IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 12 (December. 2018), ||V (II) || PP 26-32

A Novel machine learning Approach to Memory Optimization in Distributed Graph Processing Using CBIIM Model

Dr.R Priya[1], Sona Mary Louis[2]
Associate Professor and Head, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[1]
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[2]
Corresponding Author: Dr.R Priya

Abstract: Graph Processing System (GPS) is one of the open-source systems. It gives the advantages like scalability, fault-tolerance and easy program execution. The GPS is similar to the Google's proprietary systems such as pregel and apache giraph.Therefore, memory usage patterns emerge as a primaryconcern in distributed graph processing.Obtainingeffective memory optimization in vertex-centric distributed environment is critical and most of the existing research failed to obtain it. In existing system, the field of data mining has involved in those domains to optimize thememory in Distributed Graph Processing.Those techniques cannot achieve high optimization and this technique cannot apply for Real-World Graphsanalytics. The study proposes a Sequence Single-source Shortest Paths along with Enhanced RedBlackTreeEdgesdata mining framework andCBIIMframework to improve the Memory Optimization in Distributed Graph Processing.

Keywords: Distributed Graph Processing, Graph Compression, Pregel, Apache Giraph, Memory Optimization, GraphX.

Date of Submission: 29-11-2018 Date of acceptance: 13-12-2018

I. INTRODUCTION

In few years, the web applications will get incremented every day, especially in the field of social network. This network is expanding the World Wide Web (WWW) space and stores the volume of data. The social networks are such as Facebook, twitter, etc. holds above 1 billion active users, so these users every day sharing and storing a huge amount of data. It includes all activity in FB (Facebook) like, profile information storing, sharing some information from others, save some general sharing information, and create own page so increase the URLs etc. So all these data are saved in a memory area. The *google* is one of the largest search engines, as it stores huge amount of data so easily and retrieve the required information easily. It give some reports based on URL increment. The graphs vertices are to provide the increment level and realize the number of DGPS approaches, and it gives the parallel execution of algorithm. This technique divides the graphs into a number of partitions and these are assigned to the vertices. This assigning is applied on the machines, and this term is known as, "think like a vertex". This term is introduced by Pregel system. The existing systems are performed correctly but it stores the unwanted data, so is considered as the main issue of these systes.The space-efficient graph representation in vertex-centric distributed environment requires the memory optimization techniques and it deals with web-scale graphs, and in this concept also there remain the GPS issue. Each and every vertex is a single node and it maintains the list of out-edges. The partition of this out-edges to hard the compression task as vertices, so finds the neighbors, that is physical nodes. But the single-machine settings are to exploit the vertices, so it is feasible one. Then, this system based graph partition the vertex and it is independently processed to other vertices. This system achieves the memory optimization by using some representations. It needs and allows the algorithm for mining of graph elements without decompression; but this decompression includes excessive memory, so it produce the final result in the unencoded representation. The Facebook uses the memory optimization technique "Apache Giraph". This technique is used for graph search service, so automatically improves the performance and scalability. The best growth of Facebook is reached, which means it uses the memory optimization technique and gets careful and technical based improvement. Then, mostly avoids the redundancy information and ineffective information. This work describes the basic memory optimization approaches and realize the representations of weighted or out-edges in the graph. This representation is available in web-scale that is Pregel paradigm. The pregel system is to store the out-edges of each vertex independently. This policy is simply known as, "locality of reference".

II. PROBLEM DEFINITION

In the paper [1] Gonzalez J E presented concept , "PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs". The PowerGraph is a large-scale graph-structured computation model. It solves the problems are (i) highly skewed power-law degree distributions, limiting performance and scalability. This concept is to express the internal structure of the graph programs and finds out the problems. The PowerGraph abstraction is introduced a new method to distributed graph placement and structure of power-law graphs. It was used the large graph but it needs the time.

In the paper [2] Han M, KhuzaimaDaudjee proposes the concept is based on Giraph Unchained. In the year the authors have presents the "Giraph unchained: Barrierless Asynchronous Parallel Execution in Pregel-like Graph Processing Systems." In this system is used the Bulk Synchronous Parallel (BSP) model and it after implemented. This concept is to solve the BSPs problems are, (i) poor performance, (ii) globalize the synchronization. It mitigates both side message staleness and globalized the synchronization. It proposes the graphic, which implements the BAP model. It uses the Giraph model for evaluation. the graphic is to provide high performance, so improves the proposed system.

