Project Proposal

Xi Yang, xy2390
Zefeng Liu, zl2715

Introduction

PageRank (PR) is an algorithm used by Google Search to rank web pages in their search engine results.[1] PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.[2]

PageRank algorithm can be generalized to measure the importance of any type of recursive documents. It can be viewed as a node weight metric for complex networks including social networks, transportation networks, electricity networks, species networks, etc. The computing of PageRank is, therefore, a fundamental yet nontrivial problem.

In this project, we propose to develop a parallel PageRank calculation program based on the Map/Reduce framework.

Problem Formulation

The PageRank algorithm simulates a random surfer traveling within a directed graph. Given the initial weight configuration of nodes, the algorithm outputs the probability (weight) distribution which represents the likelihood of a person randomly traveling through the edges will arrive at any particular node.

Now we formulate the Map/Reduce version of the PageRank problem.

The mapper receives the pair of node and pagerank as key, and the list of adjacent nodes as value. It maps those key-value pairs to either the pairs of node and pagerank increment or the pairs of node and list of adjacent nodes. The intermediate pairs are aggregated by key and fed to the reducers.

The reducer receives the pairs emitted by the mappers and aggregates the pagerank increments and calculates the updated pagerank value.

Objective and Deliverable

A parallelism enabled efficient PageRank calculation Haskell program with performance measurements and evaluations.

---

[1] Wikipedia contributors. "PageRank." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 15 Nov. 2019. Web. 21 Nov. 2019.
[2] "Facts about Google and Competition". Archived from the original on 4 November 2011. Retrieved 12 July 2014.