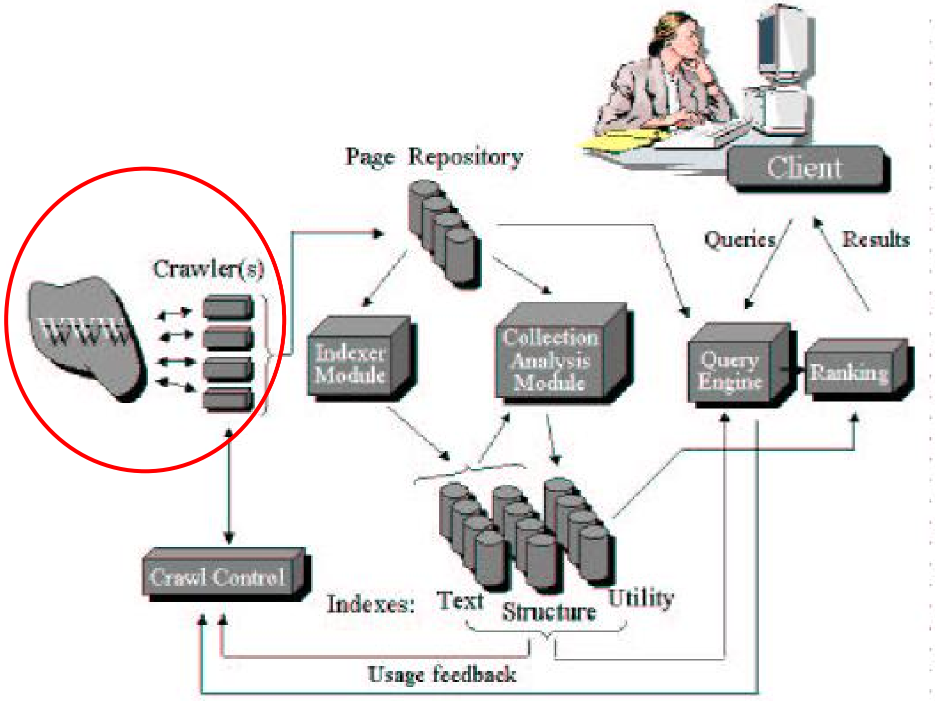
No Country for Old Web

Keeping Web Crawl data updated

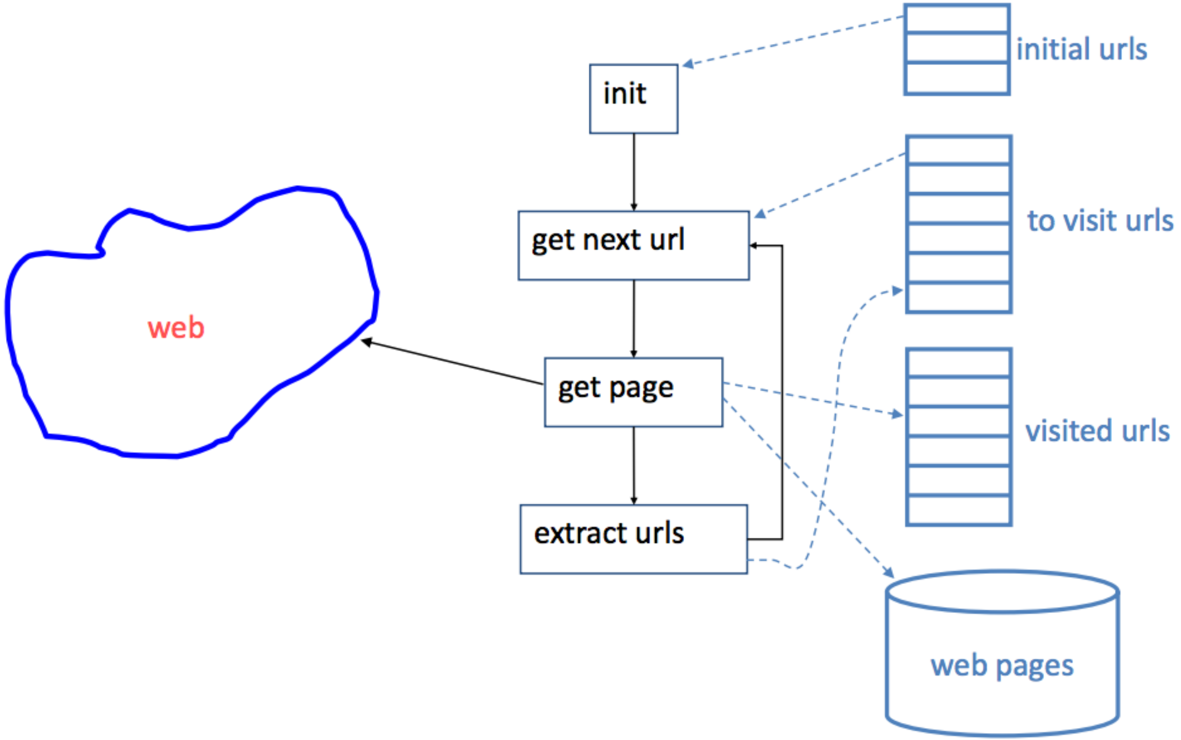
Primal Pappachan
06/05/2014

Searching the web



Arasu, Arvind, et al. "Searching the web." *ACM Transactions on Internet Technology (TOIT)* 1.1 (2001): 2-43.

