

605.744: Information Retrieval

Problem Set (Module 9)

Sabbir Ahmed

November 7, 2022

1. (30%) Give a short definition or explanation of the following concepts:

- web spam

Answer: Content on the web that is designed to be artificially favorable in retrieval even though they may be completely irrelevant to the query.

- Broders' taxonomy

Answer: Classification of search queries by users into 3 categories: informational, navigational, and transactional.

- out-degree

Answer: In a directed graph, out-degree is the number of edges going out of a vertex.

- robots exclusion protocol

Answer: Also known as robots.txt, it's used by web pages to inform crawlers on which portions to avoid indexing.

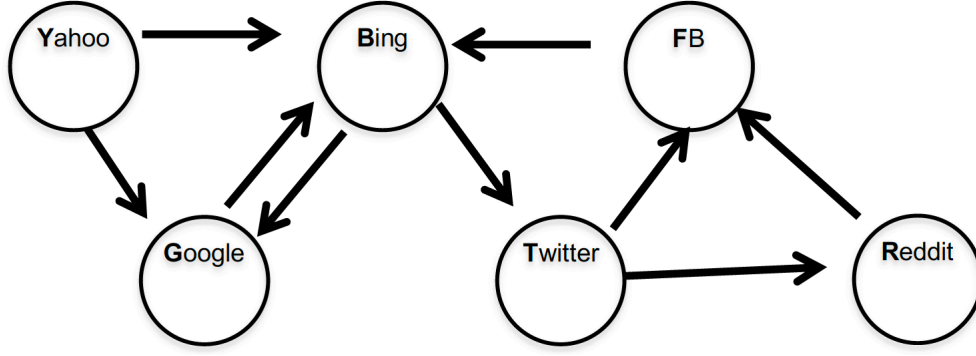
- priority queue (in the context of web crawling)

Answer: Web crawlers extract URLs from every page they scrape and store them in queues after normalizing them. The URLs can be stored in the queue depending on several factors, including how many links point to each page and their PageRank scores.

2. (20%) Describe in your own words the process described in the course text to efficiently identify near duplicate documents in a large collection.

Answer: Finding near duplicate documents is a lot more difficult than completely duplicate documents, which can be found efficiently by comparing the hashes of the entire documents. One method of identifying near-duplication is by using shingling. This process hashes n-grams of the documents into an integer between 0 and 2^{64} . These hash sets are intersected and their minimum values are compared to determine the probability of the documents being near-duplicates.

3. For this problem work with the directed web graph shown below. In the graph there are six nodes: Y, B, F, G, T, R (for the websites Yahoo, Bing, Facebook, Google, Twitter, and Reddit). Use a teleport probability of 0.20. Assume no other pages or links exist beside those shown in the figure.



- (a) (15%) Provide (i.e., write) the six recurrence equations that indicate how to iteratively calculate the PageRank score of each page at time t given scores from time $t-1$.

Answer: Using the recurrence equation:

$$PR(a) = \frac{q}{N} + (1 - q) \sum_{i=1}^n \frac{PR(p_i)}{C(p_i)} \quad (1)$$

With the values of the sites being initialized to equal probabilities:

Time	Y	B	F	G	T	R
$t = 0$	0.167	0.167	0.167	0.167	0.167	0.167

$$\begin{aligned}
 PR(Y, t_i) &= \frac{0.20}{6} + (0.80)(0) \\
 PR(B, t_i) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(Y, t_{i-1})}{C(Y)} + \frac{PR(F, t_{i-1})}{C(F)} + \frac{PR(G, t_{i-1})}{C(G)} \right) \\
 PR(F, t_i) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(R, t_{i-1})}{C(R)} + \frac{PR(T, t_{i-1})}{C(T)} \right) \\
 PR(G, t_i) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(B, t_{i-1})}{C(B)} + \frac{PR(Y, t_{i-1})}{C(Y)} \right) \\
 PR(T, t_i) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(B, t_{i-1})}{C(B)} \right) \\
 PR(R, t_i) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(T, t_{i-1})}{C(T)} \right)
 \end{aligned}$$

- (b) (25%) Using the brute-force iterative method of calculation shown in the video lecture calculate two iterations of PageRank scores for each page in the graph. Be sure to show scores at times $t=0$, $t=1$, and finally at $t=2$. Report scores using three digits of precision (e.g., 0.247, not 0.2 or 0.24696485932). Show work and do not merely provide a table of values.

