Computing PageRank Values of Wikipedia Articles using MapReduce

1. Overview

This programming assignment will enable you gain experience in implementing iterative MapReduce algorithms to estimate PageRank values of Wikipedia articles using Wikipedia dump data.

In this assignment, you will design and implement a system that calculates PageRank values of ***a subset of Wikipedia articles***. The PageRank algorithm is used by Google to rank web pages in their search engine query results. PageRank measures the importance of web pages. The underlying assumption is that more important web pages are likely to receive more links from other web pages. We will discuss the details of PageRank algorithm in class (see, *PageRank and MapReduce Algorithm.pdf*).

Most of the Wikipedia articles contain links to other articles or external web contents. We will consider only internal Wikipedia articles to calculate the PageRank.

2. Iterative PageRank Algorithm

There are many variations of the PageRank algorithms. You should implement the algorithm covered in class. Compute the PageRank based on the following formula:

Given page A, and pages T1 through Tn linking to A, PageRank is defined as:

$$PR(A) = (1-d)/n + d\ (PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn)) \quad (1)$$

Where, C(P) is the cardinality (out-degree) of page P and d (=0.85) is the damping ("random URL") factor.

The algorithm begins at step one with initial PR(A) = 1/N for each page A, where N is the total number of pages. The algorithm is then applied equation (1) iteratively until it arrives at a steady state; that is, until a subsequent iteration of the algorithm provides little or no further change in the distribution of PageRank. For this assignment, you are required to perform **10 iterations**.

3. Requirements

a. To simplify the problem, the input data file (`wikipedia.txt`) is a pre-processed file from a small subset of the Wikipedia dump (wikiTest.xml). Each row of the data file contains tab ("\t") delimited values, starting with the title of an article and then a list of titles of referred articles.

   Note, for the test file, wikipedia.txt, you can assume n is always 1000, which the number of lines of the input file.

b. Your final output file consists of a list of articles. Each row should contain the title of an article, its PageRank value, and a list of titles of referred articles.

   I also posted a simple test file (pagerankTest.txt) and its corresponding output (results.zip).