What is a Crawler?



Web Crawling Challenges

1. What to download?
2. How to refresh pages?
3. How to be unobtrusive?
4. How to download faster?

Arasu, Arvind, et al. "Searching the web." ACM Transactions on Internet Technology (TOIT) 1.1 (2001): 2-43.

Web Crawling Challenges

1. What to download?
2. How to refresh pages?
3. How to be unobtrusive?
4. How to download faster?

Arasu, Arvind, et al. "Searching the web." ACM Transactions on Internet Technology (TOIT) 1.1 (2001): 2-43.

Adding the perspective

- 20 million pages indexed by AltaVista in 1995
- One trillion URLs known by Google / Yahoo in 2008
- Page has 10 - 100 KB of textual content
- Contains roughly 100 links per page

Crawle

“... a cra ction of time pages re

. and Weber, R. R. (1998)

1. Keep th h
2. Keep th

Crawler Design issues

Batch-mode Crawler	VS	Steady Crawler
Shadowing		In-place update
Fixed Frequency		Variable Frequency

Cho, Junghoo, and Hector Garcia-Molina. "The evolution of the web and implications for an incremental crawler." (1999).

Crawler Design issues

Batch-mode Crawler

Shadowing

Fixed Frequency

Advantages

- Easy to implement
- High availability of the collection

Cho, Junghoo, and Hector Garcia-Molina. "The evolution of the web and implications for an incremental crawler." (1999).

Crawler Design issues

Advantages

- High Freshness
- Less load on network / server

Steady Crawler

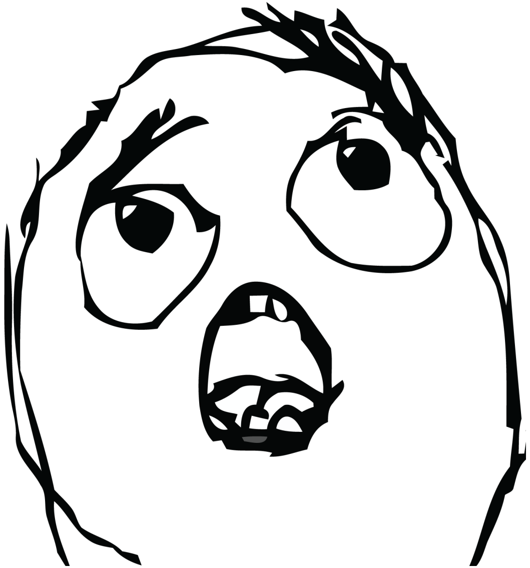
In-place update

Variable Frequency

Cho, Junghoo, and Hector Garcia-Molina. "The evolution of the Web and implications for an incremental crawler." (1999).

Re-visiting policies

- Uniform policy
- Proportional policy



Uniform Policy $>>_{\text{average freshness}}$ Proportional Policy

Cho, J. and Garcia-Molina, H. (2003). Effective page refresh policies for web crawlers. ACM Transactions on Database Systems.

Optimal Policy

- Penalize pages that change too often
- Neither uniform or proportional
- Vary access frequencies (sublinearly) with rate of change

Modelling page changes

1. Exponential distribution (Junghoo Cho; Hector Garcia-Molina (2003))
2. Statistical approach to discover parameters (Ipeirotis, P., Ntoulas, A., Cho, J., Gravano, L. (2005))

Junghoo Cho; Hector Garcia-Molina (2003). "Estimating frequency of change". . ACM Trans. Interet Technol.

