Advances and Applications in Mathematical Sciences
Volume 22, Issue 3, January 2023, Pages 799-807

GRAPH PARTITIONING USING HETEROGENEOUS DEPENDENT SELF FAST ADAPTIVE HEURISTICS FOR LARGE SCALE DATA CENTRE

R. MYNA and D. GUNASEKARAN

[1]Research Scholar
[2]Associate Professor
Department of Mathematics
PSG College of Arts and Science
E-mail: mynagopal@gmail.com
gunapsgcas@gmail.com

Abstract

Employing graph partitioning optimizes the graph performance of the data handling in large scale geologically distributed data centres. In addition, it reduces communication cost of computation model utilized for job processing. Existing graph partitioning techniques face several challenges on handling the multilevel network heterogeneities in the geo distributed data centers. In order to handle those complications, a novel Heterogeneous Dependent Fast Self Adaptive Heuristics (HDFSA) has been proposed. It is efficient in minimizing the cost of inter data center on data transfer of graph processing jobs. Further it provides better usage of the network heterogeneities. Self-adaptive Heuristics applied to dynamic graph to efficiently manage the jobs during resource failures. Heuristics generalized to compute partitions as phases for graph partitioning of large jobs with arbitrary number of constraints and strategies. Performance Evaluation of proposed self-adaptive heuristics depicts that it is capable of reducing the inter-Data Centre communication time upto 84% and it is capable of reducing the WAN usage on sparse graph by up to 85% which is compared to be fast than traditional graph partitioning methods with less runtime data centre overhead.

1. Introduction

Graph partitioning is exploiting parallelism in efficiently managing the large graphs in wide range of large scale applications, for example social

2020 Mathematics Subject Classification: 05C70.
Keywords: Graph Partitioning, Self-Fast Adaptive Heuristics, Job Processing WAN Network, Resource utilization, Large Scale Geodistributed datacenters.
Received December 30, 2021; Accepted March 10, 2022

network analysis like Face book [1], Twitter [2] and meetup [3]. Graph Partitioning is primary and vital task to maintain effective load balancing. Its determination model is capable in reducing the inter node communication. Existing model in Graph partitioning provide an important aspects in decreasing the data transmission cost and it ensure effective load balancing on processing of graph based jobs. Especially graph based application like event based online social networks involves large sets of events and its relevant data has been scattered in numerous geographically distributed heterogeneous data centers (HDCs). Meetup web application collects terabytes of user information containing events oriented data in form of image data format and video data format from various members in all part of the world [9].

Heuristics based graph partitioning strategies provides high scalability and low latency to the user based services in the social networks management. Especially Meet up has structured as heterogeneous geo-distributed Data Centres to manage event oriented data on employing graph partitioning methods. In addition, it is leads to undesirable results on distributing the data among the data centre incorporating heuristics solution in terms of privacy and cyber-security regulation aspects in graph based distributed data processing. It is unavoidable to process the job distributions in Geodistributed servers through graph based methods. Finally various heuristics based graph partitioning solutions has been identified with many technical complications towards partitioning and processing graph containing data across Geodistributed Data Centers. Initially, the data transmission between graph partitions on various servers in the Geodistributed Data Centres propagates through the entire network which is leads to more communication time than normal intra Data Centre based data communication.

Furthermore many traditional solutions employed by cloud provider for graph partitioning projects higher prices to the data user. It becomes more expensive on inter network communication traffic than on intra network communication traffic [6]. Heuristic based methods on graph partitioning balances the workload among various data partitions among data centre to reduce the vertex replication rate of the application [7] and result with high inter data centre communication size and high computation cost. To address

the above issues, heterogeneous dependent self fast adaptive heuristics has been proposed for load balancing in the geo distributed data centres.

