Week 8

Readings

"Breadth-First Search Crawling Yields High-Quality Pages" - http://portal.acm.org/citation.cfm?id=371920.371965&coll=ACM&dl=ACM&CFID=18468954&CFTOKEN=15738593

"Search Effectiveness with a Breadth-First Crawl" - http://portal.acm.org/citation.cfm?id=1390487&dl=ACM&coll=ACM&CFID=18468954&CFTOKEN=15738593

"Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering" - http://portal.acm.org/citation.cfm?id=1062768&dl=ACM&coll=ACM&CFID=18468954&CFTOKEN=15738593

"A Memory-Efficient Strategy for Exploring the Web" - http://portal.acm.org/citation.cfm?id=1248823.1249064&coll=ACM&dl=ACM&CFID=18468954&CFTOKEN=15738593

"A Web Crawler in Perl" - http://portal.acm.org/citation.cfm?id=326911.326923&coll=ACM&dl=ACM&CFID=18488853&CFTOKEN=43501364

Key ideas from the readings

- Previous research indicates that PageRank works very well in directing a crawler to download important pages early on, but breadth-first search works almost as well without the computational overhead (page 114, "Breadth-First Search Crawling Yields High-Quality Pages").
- Not only does breadth-first search find the best pages earlier, the average quality decreases over time (page 114, "Breadth-First Search Crawling Yields High-Quality Pages").
- To ease the burden placed on web servers a crawler will often wait a predefined time between requests to the same site and can even keep a queue of web sites and within each a queue of pages to visit next (page 866, "Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering"). When run in parallel as any real-world crawler would this allows the crawler to stay busy without overloading a web server.
- Another experiment shows, "Breath-first is close to the best strategies for the first 20-30% of pages, but after that it becomes less efficient." This indicates that again breadth-first search is adept at downloading the highest-quality pages early on (page 868, "Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering").
- This experiment showed that overall larger-sites-first and OPIC outperform breadth-first search and recommends larger-sites-first as the optimal strategy (page 870, "Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering").
- The queue in a breadth-first search grows very rapidly and to a much higher level than does depth-first search. However depth-first search has problems finding high-quality pages early ("A Memory-Efficient Strategy for Exploring the Web").

- Although in this experiment (due to the finite set web pages) the queue reached an upper limit before shrinking, in real world applications there would be no upper bound on the size of the queue ("A Memory-Efficient Strategy for Exploring the Web").

Technical challenges of real-world crawling

- With high probability of relative links (links referring to pages on the same host) one must be careful not to put too high of a load on the web server by requesting pages too frequently (page 115, "Breadth-First Search Crawling Yields High-Quality Pages").
- Since only a fraction of the web can feasibly be crawled, it is important to retrieve the highest-quality pages early on to ensure they are not lost later (page 755, "Search Effectiveness with a Breadth-First Crawl").
- Just crawling once is not enough; as pages are constantly changing they must be re-crawled (page 864, "Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering").
- Dynamic pages can often be separated from static html pages by filename extensions and searching for a question mark in the URL (page 867, "Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering").

Exercises

Start on final project.