A Comparative evaluation between sequential and parallel PageRank algorithm

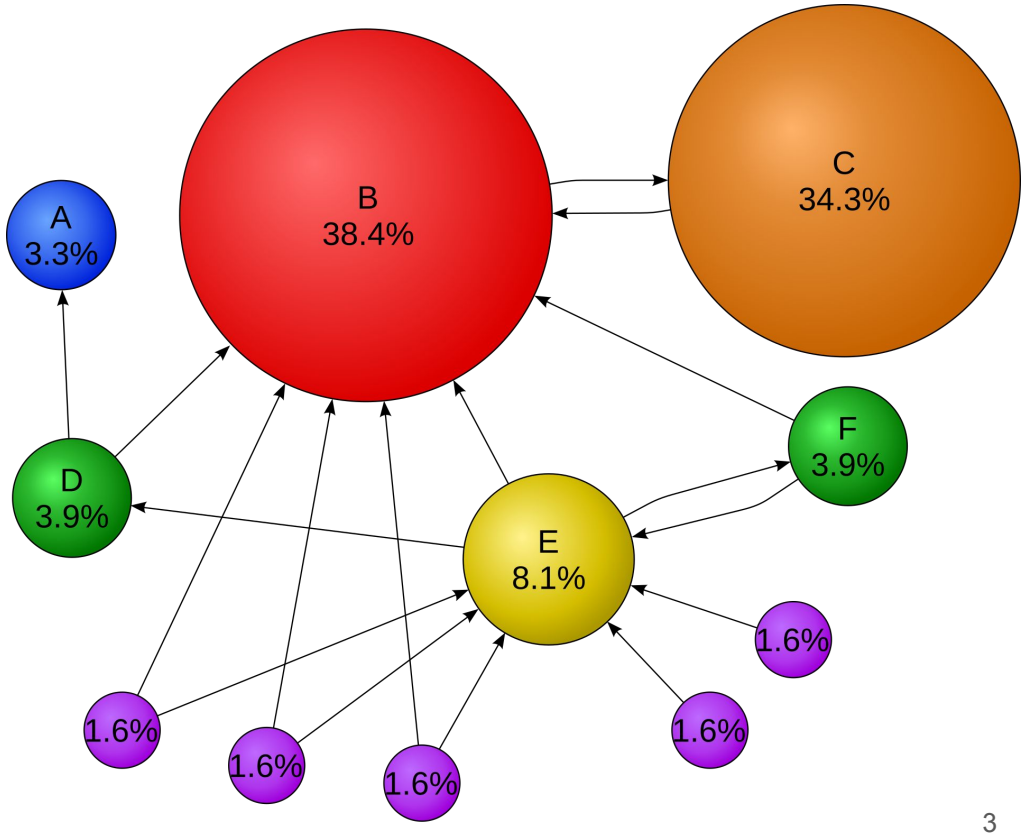
João Vitor do Amaral Spolavore, Thiago dos Santos Gonçalves