In the paper [3] Low Y presents the concept namely, "Distributed GraphLab: A framework for Machine Learning and Data mining in the cloud". In this system introduced naturally or efficiently supports the data mining and machine learning algorithms. Its main term is a GraphLab abstraction. It includes the Asynchronous, dynamic and Graph-parallel computation. So it provides the data consistency and high degree based parallel performance, it selects the shared-memory area settings. This system is to develop the graph based extensions. It includes the (i) pipelined locking, (ii) data versioning, these reduce the network congestion or traffic and it too reduces the network latency effects. It introduces the handy-Lamport snapshot algorithm. But it only supports for the Named Entity Recognition (NER) task and then, it providing the abstraction.

In the paper [4]Malewicz G proposed the concept namely, "Pregel: A System for Large-Scale Graph processing". The large graphs include web graph and various social networks. These networks are easily scaled and it includes the billions of vertices and trillions of edges and it processed efficiently. The large graph includes a sequence of iterations. First, a vertex has received the messages and sent into other vertices through outgoing edges, and it is known as 'mutate graph topology'. This system is known as "vertex-centric approach" and it includes and elaborates the set of algorithms. In this system is designed efficiently and it includes the scalability and fault-tolerant oriented implementations on the clusters. The cluster forms commodity of computers and programmed easily. The distributed system oriented details are hidden by using the abstract API (Application Programming Interface).

In the paper [5] Slihoglu, Jennifer Widom proposed the GPS concept and it which includes the graph partitioning techniques. This system basically includes the pregel system features are: (i) it uses the extended API and to create the global computations so it is efficient, (ii) it uses the dynamic repartitioning scheme so every time easily reassigns the vertices to different but this reassigns based on the messaging patterns, (iii) the term optimization provides the adjacency lists. It holds the high-degree vertices, so automatically improves the performance. In the GPS implementation uses the static and dynamic graph partitioning schemes. It describes the compilation details of high-level domain-specific programming languages and uses easy expressions.

In the paper [6] Yan D, James Cheng, Yi Lu, Wilfred Ng has proposed "effective techniques for message reduction and load balancing in distributed graph computation". The graph includes, simply networks but the massive graphs include online social networks and communication networks. So it analyzes the large graphs and many DGPS systems. The pregel is a simple basic message passing mechanism. The send and receive messages are significantly exchanged between one vertex to another vertex. This concept proposes the two message reduction techniques are,

(i) Vertex mirroring with message combining,
(ii) An additional request-respond API.

These techniques reduce the exchanged number of messages via the network. The send and receive messages only by using the single vertex. It establishes the pregel system into pregel++. Finally, the large graph uses the message reduction methods and improves the DGPS performance.

III. PROPOSED SYSTEM

The chapter discusses about the proposed methodology and the steps involved in the proposed system. The system effectively proposes a new hybrid approach model to achieve memory optimization and efficient execution in vertex-centric distributed environment, which concentrates on effective model, which is called as CBIIM (Combine BVEdges, IntervalResidualEdges and IndexedBitArrayEdges Model).

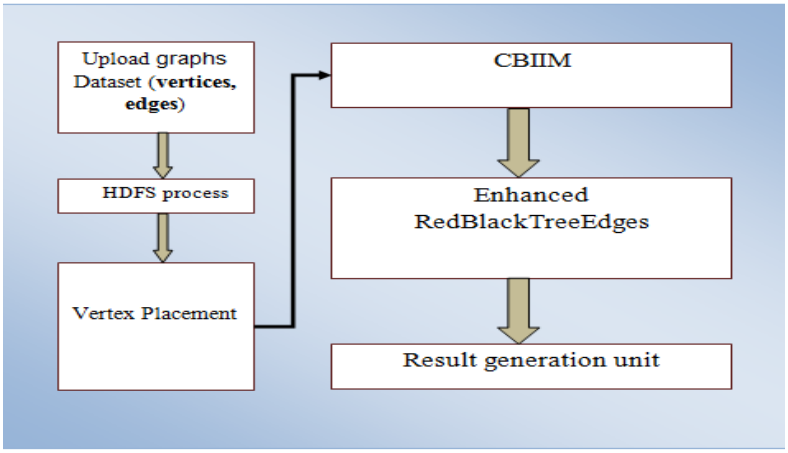

Fig 3.1 Proposed system architecture

Contribution of the Proposed Work

The followings are the contributions of the proposed system.

