Crawling results

Morten Goodwin Olsen

Current implementation status

- No sites crawled yet (:-()
- At best, crawled a site with 60 000 pages – infinite loop
- State machine fully implemented
  - Debugging and fixing – ongoing
- Current crawl (started this morning):
  The number of scenarios loaded is: 1223
  Duration: 0 h 40 m 32 s (2432 seconds)
  Seconds per page: 1.98903515937
  Pages per second: 0.502756321469

Last last scale testing

- site - number of restarts
- www.u-psud.fr - 242
- www.comune.bolzano.it - 335
- www.fingalcoco.ie - 396
- www.gobiernodecanarias.org - 379
- www.regione.vda.it - 392
- www.ag.ch - 409
- www.czestochowa.pl - 349

Remaining issues

- 1. Debug and make sure the crawler runs stable enough.
- 2. Run tests from sampling revision
  - How many crawls do we need etc.
  - How many WAMs to we run.
  - How many samples do we run.
- 3. Test rest of the Observatory

Finished test(s)

Open questions

- Still keep exhaustive scan?
-

(Nearly) Infinite loop

Suggestion 1 – Dom comparison

Extract DOM tree
Remove URL parameters
Check if current page is equal to e.g. at least 2 of the last 100
If so, don't extract links from current page

Advantage: Easy
Disadvantage: May produce false negatives. Not robust for page we do not know structure of.

Suggestion 2 – Bayesian comparison

Extract DOM tree and/or Text
Assume each tag is a property
Assume a threshold of e.g. 0.95
Check if current page is equal to e.g. at least 2 of the last 100.
If so, don't extract links from current page

Advantage: Robust – even for pages we do not know the structure of. No expert knowledge needed.
Disadvantage: May be slower than DOM. May produce false positives.

Result – When equal - Bayesian

Page 1: 0.000116243572867 equal: False
Page 2: 0.00650422686356 equal: False
Page 3: 0.00217023466918 equal: False
Page 4: 0.976367194825 equal: True
Page 5: 0.981048647154 equal: True
Page 6: 0.95328177479 equal: True
Page 7: 0.95328177479 equal: True
(all correct – to be discussed)

Result – When equal – Dom Comparison

Page 1: False
Page 2: False
Page 3: False
Page 4: True
Page 5: True
Page 6: False
Page 7: False
(two false negatives – to be discussed)

Result – When not equal - Bayesian

Page 1: 0.00760079532104 equal: False
Page 2: 0.306467606352 equal: False
Page 3: 0.00167741853722 equal: False
Page 4: 0.00167741853722 equal: False
Page 5: 0.00167741853722 equal: False
Page 6: 0.00165093792162 equal: False
Page 7: 0.00165093792162 equal: False
(all correct)

Result – When not equal – Dom comparison

Page 1: False
Page 2: False
Page 3: False
Page 4: False
Page 5: False
Page 6: False
Page 7: False
(all correct)