Web Crawling and Web Dynamics

- Knut Magne Risvik and Rolf Michelsen, *Search engines and Web dynamics*. Computer Networks, Volume 39, Issue 3, 21 June 2002, Pages 289-302.
- Junghoo Cho and Hector Garcia-Molina, *The Evolution of the Web and Implications for an Incremental Crawler*. Proceedings of 26th International Conference on Very Large Data Bases (VLDB), 10-14 September 2000, Cairo, Egypt, Pages 200-209.

Web Dynamics

- The Web is growing at a high pace (exponential)
- Documents are updated
- Dynamics of the web is shifting
  - More dynamic and real-time information available
  - News
  - The dynamics of the web creates a set of tough challenges for all search engines
- The link structure is changing
  - Using the link structure in ranking creates a slow working positive feedback loop
- XML — understand the structure and semantics of the data is a key feature for the next generation engines
- *Goal*: The local store copy must be fresh

Local Store

- *Definition* A local store copy is a snapshot of the Web at the given crawling time for each document

Types of Crawlers

- *Periodic/batch crawlers*
  - Periodically rebuild the index from scratch
- *Incremental crawlers*
  - Updates/refreshes the local collection
  - Replaces “less-important” pages with new and “more-important” pages

Trend

- Web servers are today most commonly applications serving HTML files directly from a file system upon requests.
- More advanced publication systems tying business applications to the web servers.
- The percentage of the web that is actually indexable by search engines is decreasing (the *deep web* is growing)

Web Models

- Create a model for how documents are updated, e.g. web documents are updated as independent Poisson processes
  - Average document changed once every ten days
  - 50% of all documents are changed after 50 days
- Develop a crawling strategy that maximizes the freshness of the local store.
- Mechanisms for measuring the freshness of the local store.

Cho and Garcua-Molina

- Crawling strategies
- Freshness = probability of copy in local store is up-to-date
- Age = time since late update of real document
- Interesting measure = average freshness or age over all documents and time
- Refreshing documents using uniform update frequencies is always better than using document update frequencies that are proportional to the estimated document change frequencies
- The scheduling policy optimizing freshness penalizes documents that are changed *too often*

FAST crawler

- Incremental crawler
- Star network
- Only exchanging information about discovered hyperlinks
- *Static mapping* from hyperlink information to crawler machines
- Scales linearly with document storage capacity
- Robust with regards to failure (hyperlink information for an unavailable crawler machine is queued on the sending machine until the designated receiver again becomes available)

FAST crawler (cont.)

- Scheduling algorithm prioritizing retrieval of documents most likely to have been updated on the web
- Will only retrieve new documents when old documents are removed from the local store (e.g. document does not exist on the web anymore)
- Maximizes freshness by spending as much as possible of the crawler capacity on refreshing documents that have actually changed.
- Adaptively computes estimate of the refresh frequencies
- Decreases refresh interval if document changed; otherwise increases refresh interval
- Avoids rescheduling any document more than once for each indexing cycle