Sharik Purkar

B649

Project 2

Hadoop Page Rank

Data Flow:

There are three main steps for the Hadoop Page Rank.

Here we have three MapReduce jobs to implement.

1) ***CreateGraph:***

The PageRank input Data is stored in the form of adjacency matrix as a file in the local system. We add a column to the input PageRank adjacency matrix by using MapReduce functions

Map:

Input: The input is given as *<key,value>* as (LongWritable, Text) which are the offset as key, and line in adjacency matrix as value.

Here we find out the page rank for each url. This is done by checking the number of outbounds from a particular page and then dividing the page rank by the number of outbounds.

If it is a dangling node then we calculate it by dividing the page rank by the total number of url's that are present.

Output: The output is given as (LongWritable, Text) and it represents the *<sourceUrl, Init.PageRanks#targetUrls>*.

Reducer:

Input: The input here comes from the map function of CreateGraph in the form of (LongWritable,Text) and represents the <*sourceUrl, Init.PageRanks#targetUrls>*

Output: The output is in the form of (LongWritable,Text) which are the (offset, line in file).

We then send these values to PageRank for further computation.

2) ***PageRank :***

Map:

Input: The input is given as *<key,value>* pair in the form of (offset, line in file) which comes from the result from CreateGraph.

Output: The output file is written as key value pair in LongWritable and Text format; (LongWritable, Text). The *<key,value>* pairs represent either the <*targetUrl, rankValuePerTargetUrl>* or *<sourceUrl,#targetUrls>* .

Reducer:

The output from the map is then sent to the reducer for aggregation.

Input: The input data is read as a key, value pair in the following format: (LongWritable, Text).

Output: The output from the reduce function has the key-value pair in the format (int, Text) which represent the *<url, sumofPageRankValues#targetUrls>*

3) ***CleanupResults:***

The CleanupResults is the third and final part where we use the Map and Reduce functions. Here we get the PageRank result from the PageRankReduce function and send it to a CleanupResultsMap which takes the *<key,value>* as *<offset,line in file>*. This gives us the *<url, pagerank>* which is inturn sent to the CleanupResultsReducer which gives us the final result which is the *<url, pagerank>.* This key-value pair indicates the output of our Hadoop PageRank problem which gives the respective url and pagerank related to it.

Final Output:

Output file for 5000 urls is attached as: all5000_output.txt

Output file for top 10 urls is attached in file: **pahluwal_HadoopPageRank_output.txt**

Output as given in file is:

334 9.997607666009567E-4

2234 9.993343082202914E-5

3510 9.991115726463919E-5

2036 9.969490046759656E-5

4542 9.929107305694695E-5

3674 9.915578245591296E-5

2528 9.9098574581514E-5

2058 9.89663514096048E-5

116 9.866166452421356E-4

1964 9.839417237344121E-5