My Understanding of DR. Pollett's Search Engine Code

By
Priya Gangaraju
02/10/2010

Configuration - Crawling

- Search Depth : 10
- Maximum URLs : 1000000
- Maximum links per page : 50
- Indexed file types: html, htm, jsp, cgi, shtml, php, txt

Configuration – Ranking & Searching

- Page Rank Rounds : 15
- Random surfer Alpha : .85
- Results per page : 15

Database architecture

- Consists of 5 tables
  - adjacency_matrix (ID1, ID2)
  - crawl_item (ID, URL, LINK_WORDS, PAGE_TITLE, DESCRIPTION)
  - id_item (ID, PAGE_RANK , NUM_LINKS, ROUND)
  - robot_check (URL, PATH, HAD_ROBOT_TEXT)
  - to_be_crawled (URL , DEPTH, REFERRING_ID, LINK_WORDS_REFERRER)

Crawling

1. Creates all the required tables for the database.
   - ✓ to_be_crawled, robots_check, adjacency_matrix, crawl_item.
2. Adds the seed-sites to the to_be_crawled table.
   - ✓ from seedsites.php
3. Gets a possible crawl page from to_be_crawled, check for robot.txt file, if its okay process for summary data.
   - ✓ Deletes the selected page from to_be_crawled, extracts summary data using XPath, canonicalizes urls.
4. Updates crawl_item and adjacency_matrix.

Ranking – Pagerank algorithm

1. Converts adjacency matrix into a stochastic matrix S.
2. Finds a vector v such that $S^{i}v = S^{(i+1)}v$.
3. Adjustments:
   i. Dangling nodes – nodes with no outgoing links. Assumes such nodes are connected with every other node on the web by a probability 1/total number of web-sites(S′)
   ii. Strongly connected components – Random surfer adjustment.

       $G=\alpha S' + (1-\alpha)H/\text{total number of websites}$,

       where H is the all 1 matrix.

Ranking

- Creates id_item table with page rank for each page as 1/total number of websites.
- Gets the sum of Page ranks of the dangling nodes.
- Adds (dangling nodes' page rank sum/total number of websites) to the page rank of each page.
- For random surfer correction, gets the sum of the page ranks of each page and update page rank using the formula

  $$G=\alpha S' + (1-\alpha)H/\text{total number of websites}$$

- Iterates this for 15 times.