The distributed method for graph partitioning named heterogeneous self adaptive heuristics based method on graph partitioning is employed to minimize the data centre cost for data communication between the heterogeneous resources. Further it process on reduction of time for data communication on jobs of graph processing in Geodistributed data centre while meeting the deadline constraints of Wide Area Network Usage in terms of memory and bandwidth in the geo-distributed environment. Finally budget constraints of heterogeneities of the resource with large sizes consumers more graph traffic and network bandwidth, it is effectively managed using optimization of the heuristics separately on each resources separately.

Initially, a self adaptive streaming heuristics has been proposed to minimize the inter data centre data communication size and utilize the single flow stream partitioning method to assign edges to different data centres. Next, two adaptive heuristics for partition refinement is employed to identify the network performance and increase the effectiveness of the data partitioning in the distributed data centres.

The rest of this article is sectionized into several parts is as follows. Section II provides brief introduction of traditional graph partitioning methods. Next, we define and formulate a new distributed heterogeneous graph partitioning solutions for large scale Geodistributed data centre in Section III and the proposed self adaptive graph partitioning techniques has been detailed with strategies and constraints in Section IV. Proposed model has been evaluated on basis of heuristics to heterogeneous self adaptive data in Section V and finally article is concluded in Section VI.

2. Related Work

In this part, detailed information on graph partitioning model for execution of large scale architecture has been analysed on various aspects. Typically processing of the Graphs is highly complex on single network connected resources. Distributed analytical frameworks for graph processing are generated to execute graph processing in simultaneously on large no of nodes towards inter data communications in data centres. In prior actual

graph analytics, the input graph has to be partitioned into multiple segments and it has to be executed using heuristics conditions. Figure 1 provides the structure of the heuristics model for graph partitioning.

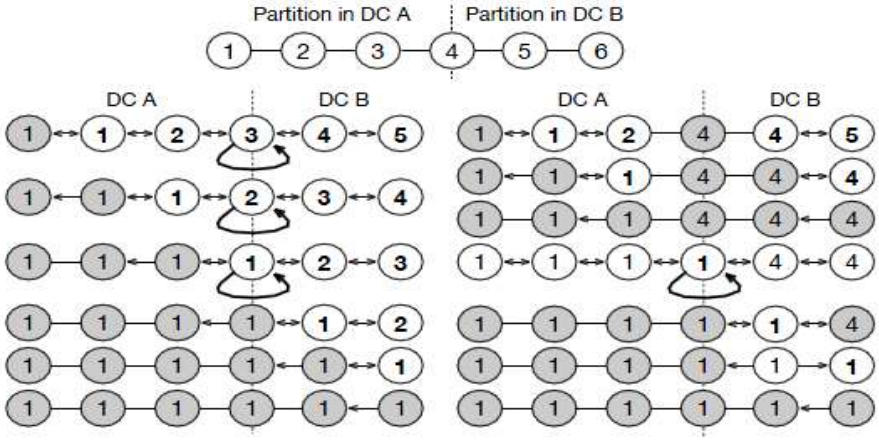

Figure 1. Graph Connected Component for Distributed Graph Partitioning.

The graph partitioning architecture will manage the data aggregation and load sharing among various data centres efficiently. Traditional state of art solutions [6] provides a data abstraction for graph generation and it aggregates the graph on basis of data aggregation heuristics [7]. The integration of data abstraction and data aggregation model allows heterogeneous resources uses the data vertex and data edges as effective data management analytics applications and it is capable of handling complex structures.

3. Graph Partitioning – Problem Definition

In this part, graph partitioning problem has been defined and its solution to manage the large scale data centre is carried out on employing Heterogeneous Self adaptive graph partitioning method. The proposed model has been highlighted with complex design towards efficiently handling large graph representations through adaptive heuristics. The design solution is employed across multiple Geodistributed datacenters paradigm. In this design, fundamental challenge arises in processing of graph with raw input from globally operated datacenters connected using wide area network [10]. The input graph store data and computes on traditional graph analytics

frameworks with various heuristics faces numerous challenges which is as follows

- Graph Bisection occurs due to data heterogeneity and resource heterogeneity in geodistributed data centre.
- Partitioning of dynamic size graph will leads to high data
- It leads to NP hard problem on selecting optimal partition

