Advanced Crawling Techniques

Olli-Pekka Rinta-Koski

November 5, 2003

Selective Crawling (1)

- No crawler can harvest the entire Web
  - Time required to complete a large crawl can be significant
  - Crawler should be able to fetch the most important subset of pages in the time available

Selective Crawling (2)

- Selectivity can be added by making URL queue insertions/extractions depend on scoring function $\longrightarrow$ *best-first* search
  - Boolean: $s(u) = 1$ if document is relevant, $s(u) = 0$ if not
  - Real-valued: $0 \leq s(u) \leq 1$ (e.g. conditional probability)
  - $s(u) \doteq s_\theta^{(\xi)}(u)$ depends on URL u, relevance criterion ξ and parameters θ, not on e.g. crawler state or time
- Scoring possibilities:
  - **Depth** – penalize pages deep within site hierarchy
  - **Popularity** – threshold using indegree
  - **PageRank** – estimated using partial knowledge

Efficiency Measures for Selective Crawling

- Comparison with "ideal" crawler at time t
  - Ideal crawler fetches pages in order of relevance: $r_t = t$
  - Random crawler will only have $r_t = t/N$
  - Selective crawler should be somewhere in between
- Fixed importance target
  - E.g. fetch quickly pages with ≥ 20 backlinks:
    efficiency $= q_t(G)/t$

Focused Crawling

A *focused crawler* searches for information related to certain topics instead of using generic quality measures.

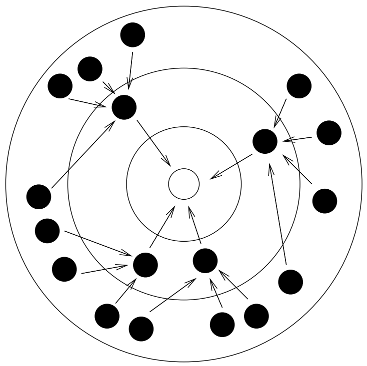
- Relevance prediction
  - Parent based

    Topic locality: outlinks are likely to lead to similar pages
    - Hard focusing: ≥ 1 link has to be relevant
    - Soft focusing: duplicate links lead to re-evaluation of relevance
  - Anchor based

    Link anchor text is often very informative

Context graphs

- Layer 0 contains node u (target document), layer i contains parents of nodes in $i-1$
- System is trained with a data set of documents with known distance to target
- After training, system predicts the layer of a new document and selects a queue for the link
- Crawler gets links from first nonempty queue $\rightarrow$ relevant information is found more rapidly



Reinforcement learning

- Unsupervised: agents learn from received rewards (or punishments)
- Agent has a state-dependent probability distribution over possible actions, as dictated by its *policy*
- An optimal policy should maximize expected rewards over time
- A crawler is rewarded for downloading relevant documents
- Standard implementations may not be able to handle the huge sets of states and actions in Web search

Reinforcement learning (2)

LASER

http://www-2.cs.cmu.edu/~jab/cv/pubs/boyan.laser.pdf

- Compute relevance scores $r_0(u)$ based on the query using TFIDF (distance between query and document)
  - Knowledge about HTML tags is used to emphasize titles, headlines etc.
- Propagate scores in Web graph to influence linking pages and iterate until convergence (≈ 5 iterations)
  - Documents contribute to the relevance of u in inverse proportion to their distance

Intelligent Web agents

- **Fish** algorithm

  Agents:

  - gain energy by collecting relevant documents
  - use energy by using network resources

  Shark algorithm: Fish + relevance score from anchor text and ancestor scores

- **CiteSeer** `http://citeseer.nj.nec.com/cs`

  - Search engine focused on computer science literature
  - Fetches PS/PDF papers from the Web and builds a citation index
  - Uses citation index for link analysis

Distributed Crawling

- A single crawler cannot fetch large amounts of data rapidly
- A scalable system can be built by distributing the crawling activity to multiple processes
- However, separate and independent crawling processes might fetch the same documents
- Performance measures:
  - **Communication overhead**
    % of bandwidth spent to coordinate activities
  - **Overlap** % of duplicate documents
  - **Coverage** % of reachable documents actually downloaded
  - **Quality** Scoring functions may depend on link structure, which may be lost in distributed crawling

Crawler interaction

- **Partitioning** = splitting URLs into subsets
- **Coordination** = how crawlers agree on the "ownership" of different page subsets
  - Overlap of independent crawlers depends only on initial seeds
  - *Static coordination:* the Web is partitioned into subgraphs before crawling starts
  - *Dynamic coordination:* crawler partitions are modified during crawling
- **Confinement** = how strictly statically coordinated crawlers stay within their partitions *(firewall, crossover, exchange)*

Web Dynamics

- Knowing the rate of change of Web subsets is crucial in allocating crawling resources
- Some definitions:
  - β**-current:** document was current before "grace period" β
  - (α, β)**-current:** documents are β-current with $P \geq \alpha$
- **Lifetime** T – time between creation and removal of page
- **Age** t – time since creation of page
- **Survivorship function** $S(t) \doteq 1 - F(t) = P(T > t)$
  $F(t) =$ cdf of lifetime
- **Change rate** $\lambda(t)$

Web Dynamics (2)

- Document age difficult to estimate by sampling *(aliasing)*
  - `Last-Modified` HTTP header may be available
- Assuming a constant change rate, it is possible to estimate lifetime from observed age
- However, the Web is growing with time

Other measures of recency

- **Freshness**
  $\phi(t) = 1$ if document is up-to-date at time t, 0 if not
  i.e. β-current for $\beta = 0$
- **Index age** is the age of the document if the local copy is outdated, 0 if it is fresh

Recency and synchronization policies

- All sites do not change at the same rate
- `.com` sites change more often than `.edu` or `.gov` sites
- Design of synchronization policy with:
  N documents of interest
  λ_i change rate, I_i refresh interval $\quad i \in [1, \ldots, N]$
  Limited bandwidth constrains refresh interval
- Goal: maximize freshness given constraints
- Choices:
  - **Constant interval** – not good
  - **Interval proportional to change rate** – even worse
  - **Solution of constrained optimization problem** – best

Homework Problems

1. Describe some advantages and disadvantages of a geographically distributed crawler.
2. Give the most relevant factors (in your opinion) when judging search engine response to a query.