Proposed Future Work

1. Negotiate on a right crawling policy between the crawler and the website
2. Consideration for Web page quality in crawling policy
3. Adaptive schemes for estimating access frequencies of web pages
4. Crawler Parallelization

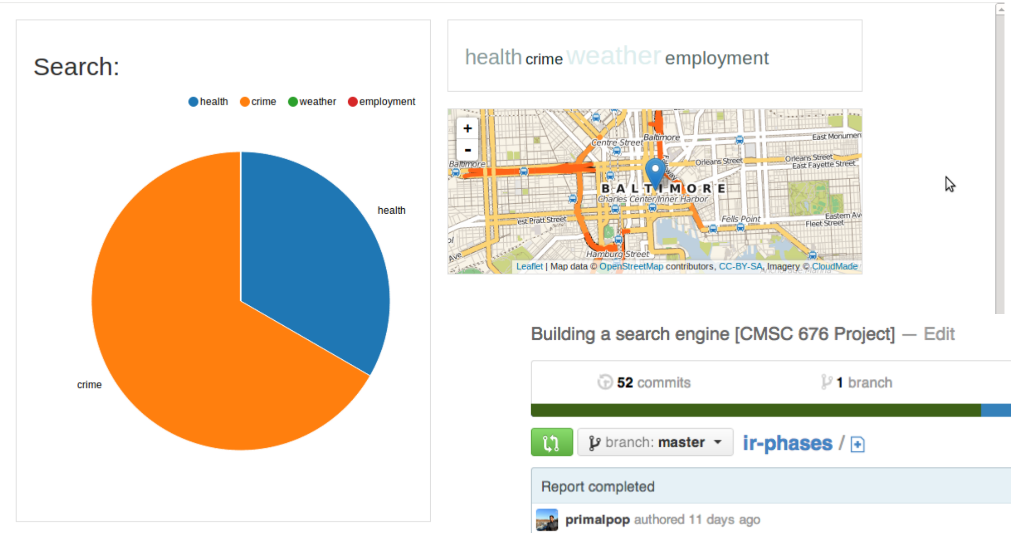
Recent Work

- “User-Centric Web Crawling” WWW 2005
- “Crawl Ordering By Search Impact” WSDM 2008
- “Recrawl Scheduling Based on Information Longevity” WWW 2008
- Google’s sitemap protocol

What still needs to be done?

1. Personal and time-sensitive data - social networks, blogs
2. Personalized content - generic user versus crawl based on different user profiles
3. Scalability and Performance issues - no general purpose solution

How I got interested

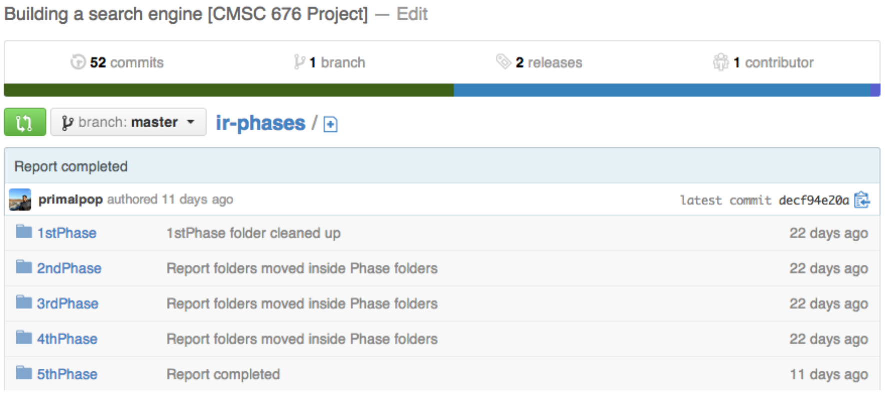


Search:

health crime weather employment

Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © CloudMade

Common Crawl



Building a search engine [CMSC 676 Project] — Edit

52 commits · 1 branch · 2 releases · 1 contributor

Report completed

primalpop authored 11 days ago · latest commit decf94e20a

1stPhase	1stPhase folder cleaned up	22 days ago
2ndPhase	Report folders moved inside Phase folders	22 days ago
3rdPhase	Report folders moved inside Phase folders	22 days ago
4thPhase	Report folders moved inside Phase folders	22 days ago
5thPhase	Report completed	11 days ago

References

1. Cho, Junghoo, and Hector Garcia-Molina. "*Dealing with Web Data: History and Look ahead.*" PVLDB 3.1 (2010): 4.
2. Olston, Christopher, and Marc Najork. "*Web crawling.*" Foundations and Trends in Information Retrieval 4.3 (2010): 175-246.
3. Cho, Junghoo, and Hector Garcia-Molina. "*Effective page refresh policies for web crawlers.*" ACM Transactions on Database Systems (TODS) 28.4 (2003): 390-426.
4. Ntoulas, Alexandros, Junghoo Cho, and Christopher Olston. "*What's new on the web?: the evolution of the web from a search engine perspective.*" Proceedings of the 13th international conference on World Wide Web. ACM, 2004.
5. Shestakov, Denis. "*Current challenges in web crawling.*" Web Engineering. Springer Berlin Heidelberg, 2013. 518-521.

References

1. *Web Crawler* (Wikipedia) http://en.wikipedia.org/wiki/Web_crawler