Google’s Page Rank Algorithm

- Sanket Achari

Objective

- Implement PageRank algorithm to find the most important Wikipedia pages present in the 10 GB Wikipedia dataset.

- Dataset : xml file

Technologies Used

- Hadoop’s Map Reduce framework for parallel data processing
- Amazon web services such as Elastic Map Reduce(EMR), Simple Storage Service (S3)
- Programming language: Java
- IDE: IntelliJ

About Page Rank

- Definition of Page Rank:
  - PageRank is a function that assigns a real number to each page in the Web. The intent is that the higher the PageRank of a page, the more important it is. The equation is as follows:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

- Where d = 0.85, p1 ,p2 ,...,pN are the pages under consideration, M(pi) is the set of pages that link to pi , L(pj) is the number of outbound links on page pj , and N is the total number of pages.

About Page Rank

- At the initial point, each page is initialized with PageRank 1/N and the sum of the PageRank is 1.
- But the sum will gradually decrease with the iterations due to the PageRank leaking in the sink nodes.

- Sink nodes : A node/page which doesn’t have outgoing links.

Implementation

- 5 steps:

1. Extract wikipages, preprocessing : remove red links, self links, duplicate links
2. Prepare adjacency graph: Page and the number of pages linked to it
3. Count the total number of nodes N i.e. pages
4. Apply the page rank function, iterate this step 8 times
5. Thresholding & Sorting: Pages which are below threshold i.e. 5/N can be neglected. Sort the pages in decreasing order of their page ranks.

- Store the output in the text file. This output contains page and its rank.

Challenges Faced

- Didn’t know about Hadoop and AWS. Took some time to learn

- Faced difficulties during implementation of Map Reduce jobs
  - Remove duplicate links
  - Preparing adjacency graph

Improvements

- Page Rank algorithm favors old pages.
- If new page is added to data set then, since it was not referred by other pages its rank would be less even though the page contains important information.

- Need to study advanced algorithms

Miscellaneous

- Individual project
- Took 3 weeks to complete
- https://github.com/sanketachari/PageRank_Implementation.git