IRWS PageRank HITS

Davide Pizzolato - 881347

April 2024

1 Introduction

This document is intended to provide a description of how I coded a time and space efficient version of PageRank, HITS and InDegree in c++.

2 Compilation

You can compile the software using the Makefile, it will compile it with the maximum optimizations in C++ 20.

3 Usage

You should call the software with two parameters:

- **Graph Path:** The path of the graph file
- **K:** The top-k nodes to compute

Example: `./IR_Project ./data/web-Stanford.txt 100`

4 Code explanation

4.1 Sorter class

This class divides the input in blocks that can be kept in memory, and then perform a BSBI (it sorts the blocks and then merge the block). The class provide the data like an iterator through the method next();

4.2 writeMatrixFiles()

This method transform the input from the Sorter class in a sparse matrix representation, creating a columns file, a rows file and a dangling nodes file.

4.3 multiply()

This class multiply a sparse matrix by a dense vector.

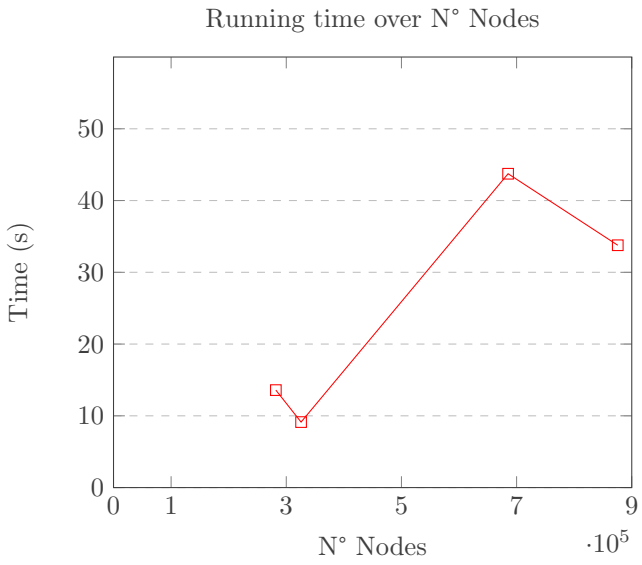
It require a start and an end indexes, implemented in order to parallelize the computation. It also accept a multiplier (that is used to normalize the vector) and a value that is added to every element of the output vector.

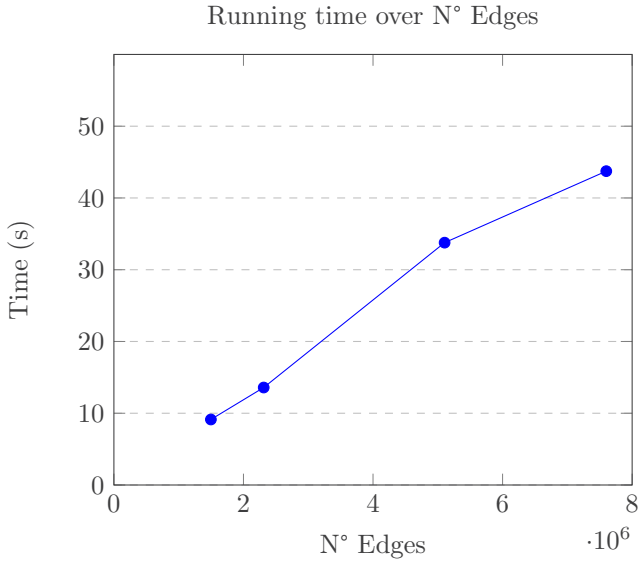
It returns the summation of the output vector.

5 Benchmark

This tests was performed on a VPS with 1 vCore, 2Gb of RAM and Ubuntu 22.04 that at the time was not doing other tasks. With just one core we cannot appreciate the improvement given by the multi-threading but it was the only environment available for the tests.

Name	N° Nodes	N° Edges	Time required (s)
web-NotreDame	325,729	1,497,134	9.129
web-Stanford	281,903	2,312,497	13.584
web-Google	875,713	5,105,039	33.771
web-BerkStan	685,230	7,600,595	43.735





6 Conclusion

From the two graphs we can see that the computational complexity is more related on the number of edges than the number of node.

This was to be expected because the two most expensive operation (sparse matrix construction and matrix-vector product) depends on the number of edges