INDIANA UNIVERSITY, SCHOOL OF INFORMATICS AND COMPUTING

Distributed Systems
Project 1

(Due Sept 11th, 2016)

Ethan Li, Rohit Patil

The description of the main steps and data flow in your program.

There are three main steps.

The first main step is to load the data from the input file. The data is stored in a HashMap with the formation of adjacency matrix. The keys in the hashmap are URLs and the values are the outgoing URLs associated with the keys.

The second step is calculation. It will calculate page rank values for M times set by users. In a single iteration, for URLs with outgoing urls, their values will be set by the values of their outgoing urls. For example, if url 1 has two outgoing urls which are url 2 and url 3, then prNew(1) = (1-d) / N + d * (pr(2)/L(2) + pr(3)/L(3) + danglingValue).
Here, prNew(i) is the new rank value of URL i, pr(i) is the previous rank value of URL i, L(i) is the out-degree of URL i, N is the number of URLs, d is damping factor, and danglingValue is the sum of the previous rank values of all dangling nodes divided by N. For those dangling nodes, their new page rank values are prNew = (1-d) / N + d * (danglingValue). After M iterations, the page rank values are recalculated and towards to convergence.

Finally, sorting the URLs by their ranks. And then output the top 10 with their rank values to a file.

In this process, the data is read from input file and being stored and calculated in memory, then written back to disk.

The output is:

```
Top 10 URLs with Highest Page Rank values.
1 : 0.050493205916884214
8 : 0.031062730432822802
19 : 0.031062730432822802
2 : 0.028395678056541068
7 : 0.028395678056541068
4 : 0.028373446705690957
11 : 0.028373446705690957
5 : 0.021334682524375016
17 : 0.01768554434656487
20 : 0.014610944238783032
```

Figure 0.1: The top 10 URLs in pagerank.input.1000.urls.19 file