Cultural Assignment **Joseph Martinsen**
CSCE 221-506 Due Date: 01/24/2017
Dr. Teresa Leyk

Analysis of Selection of Query

Introduction

In this day and age, the world wide web has truly gone global. In 2013 it was reported that Google, Amazon, Facebook, and Microsoft had 1,200 petabytes stored online [1]. This number has only grown since then. With all this data, the designing and implementation of algorithms in order to search through all this information has proved to be the perfect solution to this issue.

Algorithm Description and Implementation

"The objective of a search engine is to investigate good superiority resultant pages" [2, pg. 1] taking into account the fact that spam filtering is very important but not the only factor. The algorithm described in the article attempts to return to the user the best search results by combining trust with authority.

The biggest wrinkle in the goal of returning the best results is correctly identifying and then filtering out spam. As Dr. Pruthi put it, the biggest challenge for a search engine is the fight with web spam [2, pg. 2]. In order to combat this, pages are assigned a Trust Rank as follows. Extremely trustworthy sites are assigned to generate seed nodes. Each of these selected sites are assigned a non-zero trust score while the rest of the site are set to a zero trust score as their initial value. Next, a biased Page Rank is second-hand to circulate preliminary trust values to linked to the outward sites [2, pg. 2]. This method results in better websites having a good trust score while spam websites will have lower trust marks. The following equation is the Trust Rank formula.

$$TR(i) = d \sum_{j:j \to i} \frac{TR(j)}{O(j)} + \begin{cases} (1-d)\dfrac{1}{|\tau|} & \text{if } i \in \tau, \\ 0 & \text{if } i \notin \tau, \end{cases} \tag{1}$$

Where TR(i) is the Trust Rank score for page i and τ is the seed set. TR(i) will be initialized as $(1-d)\dfrac{1}{|\tau|}$ if i is in the seed set and 0 if it is not.

The relative spam mass of a given page i is described in the following equation.

$$SP(i) = \frac{PR(i) - TR(i)}{PR(i)} \tag{2}$$

This equation return the division of a page i's Page Rank due to the contribution of linked spam pages. This incorporation of Trust Rank into Page Rank allows for spam pages to receive penalties while spam free web pages do not receive any rank deductions [2].

Algorithm Summary

In this article, there was not as much focus on any data structures as compared to the algorithm but it from my understanding of the algorithm, the algorithm would be deemed useless without some sort of data structure that allows for a user to easily perform queries over.

As for the algorithm, an example should be able to show how the algorithm functions. The example uses Yahoo! Research Web collection data-set which will be referred to as UK-2007 [3].This data set contains 100 million total pages, 114,529 hosts, 6,479 labeled as trust worthy sources, and 6% are spam. Queries used are highly used spammed words such as money, home-loan, and lottery.

The following are the procedures and results of the study conducted by Dr. Pruthi as I interpreted (very intriguing article but poorly used grammar).

Step 1: Identify the most famous and current 100 terms usually used as meta-keywords in website header files used by the sites marked as spam (i.e money, home-loan, and lottery)

Step 2: Add up the number of results for all the queries

Step 3: Select the 90 most popular queries from AOL Search

Step 4: Follow the procedure for different types of comparisons.

The results found that that the combination of Trust Rank and Page Rank returned an improved filtered result for the top 10 search results compared to regular searches.

References

[1] Mitchell, Gareth. *How Much Data Is on the Internet?* Science Focus. Immediate Media Co, 23 Jan. 2013. Web. 24 Jan. 2017.

[2] Pruthi, Dr. Jyoti *Selection of Query - Utilize Trust-Based Algorithms to Propagate Trust.* India: 2016 International Conference on Computing for Sustainable Global Development (INDIACom), 2016

[3] Yahoo! Research. Web Collection UK-2007. http://barcelona.research.yahoo.net/webspam/datasets/uk2007.