Answer:

$$\begin{aligned}
 PR(Y, t_1) &= \frac{0.20}{6} + 0 \\
 &= \frac{1}{30} \\
 &= 0.033
 \end{aligned}$$

$$\begin{aligned}
PR(B, t_1) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(Y, t_0)}{C(Y)} + \frac{PR(F, t_0)}{C(F)} + \frac{PR(G, t_0)}{C(G)} \right) \\
&= \frac{1}{30} + (0.80) \frac{1}{6} \left(\frac{1}{2} + \frac{1}{1} + \frac{1}{1} \right) \\
&= \frac{1}{30} + \frac{1}{3} \\
&= 0.367 \\
PR(B, t_2) &= \frac{1}{30} + (0.80) \left(\frac{0.033}{2} + \frac{0.233}{1} + \frac{0.167}{1} \right) \\
&= 0.367
\end{aligned}$$

$$\begin{aligned}
PR(F, t_1) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(R, t_0)}{C(R)} + \frac{PR(T, t_0)}{C(T)} \right) \\
&= \frac{1}{30} + (0.80) \frac{1}{6} \left(\frac{1}{1} + \frac{1}{2} \right) \\
&= \frac{1}{30} + \frac{1}{5} \\
&= 0.233 \\
PR(F, t_2) &= \frac{1}{30} + (0.80) \left(\frac{0.100}{1} + \frac{0.100}{2} \right) \\
&= 0.153
\end{aligned}$$

$$\begin{aligned}
PR(G, t_1) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(B, t_0)}{C(B)} + \frac{PR(Y, t_0)}{C(Y)} \right) \\
&= \frac{1}{30} + (0.80) \frac{1}{6} \left(\frac{1}{2} + \frac{1}{2} \right) \\
&= 0.167 \\
PR(G, t_2) &= \frac{1}{30} + (0.80) \left(\frac{0.367}{2} + \frac{0.033}{2} \right) \\
&= 0.193
\end{aligned}$$

$$\begin{aligned}
PR(T, t_1) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(B, t_0)}{C(B)} \right) \\
&= \frac{1}{30} + (0.80) \frac{1}{6} \left(\frac{1}{2} \right) \\
&= \frac{1}{30} + \frac{1}{15} \\
&= 0.100 \\
PR(T, t_2) &= \frac{1}{30} + (0.80) \left(\frac{0.367}{2} \right) \\
&= 0.180
\end{aligned}$$

$$\begin{aligned}
PR(R, t_1) &= \frac{0.20}{6} + (0.80) \left(\frac{PR(T, t_0)}{C(T)} \right) \\
&= \frac{1}{30} + (0.80) \frac{1}{6} \left(\frac{1}{2} \right) \\
&= \frac{1}{30} + \frac{1}{15} \\
&= 0.100 \\
PR(R, t_2) &= \frac{1}{30} + (0.80) \left(\frac{0.100}{2} \right) \\
&= 0.073
\end{aligned}$$

	Y	B	F	G	T	R
t = 0	0.167	0.167	0.167	0.167	0.167	0.167
t = 1	0.033	0.367	0.233	0.167	0.100	0.100
t = 2	0.033	0.367	0.153	0.193	0.180	0.073

- (c) (5%) Which page (or pages) has/have the lowest PageRank score after two iterations?
Answer: Yahoo has the lowest PageRank score after two iterations with a value of 0.033.
- (d) (5%) Which page (or pages) has/have the highest PageRank score after two iterations?
Answer: Bing has the highest PageRank score after two iterations with a value of 0.367.