1

Description of the Computational Object

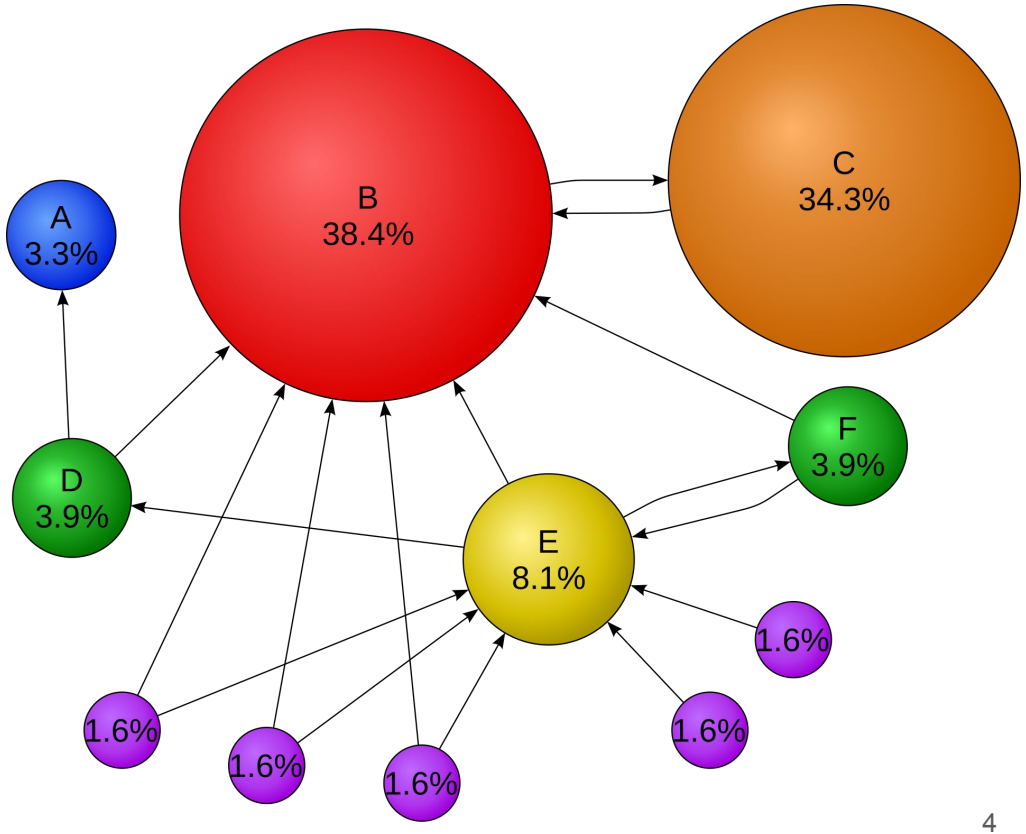
Description of the Computational Object

- The PageRank algorithm was created by Larry Page and Sergey Brin. It consists of a method of ranking web pages by assigning a numerical score based on the number and “quality” of incoming links, where links from more important pages contribute more to a page’s rank.



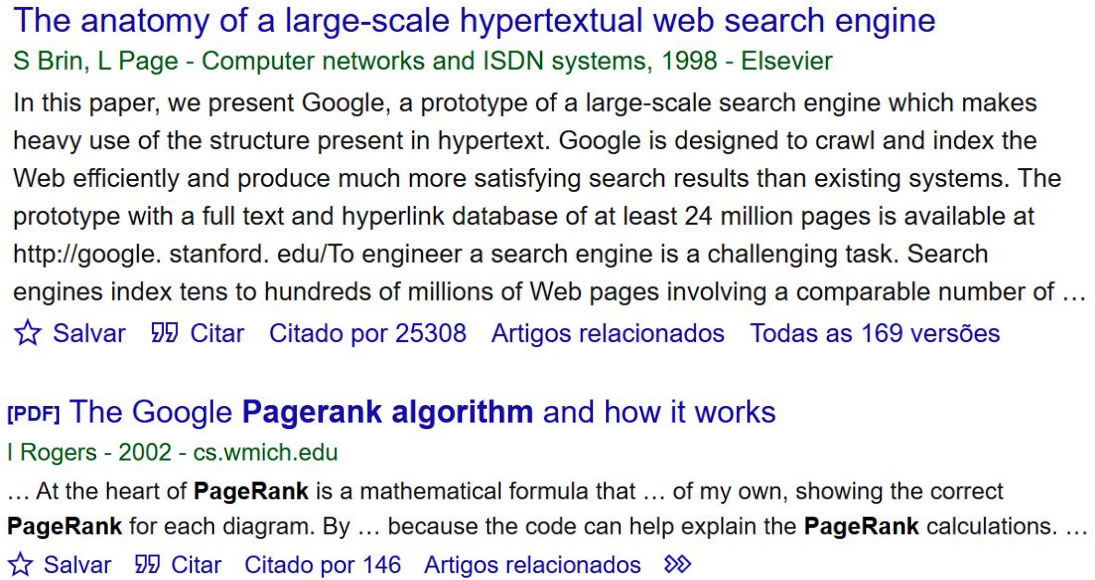
Description of the Computational Object

- The PageRank algorithm was created by Larry Page and Sergey Brin. It consists of a method of ranking web pages by assigning a numerical score based on the number and “quality” of incoming links, where links from more important pages contribute more to a page’s rank.
- The PageRank of any given page is defined recursively and depends on the number and PageRank metric of all pages that link to it.