- To obtain effective memory optimization in vertex-centric distributed environment is critical and most of the existing research failed in it. So proposed light weight CBIIM along with RedBlackTreeEdges framework achieve memory optimization and efficient execution in vertex-centric environment.
- A HDFS process has been applied for load data set values in to Hadoop distributed file system because it's holds very large amount of data and provides easier access and parallel processing.
- Sequence Single-source Shortest Paths algorithm has been used for to find the effective shortest path from a single source vertex to all other vertices in the given graph.
- CBIIM unit has been applied to process allows for efficient mutations on the graph elements and it also support for weighted graphs.
- Proposed EnhancedRedBlackTreeEdges significantly improves memory overhead and memory bottleneck

IV. METHODOLOGIES

This chapter brings the overall algorithm and technique involved in the proposed system. This gives the brief idea about the steps and process involved in the proposal. The proposed system involved with 4 major is HDFS, Sequence Single-source Shortest Paths,CBIIM and RedBlackTreeEdges. The first process is the data collection and uploads in to distributed file system, where this is collected from various graph web sources which contains vertices and edges. Each step is discussed in the following chapters.

HDFS process

Now the most important process in which it initially has to load data from local file system to Hadoop. After that loading process will completed HDFS can process that data.

Steps 1: create the directory in HDFS
Steps 2: load the data in Hadoop system
Fs –copy from local source-path to destination-path .
In this source-path will be the local files and destination-path will be the Hadoop file system path.
Step 4: View the file in HDFS
Step 5: Start all command
Step 6: Initiate node manager
Step 7: Initiate resource manager

Sequence Single-source Shortest Paths

This algorithm focuses on finding the shortest path (between single source vertexes) by using the shortest path algorithm and every other vertex in the graph. These algorithm steps are listed out one by one.

Algorithm steps

Step 1: Start the process

Step 2: if supersteps == 0 then //assign the values
Step 3:vertex.setValue (1);
Step 4:minDist←----- is Source (vertex)? 0: 1;
Step 5: for message in messages do
MinDist min (minDist, message);
Step 6: if minDist<vertex.getValue () then
Step 7:vertex.setValue (minDist);
Step 8: for edge in vertex.getEdges () do
Send Message (edge, minDist + edge.getValue ());
Step 9:voteToHalt ();

Pseudocode

```
MinGraphDistance_calculate (Intdist [array], Boolean sptSet [])
   {
      // Initialize minimum, value
Int min = Integer.MAX_VALUE, min_index=-1; //assigning
        For (int v = 0; v < V; v++)
          If (sptSet[v] == false &&dist[v] <= min)
          {
             Min = dist[v];
min_index = v;
          }
           Return min_index;
   }
```

Enhanced RedBlackTreeEdges

This algorithm proposed to implement minimize or eliminate space overhead in a vertex-centric distributed environment. This algorithm every path from the root to a leaf has the same length so it makes a Perfect balance. Here working procedure list out by here step by step details.

Pseudocode

Step 1: Recursively traverse left sub tree.
Step 2: Visit root node.
Step 3: Initiate search process
Struct node *current = root; //assign root node as structure node
Step 4: process with while loop till loop end
While (current->data! = data)
If (current! = NULL)
//go to left tree
If (current->data > data)
Current = current->leftChild;
//else goes to right tree
Else
Current = current->right Child
Return current;

Step 5: End of the process.

V. RESULT AND ANALYSIS

Data Sets

In the section data set is generally describes it is a collection of data. It mostly corresponds to the contents of a single database table or a single statistical data matrix. The table is a collection of rows and columns. So every column of the table represents particular variable. Each row corresponds to a given member of the data set. A term Hadoop is a big collection of datasets. Then the proposed system is takes the dataset from ten different social media network based applications. Because the social media data is a very fast and every day

increasing data it includes the huge data. So, for this experiment is to select this dataset by using the effective Hadoop framework.

graph	vertices	edges	type
uk-2007-05@100000	100,000	3,050,615	web
uk-2007-05@1000000	1,000,000	41,247,159	web
ljournal-2008	5,363,260	79,023,142	social
indochina-2004	7,414,866	194,109,311	web
hollywood-2011	2,180,759	228,985,632	social
uk-2002	18,520,486	298,113,762	web
arabic-2005	22,744,080	639,999,458	web
uk-2005	39,459,925	936,364,282	web
twitter-2010	41,652,230	1,468,365,182	social
sk-2005	50,636,154	1,949,412,601	web

Table 5.1 Dataset of our experimental setting with a total of ten publicly available web and social network graphs.

