Literature survey for Parallelization of Pagerank Algorithm

Group-8

Nishant Kumar Dhvanil Parikh Suyash Ghuge Shreyas Shankar

The performance of PageRank is bounded by the external memory (i.e., DRAM) accesses. Thus, many prior works focus on optimizing the memory performance. In, a graph reordering approach is proposed to improve locality and reduce cache misses. The reordering approach identifies the optimal permutation among all the vertices in a given graph by keeping the vertices that will be frequently accessed together. However, the pre-processing overhead of is non-trivial. In, an FPGA design to accelerate the PageRank algorithm is developed. Similar as our partitioning approach, the design in also partitions all the vertices into vertex sets such that the data of each vertex set can fit in the on-chip BRAMs of FPGA. However, the FPGA accelerator can only process one partition at a time, while our design can process distinct partitions in parallel. In, Beamer et. al. propose the propagation blocking approach, which first stores propagations (i.e., messages) in cached bins and accumulates them before writing into DRAM. This approach reduces the number of memory accesses but results in additional memory requirement. Moreover, the design in does not optimize the data layout; thus, random memory accesses still occur when writing messages into the memory.

Comparative Study of Page Rank and Weighted Page Rank Algorithm
(http://www.rroij.com/open-access/comparative-study-of-page-rank-and-weightedpage-rank-algorithm.php?aid=45124)

PAGERANK
1) *Advantages*:

- Less Time consuming:- As pageRank is a query independent algorithm i.e. it precomputes the rank score so it takes very less time
- Feasibility:-This algorithm is more feasible as it computes rank score at indexing time not at query time.
- Importance: - It returns important pages as Rank is calculated on the basis of the popularity of a page
- Less susceptibility to localized links: - For calculating rank value of a page, it consider the entire web graph, rather than a small subset, it is less susceptible to localized link spam.

2) *Disadvantages*:

- The main disadvantage is that it favours older pages, because a new page, even a very good one, will not have many links unless it is part of an existing web site.

- Relevancy of the resultant pages to the user query is very less as it does not consider the content of web page.
- Dangling link: This occurs when a page contains a link such that the hypertext points to a page with no outgoing links. Such a link is known as Dangling Link
- Rank Sinks: The Rank sinks problem occurs when in a network pages get in infinite link cycles.

WEIGHTED PAGERANK

1) *Advantages*:

- Quality: Quality of the pages returned by this algorithm is high as compared to pageRank algorithm
- Efficiency: It is more efficient than pageRank because rank value of a page is divided among it's outlink pages according to importance of that page.

2) *Disadvantages*:

- Less Relevant: As this algorithm considers only link structure not the content of the page, it returns less relevant pages to the user query.

A Parallel Data Mining Algorithm for PageRank Computation
(https://www.researchgate.net/publication/310801114_A_Parallel_Data_Mining_Algorithm_for_PageRank_Computation)

Advantages:

- CUDA language gives more access to thread parallelism.
- This implementation uses a high number of threads in core.
- Intercom **Partition-Based Parallel PageRank Algorithm**
- Munication of cores via shared memory.
- The results obtained show that the proposed model outperforms the difficulties in any scalable web graph and gives the best results compared to the CPU version.

Disadvantages:

- Dead Ends: Dead Ends are simply pages with no outgoing links.
- Spider Traps: Another problem in PageRank is Spider Traps. A group of pages is a spider trap if there are no links from within the group to outside the group.
- Circular References: If you have circle references in your website, then it will reduce your front page's PageRank.

Partition-Based Parallel PageRank Algorithm
(https://ieeexplore.ieee.org/document/1488928)

Advantages:

- For an EDSCC (Equally dense and strongly connected cluster), the total packet size will be equal to the size of the destination rank vector .
- For an ideal case of an input web graph that can be partitioned into several EDSCCs, the I/O cost is constant while the synchronization cost is zero. Therefore, the running time of both Partition-based and Split-Accumulate algorithms will linearly decrease when the number of processors increases.

Disadvantages:

- The worst case occurs when all nodes in the input web graph are fully connected. Each processor has to transmit every new computed score to other processors for synchronization.
- For the worst case, the and synchronization cost will be varied in accordant with the number of processors. If we continually increase the number of processors, the running time will also continually increase, or the overall speedup performance will be rapidly drop

An Improved Page Rank Algorithm based on Optimized Normalization Technique
(https://ijcsit.com/docs/Volume%202/vol2issue5/ijcsit2011020570.pdf)

Advantages:

- It is the query independent algorithm that assigns a value to every document independent of query.
- It is Content independent Algorithm.
- It concerns with static quality of a web page.
- Page Rank value can be computed offline using only web graph.
- Page Rank is based upon the linking structure of the whole web.Page
- Rank does not rank website as a whole but it is determined for each page individually.
- More the outbound links on a page T, less will page. A benefit from a link to it.
- Page Rank is a model of user's behavior

Disadvantages:

- The researchers can plan to explore more on the page rank algorithm based on damping factor to enhance the performance of the proposed scheme.