Description of the Computational Object

- With this in mind, there have been many implementations and variations of the PageRank algorithm and scientific papers that discuss its efficiency.



The anatomy of a large-scale hypertextual web search engine

S Brin, L Page - Computer networks and ISDN systems, 1998 - Elsevier

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google. stanford. edu/To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of Web pages involving a comparable number of …

☆ Salvar 99 Citar Citado por 25308 Artigos relacionados Todas as 169 versões

[PDF] The Google **Pagerank algorithm** and how it works

I Rogers - 2002 - cs.wmich.edu

… At the heart of **PageRank** is a mathematical formula that … of my own, showing the correct **PageRank** for each diagram. By … because the code can help explain the **PageRank** calculations. …

☆ Salvar 99 Citar Citado por 146 Artigos relacionados »

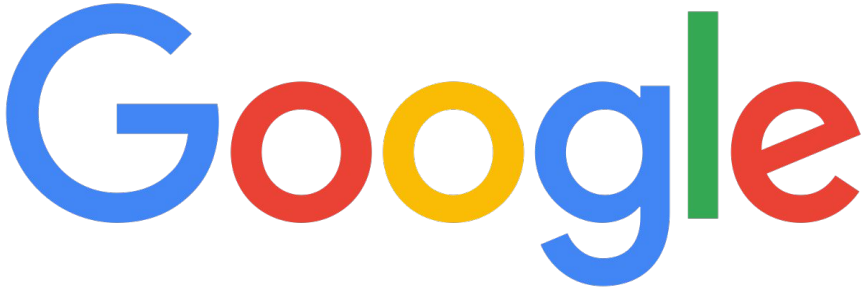
Description of the Computational Object

- With this in mind, there have been many implementations and variations of the PageRank algorithm and scientific papers that discuss its efficiency.

- Most importantly, it was the main algorithm used by Google to rank its pages to order search results until a few years ago.

The anatomy of a large-scale hypertextual web search engine
S Brin, L Page - Computer networks and ISDN systems, 1998 - Elsevier
In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google. stanford. edu/To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of Web pages involving a comparable number of …
☆ Salvar 99 Citar Citado por 25308 Artigos relacionados Todas as 169 versões

[PDF] The Google **Pagerank algorithm** and how it works
I Rogers - 2002 - cs.wmich.edu
… At the heart of **PageRank** is a mathematical formula that … of my own, showing the correct **PageRank** for each diagram. By … because the code can help explain the **PageRank** calculations. …
☆ Salvar 99 Citar Citado por 146 Artigos relacionados »



2
Choice of analysis method

Choice of analysis method

- With this in mind, we have decided to compare the PageRank algorithm’s efficiency when executed sequentially and in parallel.
- To do that, we have decided to take a **measurement** approach to compare the algorithm’s efficiency.

Choice of analysis method

- With this in mind, we have decided to compare the PageRank algorithm's efficiency when executed sequentially and in parallel.
- To do that, we have decided to take a **measurement** approach to compare the algorithm's efficiency.
- We want to visualize how the parallelization can *(if possible)* reduce the time taken by the sequential PageRank, and by how much the time can be reduced.
- And while doing that, we want to collect system metrics to figure out how the parallelization affects cache hits, misses, and CPU usage.

Choice of analysis method

- We plan on using a computer from the *Laboratório de Computação Paralela e Distribuída*'s available computers (PCAD).
- We will run the benchmark 10 times. 5 times profiling with the Intel® VTune™ Profiler, and 5 times without the Intel® VTune™ Profiler to get the time spent and energy consumption metrics.

Choice of analysis method

- We plan on using a computer from the *Laboratório de Computação Paralela e Distribuída*'s available computers (PCAD).
- We will run the benchmark 10 times. 5 times profiling with the Intel® VTune™ Profiler, and 5 times without the Intel® VTune™ Profiler to get the time spent and energy consumption metrics.
- We want to use the Intel® VTune™ Profiler to collect the following metrics:
  - Estimated Energy Consumption (Watts)
  - L1, L2, L3 Cache Hit Rate (%)
  - CPU usage (%)
- If needed, we will consider more metrics as we get further into our project.