The dataset indicates the significant improvement on the space-efficiency for the important proposed techniques. In the concept only collects the memory related contents or datasets, because the proposed work is related with the memory optimization. So the dataset is fully adjacency lists of the graphs. It is connected with lot of neighbors with consecutive ids.

Experimental Results

This section describes the implementation process. Implementation is the realization of an application, or execution of plan, idea, model, design of a research. This section explains the software, datasets and modules which are used to develop the research. Then the experiment is performed on an AMD Sempron (tm) 140 Processor 2.70 GHz with RAM capacity 3 GB. The algorithms are implemented in Java and are run under Hadoop platform.

In the evaluation by measuring the time needed for the proposed algorithm based execution which operates on the weighted graphs. Then using the out-edge representation its execution by using the search shortest paths algorithm via Function.

Results and Analysis

The experiments are basically designed so that the different parts of the work could be evaluated easily and effectively. To this aim, first the memory Optimized important terms are discussed. Second, all the dataset is processed and apply the different graphical approaches are theoretically analyzed over the social media based application dataset completely. Finally, algorithms are implemented this proposed work was implemented using the java. The performance of this proposed work scheme was compared with the existing algorithms based on the following parameters.

Memory Performance Evaluation

This experiment has been done through the Dataset. The dataset is processed by our effective proposed framework. The proposed system Memory-Optimized improvement in terms can be analyzed with different set of data's and results are shown in Figures

Metrics	Dataset (No of vertex)	Existing	Proposed Framework
Memory-Optimized (%)	**DS1(100)**	95	88
	DS2(150)	93	80
	DS3(200)	93	78
	DS4(250)	90	75

Table 5.2: Memory-Optimized Comparison

Performance comparison of proposed CBIIM with existing approaches based On Detection Memory-Optimized Result.

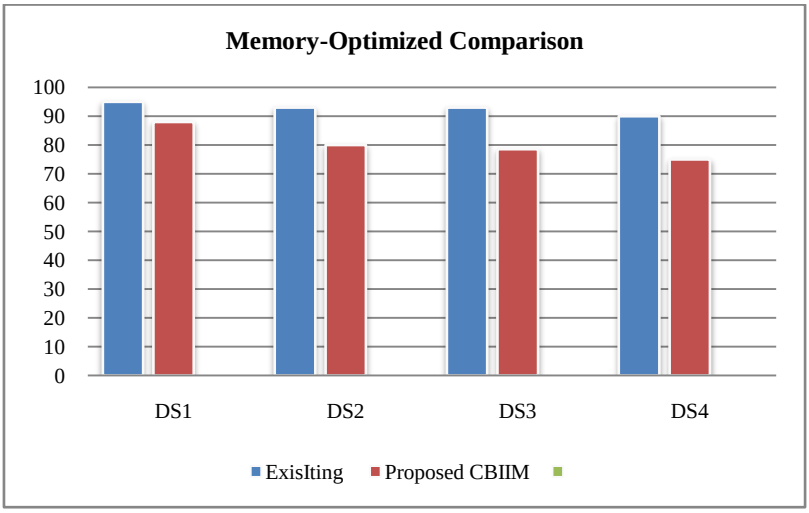

Table 5.3: Memory-Optimized Comparison

From the results shown in the graphs, it can be observed that the proposed CBIIM based approaches provides better Memory-Optimized result. The system finally performs the analysis to show the Memory-Optimized of the proposed system.

Time comparison between existing and proposed

It is observed that the proposed system consumes only less time than the existing. Performanceis very promising compared to the existing methods that have been explored in the previous chapters. The figure below provides the time comparison chart.

