)

21

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

22

Dynamics of Web page changes

changes in page contents [Fetterly et. al.]

- measure of change:
  - remove HTML mark-up
  - shingle content of pages
    - shingle size 5
  - make sketch from shingling
    - feature vector for each download of each page
- findings
  - large documents change more often and more extensively
  - past change of page good predictor for future
  - more detailed analysis

23

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

24

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

25

How use characteristic in crawling?

- Keep page characteristics in URL data
- page re-crawl time based on char’istics
  - how often changing
    - what type of content changing
  - top-level domain
  - popular pages
  - large documents
  - others?

Other uses for characteristics?

26