4. Heterogeneous Self Fast Adaptive Heuristics Proposed Technique

In this part, a heterogeneous self fast adaptive heuristics model generates partitioned graph for NP hard data distributions. First, task or load to be processed is converted into graph containing vertices and edges. The graph is partitioned towards execution on basis of edge partitioning and vertex partitions on geographically distributed data centres has been represented in terms of graph is as represented as

Graph $G = (V, E)$

V = Vertex nodes

E = Node Edges

Undirected Graph Weight $G_w = (NW, EW)$

NW = Vertex weights

EW = edge weights

Where N represents data, NW represents data communication cost

Edge $EW(i, j)$ means data di from Node I is send with edge weight EW to Node j.

Cross edge partitions of initial vertices and edges of undirected graph is partitioned into sub-graphs S denote

$G_S = (VS, ES)$

VS – Vertices Set of sub-graph

ES – Edge set of Sub graph

Each Partition of graph is represented as Sub graph or partition P

$$P_k = (p_{k1}, p_{k2} \cdots p_{kn})$$

Set Condition as partitions with zero duplicate edges as

$$E(pi) n E(pj) = \Phi$$

To avoid the Replication

$$E(p_i) < C$$

Where C is the partition with maximum edges

Criterion for Selecting the Edge or Vertex

$V = \arg\max \mu(v_i)$ where $V_i \text{€} N(p_k)$.

The graph edge partitioning hindrance requires a balanced p-edge graph partitioning determination to minimize the Radio Frequency of the data center. Further graph policies uses multiple heuristics is to manage with the received graph data after replication partition. Graph partitioning algorithms with heuristic solution produces the high partitioning accuracy. The following algorithm represents the graph handling procedure using heuristics on the heterogeneous distributed data centre.

Algorithm 1. Heterogeneous Self Fast Adaptive Heuristics

Set Graph execution mode to Modularity M,

Modularity updates of graph is $M \leftarrow 0$,

Current Graph Processing error $\delta \leftarrow \infty$,

While $M\delta > \infty$ do

If (Optimal Heuristics Solution $M_p V$) then

Edge update of Graph $E(p_i) = |V|^* d$

$$M_p \delta_k + 1 \leftarrow 2|E|,$$

$$M_p \delta_k + 1 \leftarrow 2p^* |E(p_i)|$$

The edges between Normal vertex V and the vertices V in partition p_k are allocated to global error free partition

Else

The vertex with the highest value of $\mu(v_i)$ in neighbour vertex set $N(p_k)$ is detremined as the optimal vertex v

There are more edges between normal edge e and the vertices in P

The local partition will occurs complicated with the updation through addition of the optimal vertex V from $N(p_k)$. The determining criterion $\mu(v_i)$ is based on the modularity changes.

5. Simulation Results

In this part, the heuristics based graph partitioning model employed will partition the data in the distributed data centre into unstructured parturitions and it traverse with graph based workloads on basis of adaptive heuristics. The simulated environment is self configured with dynamic high frequency specification as graph based data partition model. The spanned vertices of the dynamic graph are overestimated and it is processed using self adaptive constraints to avoid replications. Modularity is computed on dynamic graph to reduce the latency of the computation in each time window.