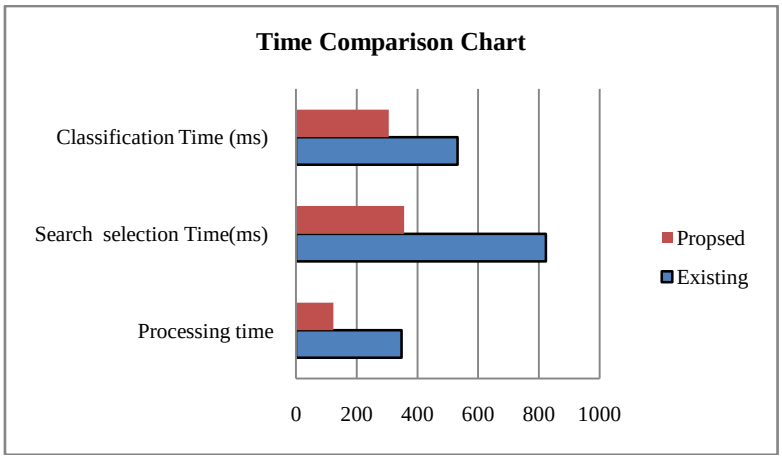

Figure 5.4 Time comparison between existing and proposed

.

The next chapter deals with the presentation of the conclusion and enhancements

VI. CONCLUSION

The study proposed a new memory optimization scheme for distributed graph processing. The system solves the main problem in the literature, which is memory requirement. The study overcomes the above problem by applying the realized memory optimization algorithm. The proposed system handles the large category dataset more rapidly, accurately and effectively. It keeps the good scalability at the same time. This system effectively labels the out-edge representations. The proposed system includes the three techniques and are implemented, these are known as compressed techniques, so compressed the out-edge representations for distributed graph processing. Finally, proposed enhanced RedBlackTreeEdges, it shows the significant improvement in the performances. The experimental results are evaluated by using the Java. The experimental result shows that integrated extended proposed algorithm shows better quality assessment compared to traditional optimization techniques. From the experimental results, the execution time is calculated it is almost reduced than the existing system.

The framework of the proposed model that can be used to analyze the existing work, identify gaps and provide scope for further works. The researchers may use the model to identify the existing area of research in the field of data mining in other dataset and use of other research algorithms. As further work, improvements can easily be done since the coding is mainly structured or modular in nature. In the system can changing the existing modules or adding new modules can append improvements. Further enhancements can be made to the application by expanding the existing modules future research system can apply memory optimization handles for distributed graph processing. This future research can apply high dimensional data set along with any new and effective memory optimization tool, so easily optimizes any types of memory for their distributed graph processing.

REFERENCES

[1]. Gonzalez J E, Y. Low, Gu H, Bickson D, and Guestrin C, "PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs," in Proc. of the 10th USENIX Symposium on Operating Systems Design and Implementation, Hollywood, CA, USA, October 8-10, pp. 17–30.(2012)

[2]. Han M and Daudjee K, "Giraph Unchained: Barrierless Asynchronous Parallel Execution in Pregel-like Graph Processing Systems," Proc. VLDB Endow., vol. 8, no. 9, pp. 950–961, (May 2015).

[3]. Low Y, Gonzalez J, Kyrola A, Bickson D, Guestrin C, and Hellerstein J M, "Distributed GraphLab: A Framework for Machine Learning in the Cloud," Proc. of the VLDB Endowment, vol. 5, no. 8, pp. 716–727. (2012)

[4]. Malewicz G, Austern M H,Bik A J C, Dehnert J C, Horn I, Leiser N, and Czajkowski G, "Pregel: A System for Large-Scale Graph Processing," in Proc. of the ACM SIGMOD Int. Conf. on Management of Data, Indianapolis, Indiana, USA, June 6-10, pp. 135–146.(2010).

[5]. Salihoglu S and Widom J, "GPS: a graph processing system," in Proc. of the 25th Int. Conf. on Scientific and Statistical Database Management, Baltimore, MD, USA, July 29 – 31, pp. 22:1– 22:12. (2013).

[6]. Yan D, Cheng J, Lu Y, and Ng W, "Effective Techniques for Message Reduction and Load Balancing in Distributed Graph Computation," in Proc. of the 24th Int. Conf. on World Wide Web, Florence, Italy, May 18-22, pp. 1307–1317.(2015)

Dr.R Priya. " A Novel machine learning Approach to Memory Optimization in Distributed Graph Processing Using CBIIM Model."IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 12, 2018, pp. 26-32.