Hadoop PageRank

B649 Project 2

1) The description of the main steps and data flow in our program.

Here we have to calculate the page rank of a list of source URLs. We have been provided the source URLs along with their target URLs list i.e. the Adjacency Matrix.

The project has been divided into 3 MapReduce jobs.

1. Create Graph phase :
   Input: <sourceURL, PageRank#TargetURLs>
   Output: <sourceURL, PageRank#TargetURLs>

2. PageRank phase :
   Input: <sourceURL, PageRank#TargetURLs>
   Output: <sourceURL, #targetURLs> or <targetURL, rankValuePerTargetURL>

3. CleanUp Results:
   Input: <sourceURL, sumOfPageRankValues#TargetURLs>

2) PageRank phase:

This phase has two parts:

a) PageRankMap:
   In this phase, we take in the key value pair as: <sourceURL, PageRank#TargetURLs>.
   For each key i.e. for each source node, the rank value is calculated using its probability of occurrence in the tragetURL list.
   If it is found that the node is a deadnode, then its rankvalue is calculated as (1/Total_number_of_nodes) and this rank value is scattered throughout for all the nodes. Once, this page rank of each node in the targetURL list is calculated, the node and its rank value is then passed on as the output as: <sourceURL, rankValuePerURL >.
   In the case where the is the list of TargetURL, i.e. when the node is not a dead node, the rankValue for that node is calculated using the formula (1/NoOfOutgoingNeighbors).
   The node containing the targetURL list is then output in a keyValuePair as:
   <sourceURL, #targetURLs>

b) PageRankReduce:
   In the Reduce phase, we take the input key value pair as:

<sourceURL, #targetURLs >
In the reduce phase, we later iterate over all the sourceURL values simultaneously taking the sum of all its rankValues which we received from the Mapper.
This total sum of the rank for that particular URL is then written to the console using the keyValue pair as:
<targetURL, rankValuePerTargetURL>

3) The output file (yketkar_HadoopPageRank_output.txt) which contains the first 10 urls along with their ranks.

Rank	Source URL Number	It's PageRanks
1	4	0.1206985439094576
2	34	0.10776863798237155
3	0	0.09651067430015867
4	20	0.0773080478756419
5	2	0.03690483528905177
6	146	0.03517334419309333
7	3424	0.030985948953611255
8	14	0.016459328448425656
9	16	0.011379122999302936
10	12	0.010968028739305422