Web Search Ranking

(COSC 488)

Nazli Goharian
nazli@cs.georgetown.edu

1

Evaluation of Web Search Engines: High Precision Search

- Traditional IR systems are evaluated based on precision and recall.
- Web search engines are evaluated based on top *N* documents.
  - Recall estimation is very difficult
  - Precision is of limited concern, as many users do not look beyond 1st screen.

=> *How fast and accurate the first results screen is generated?*

2

Web Page Ranking

- Evidence of quality for ranking:
  - Domain names -- *.edu,..*
  - Text content -- *term count, BM 25,…*
  - Links – *anchor text, number of in/out links, (Alg.: HITS, PageRank)*
  - Web usage log– *clicktrough data, eye tracking, geographical info (IP address, language,..), query history,..*
  - Query patterns – *certain day,time…for improving efficiency & quality*
  - Page layout – *title, font size, html tags positions on page…*
  - A problem: Web spam

3

Anchor Text

- Short, 2-3 terms, describe the linked/destination page.
- May/may not be a different point of view than the author's.
- Anchor text of links to a doc d_i included in index for d_i
- Extended anchor text (text surrounding anchor text) may also be used
- Generally weighted based on frequency (notion of *idf*)
- Spamming problem

6

Localized Search

- Using geographic information to modify the ranking of results (in addition to SC scores, link based scores,…).
- Geographic information maybe derived from:
  - Location of device sending the query
  - Context of query
    - *restaurant near Al Capone's home's town*
    - *restaurant Near White Sox stadium*
  - Geographic location in the query
    - *Chicago restaurants*
  - Geographic location in a document metadata

7

Link-based Ranking: Authorities and Hub (HITS)

- (HITS: Hyperlink-Induced Topic Search, 1999) – Kleinberg
- Links can indicate popularity
- Assigning each retrieved web page two scores: Authority and Hub scores (thus, query dependant & query independent)
  - Authority page: an authoritative source on a given topic
  - Hub page: page listing pointers to authority pages on a topic
  - Authority score: summation of scores of all the hubs pointing to that authority page
  - Hub score: summation of scores of all authority pages the hub is pointing to

8

Computing Authority and Hub Scores

- Retrieve all pages containing the query term *t*. This is called *root set*. *(~200 pgs)*
- Create a set including union of *root set* pages, pages that point to root set pages, and pages that root set pages point to. This is called *base set*.
- Using the *base set s* to compute the hub and authority scores.
- An iterative algorithm:
  - Initialize hubs and authorities with a score, ex. 1
  - Update H(p) and A(p) $H(p) = \sum_{u \in S | p \to u} A(u)$ $A(p) = \sum_{u \in S | u \to p} H(u)$

9

Link-based Ranking: Page Rank

- Mid 90's by Larry Page & Sergey Brin
- A scoring mechanism in Web search (trade marked by Google and patented by Stanford)
- Generally calculated at the time of crawling (query independent)
- Using incoming and outgoing links as an indicator of *popularity*, adjusts Web page score
- *Popular page* is defined as a page that
  - Many Web pages link to it (*inlinks*)
  - Important (popular) pages link to it

10

Page Rank

$$PageRank(A) = \frac{(1-d)}{N} + d \sum_{D_1 \ldots D_n} \frac{PageRank(D_i)}{C(D_i)}$$

- PageRank of (A) is defined based on some ratio of PageRank score of each page D_i linking into A

$C(D_i)$: number of links out from page D_i

d : damping factor (from 0-1; commonly 0.85; ~15% cases are random visits)

N: total number of pages

An Iterative Algorithm:

Initially all pages are assigned an arbitrary page rank (1/n), summing to 1

Iteratively calculate the scores until the new scores do not change significantly

To converge faster, may initialize page ranks based on number of inlinks, log info, etc.

11

Web Page Ranking

- Considering both *query dependant* and *query independent* scores (captured during indexing), a ***global score*** is generated for each page:
  - For retrieved results based on query dependant ranking (ex. BM25), rank using Page Rank

  Or,
  - Use a linear combination of various relevance evidence (textual, BM25, link,….)

SC(D, Q) = a BM25 (Q,D) + (1-a) PageRank (D)

12