Choice of analysis method

- The implementation of the PageRank algorithm we chose was from the GAP Benchmark Suite.

Choice of analysis method

- The implementation of the PageRank algorithm we chose was from the GAP Benchmark Suite.

- This benchmark suite was proposed by students from Berkeley University, and was widely used on many research papers.

The **GAP benchmark suite**
S Beamer, K Asanović, D Patterson - arXiv preprint arXiv:1508.03619, 2015 - arxiv.org
… We present a graph processing **benchmark suite** with the goal of helping to standardize graph processing evaluations. Fewer differences between graph processing evaluations will …
☆ Salvar Citar Citado por 654 Artigos relacionados Todas as 5 versões »

Choice of analysis method

- The implementation of the PageRank algorithm we chose was from the GAP Benchmark Suite.

- This benchmark suite was proposed by students from Berkeley University, and was widely used on many research papers.

The **GAP benchmark suite**
S Beamer, K Asanović, D Patterson - arXiv preprint arXiv:1508.03619, 2015 - arxiv.org
… We present a graph processing **benchmark suite** with the goal of helping to standardize graph processing evaluations. Fewer differences between graph processing evaluations will …
☆ Salvar Citar Citado por 654 Artigos relacionados Todas as 5 versões »

- And for graphs to use as input, we will use Stanford University's Large Network Dataset Collection.

3

Justification of computational object choice

Justification of computational object choice

- We chose this specific algorithm because we are having another class named Parallel and Distributed Programming (INF-01008), and we learn different methods to parallelize algorithms.
- Because of that, we prefer doing research on an algorithm that we feel the most comfortable with.

Justification of computational object choice

- We chose this specific algorithm because we are having another class named Parallel and Distributed Programming (INF-01008), and we learn different methods to parallelize algorithms.
- Because of that, we prefer to conduct research on an algorithm that we feel most comfortable with and that will allow us to complete the final projects of both courses at once.
- Also, Thiago used to be a research intern on LPPD, and worked directly on PCAD and CPU metrics collection.

4

Schedule for next stages.

Schedule for next stages.

2025-09-08 to **2025-09-15 (1 week):** PageRank parallelization using OpenMP.

Schedule for next stages.

2025-09-08 to **2025-09-15 (1 week):** PageRank parallelization using OpenMP.

2025-09-17 to **2025-09-29 (12 days):** Run the benchmarks on PCAD to obtain metrics for the sequential and parallelized PageRank.

Schedule for next stages.

2025-09-08 to **2025-09-15 (1 week):** PageRank parallelization using OpenMP.

2025-09-17 to **2025-09-29 (12 days):** Run the benchmarks on PCAD to obtain metrics for the sequential and parallelized PageRank.

2025-10-13 to **2025-11-13 (1 month):** Analyse the results obtained by the previous step, document our insights , create notebooks to unify our material and use scripts to create plots and graphics to help visualize the results.

Schedule for next stages.

2025-09-08 to **2025-09-15 (1 week):** PageRank parallelization using OpenMP.

2025-09-17 to **2025-09-29 (12 days):** Run the benchmarks on PCAD to obtain metrics for the sequential and parallelized PageRank.

2025-10-13 to **2025-11-13 (1 month):** Analyse the results obtained by the previous step, document our insights , create notebooks to unify our material and use scripts to create plots and graphics to help visualize the results.

2025-11-17 to **2025-11-23 (6 days):** Finalize the report, organize the relevant archives and objects of our project, send to evaluation.

References

- PageRank Algorithm Explained https://medium.com/biased-algorithms/pagerank-algorithm-explained-5f5c6a8c6696
- PageRank Wikepedia - https://en.wikipedia.org/wiki/PageRank
- The GAP Benchmark Suite - https://arxiv.org/abs/1508.03619
- The PageRank Algorithm and How it Works - https://cs.wmich.edu/gupta/teaching/cs3310/lectureNotes_cs3310/Pagerank%20Explained%20Correctly%20with%20Examples_www.cs.princeton.edu_~chazelle_courses_BIB_pagerank.pdf
- The Anatomy of a Large Scale Hypertextual Web Search Engine - https://www.sciencedirect.com/science/article/pii/S016975529800110X