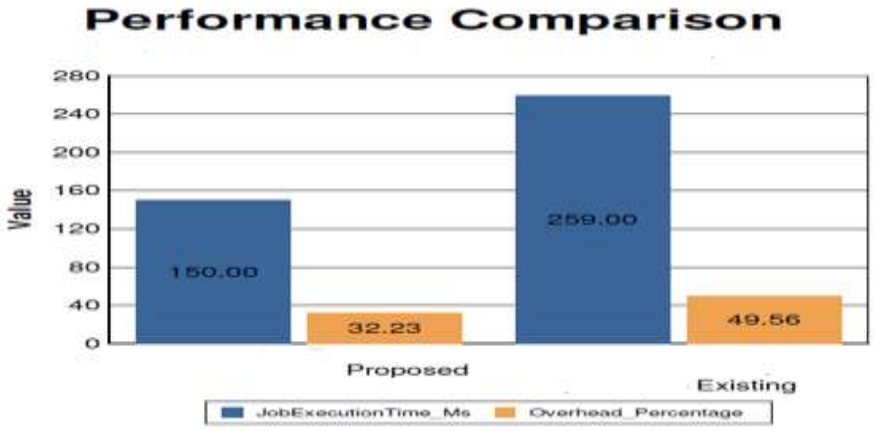

Figure 2. Performance evaluation of the Self Adaptive Heuristic model

against traditional model.

The data centre consist of dynamic job is processed in terms of quality of service and Communication cost. Further model is has been processed using heuristics to generate the optimal partitions to process. In these sub graph reconfigurations (stability) of the partition is carried out without adjusting parameters on resource scheduling strategy of dynamic heterogeneous environments. Probability of the vertex can be core vertices to oscillate over large degree in the graph with high data dynamics. The performance evaluation of the self adaptive heuristics model is established by computing the performance of the graph partitioning mechanism to generate the optimal partitions towards resource adjustment against various tasks. The figure 2 explains the performance results for different partitioning heuristics and those represented in the table 1

Table 1. Performance evaluation of the Dynamic model for graph partitioning using adaptive heuristics.

Technique	Time for Job execution in ms	Overhead (%)
Adaptive Heuristics based graph partitioning	165ms	33.63
Multistage Edge partitioning	269ms	45.76

From the above performance results, difference between the partitioning methods on different degree of the data communication on dynamic workload has been clearly presented. Partition on data distributions partitions the graph data simultaneously, which will be adaptable for large-scale graph partitioning.

Conclusion

In this paper, a heterogeneous self fast adaptive heuristics has been proposed with detailed designed and simulated to minimize impacts of overhead and latency on processing large scale graph jobs in Data Centres. Proposed model includes two different optimization phases. In First phase

local partitions utilizes the multiple partitions of the graph to reduce communication and second phase, uses global partitions using modularity constraints in optimal partition prediction heuristics to refines graph partitioning on communication cost.

References

[1] A. Ching, S. Edunov, M. Kabiljo, D. Logothetis and S. Muthukrishnan, One trillion edges: Graph processing at facebook-scale, VLDB. 8(12) (2015), 1804-1815.

[2] N. Elyasi, C. Choi and A. Sivasubramaniam, Large-scale graph processing on emerging storage devices, in FAST'19, (2019), 309-316.

[3] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson and C. Guestrin, Powergraph: Distributed graph-parallel computation on natural graphs, in OSDI'12, 17-30.

[4] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola and J. M. Hellerstein, Distributed graph lab: A framework for machine learning and data mining in the cloud, VLDB, 5(8) (2012), 716-727.

[5] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser and G. Czajkowski, Pregel: A system for large-scale graph processing, in SIGMOD '10, pp. 135-146.

[6] C. Mayer, M. A. Tariq, C. Li and K. Rothermel, Graph: Heterogeneity aware graph computation with adaptive partitioning, in Proc. of IEEE ICDCS, 2016.

[7] E. Minkov, W. W. Cohen and A. Y. Ng, Contextual search and name disambiguation in email using graphs, in SIGIR '06, (2006), 27-34.

[8] Q. Pu, G. Ananthanarayanan, P. Bodik, S. Kandula, A. Akella, P. Bahl and I. Stoica, Low latency geo-distributed data analytics, in SIGCOMM '15, (2015), 421-434.

[9] J. Ugander and L. Backstrom, Balanced label propagation for partitioning massive graphs, in WSDM '13, 507-516.

[10] L. Zhu, A. Galstyan, J. Cheng and K. Lerman, Tripartite graph clustering for dynamic sentiment analysis on social media, in SIGMOD '14, 1531-1542.