Quantitative Analysis

Shreysa Sharma

10/27/2017

Table 1 G1 Page Rank results

PR_Scores	PR_Links
0.047385	International_Standard_Book_Number
0.038873	Digital_object_identifier
0.030490	International_Standard_Serial_Number
0.021511	National_Diet_Library
0.017980	Bibcode
0.017648	PubMed_Identifier
0.014713	World_War_II
0.014404	Canada
0.013421	Japan
0.012906	OCLC

Table 1 contains top 10 Page rank results for BFS Search. The source ratio is 0 and the Sink ratio is 0.067

Table 2 G1 In-Links results

Num_inlinks	In_links
433	International_Standard_Book_Number
355	Tropical_cyclone
353	Digital_object_identifier
246	Bibcode
244	Wayback_Machine
222	National_Oceanic_and_Atmospheric_Administration
202	Extratropical_cyclone
199	National_Hurricane_Center
191	Pascal_(unit)
190	Storm_surge

Table 2 contains top 10 in-link results for BFS Search.

Analysis of G1 results

If we look at the results in the file G1_results.txt, the top 10 page rank results refer to International_Standard_Book_Number, Digital_object_identifier, International_Standard_Serial_Number, Canada and so on.

The International_Standard_Book_Number appears on top of the page rank and in-link lists. Most of the pages on Wikipedia have information derived from books. The books have an ISBN number that links to the ISBN page. As the number of URLS computed here are 1000, there are more references to the ISBN page. As a result the ISBN page has the most page rank score. 3 out of 10 links in page rank list and in-links are common. This suggests that page rank does take into account the number of in-links for its calculation.

Given the start page for the crawler "Tropical_Cyclone", the information being collected is most likely related to what a tropical cyclone is, how and what conditions make it it occur, what are the geographical places that it occurs most likely, which were the recent places that were hit by a tropical cyclone and so on. Also, the page rank basically suggests the denser result set where one can find more relevant information. A major part of the information on Wikipedia is derived from books and thats the reason for ISBN having the highest page rank score in the list since each book also comes with a ISBN number and each ISBN number links to the ISBN page. Therefore, after reading both the page rank list and in-link list, I feel that in this case, Page Rank is a better approach as it suggests links that can provide more information about the starting page.

Table 3 G2 Page Rank results

PR_Scores	PR_Links
0.053355	Digital_object_identifier
0.051909	International_Standard_Book_Number
0.032169	Bibcode
0.031421	PubMed_Identifier
0.022164	Canada
0.017076	United_States_dollar
0.010541	United_States
0.005918	Earth
0.005228	Puerto_Rico
0.004937	Atlantic_Ocean

Table 3 contains top 10 Page rank results for DFS search. The source ratio is 0 and the Sink ratio is 0.006.

Table 4 G2 In-Links results

Num_In_links	In_links
579	International_Standard_Book_Number
543	Digital_object_identifier
364	Bibcode
284	PubMed_Identifier
261	United_States
242	Tropical_cyclone
208	Earth
192	Extratropical_cyclone
173	NASA
165	Atlantic_Ocean

Table 4 contains top 10 in-link results for DFS Search.

Analysis of G2 results

If we look at the results in the file G2_results.txt, the top 10 page rank results and in-link results refer to Digital_object_identifier, International_Standard_Book_Number, Bibcode, PubMed_identifier, Atlantic_Ocean and so on.

The initial page rank links in the G2_results are approximately the same as BFS results. However, the links listed later, such as Puerto-Rico, Atlantic_ocean etc provide information of the order which place was the recently hit by a tropical cyclone (Puerto-Rico), or where are the ocurrances more common. In this graph

the information being captured is further away from the start crawl topic, this is expected as this is DFS search so we are giving preference to distant relationships more than closer relationships.

The results in the page rank and in-links have 4 links in common out of 10. However if one visits the earliar links in either of the lists, it is easy to infer that the page rank has higher importance as these are terms that are listed in majority pages and hence results in higher page rank or in-links list. Going by the start page the user is interested in, in this case "Tropical_Cyclone", what is a tropical cyclone, why does it occur and so on, then the page rank links can provide a volume of details about the start point as they have pages closely knit together giving out that information.

Furthermore, if the user is inclined towards gaining insight about distant relationships to the start topic, such as which places were recently hit, or what are the effects of such cyclones the page rank results again provide better valued result set when used in conjunction with the G2 graph.

Analysis of Page Rank links of G1 and G2

The top 10 page rank links in G1_results and G2_results have 5 links in common. These links refer to the ISBN, Digital_Object_Identifier and so on. As discussed in the analysis of G1 results, a lot of pages pick up information from books and therefore have ISBN numbers of these books listed. Similar is the case with the other few links that have the highest page rank values in both G1 and G2 results. Later links in G1 and G2 are although different but provide the same kind of information such as links to places like Puerto-Rico which was recently hit by a cyclone or Japan which lies in quite a cyclone prone region.

After implementation of the program and reading through the links, I have come to a conclusion that the following factors result in high page rank values:

(i) The number of incoming links to the page in question.

(ii) The quality of the links that have outgoing links to the page in question.

Page rank on G1 graph provide information closer to the start crawl doc id and G2 graph captures distant relationships, so based on the kind of information being seeked we can use the relavant graph in conjunction with the Page Rank.

Analysis of in-links list of G1 and G2

5 out of 10 links in the in-links list for G1 and G2 are same. The links refer to pages that have the highest number of in-links for the given query in BFS and DFS searches. Given the links in both the set, the G2 results being a BFS search clearly capture more relevant details more closer to the starting link. As a result we can see that more relevant links such as National_Oceanic_and_Atmospheric_Administration, National_Hurricane_Center that study the tropical cyclones closely are present in the in-links list. The G2 in-link results being generated from a DFS search go deeper into the initial links and capture their nitty gritty details such as link to NASA which in some way has research studies related to tropical cyclones. These do not directly relate to the starting point but appear in the list as DFS captures deeper information in the initial links therefore covering less on the entire surface links. Owing to above mentioned reasons, it is clear that BFS search i.e. G1 provides a better result set in terms of in-links for the starting point than DFS generated in-links i.e. G2.

Conclusion

Going by the assumption, when the user starts from the "tropical_cyclone" page, one is more likely to be interested in pages that provide more information and have links that are more relevant. Therefore for both the G1 and G2, I feel that the Page Rank list provides a better result set than the in-links.