Parallel Computing for Analysis of Large Scale Social Network

Badrun Nahar Khan
American International University-Bangladesh
Dhaka, Bangladesh

Noor Mohammad Zahid
American International University-Bangladesh
Dhaka, Bangladesh

ABSTRACT
The importance of social network analysis is realized as an inevitable tool in forthcoming years. This is due to the unprecedented growth of social-related data, boosted by the proliferation of social media websites and the embedded heterogeneity and complexity. The data generated from social network 10-15% data among them are structured and 85-90% data are unstructured. The unstructured data are useless. We will need additional technique to process those unstructured data. This paper focuses on parallel computational techniques for social network analysis. In particular, a brief discussion of some existing parallel algorithms is carried out and a new parallel computational technique is proposed to achieve parallelism.

General Terms
Parallel algorithm

Keywords
Parallelism, vertices, edges, graph, architecture and so on.

1. INTRODUCTION
In parallel computing, different algorithms are used across multiple processors. Usually this means distributed computing, where a computer splits a computation into chunks and tells other computers what chunks to evaluate. So a machine learning application would be fitting the parameters of a model using a computer cluster, instead of using just one computer's processor. By the way, parallel computing is tremendously overused in machine learning.

With the rapid growth of social science people are using social network more and more. In Facebook there are 1.59 billion accounts (approximately $1/5^{th}$ of worlds total population), 30 million FB users updating their status at least once each day, 10+ million videos uploaded every month, 1+ billion content pieces shared every week and more than 1 billion photos uploaded every month. So a huge number of data is created and we need to analyze those data [1].

During the last few years, different works have been done to achieve parallelism in social network analysis but there are some issues that need to be addressed. These are:

i. It is necessary to have a specific network infrastructure or any data transmission media to transfer data from user to server.
ii. When a social media user likes or follows any specific page they didn't mention about how any recommender system will work.
iii. Need to follow a specific parallel machine learning technique to achieve parallelism.
iv. Need to have proper idea how to overcome any barrier while achieving parallelism.

To overcome these shortcomings, we have conducted a study. Main aim of our study is to achieve parallelism to analyze large scale data extracted from social networks.
memory space, there is no need to partition the graph, and we can avoid the overhead of message passing. However, attaining good performance is still a challenge, as a large class of graph algorithms are combinatorial in nature, and involve a significant number of non-contiguous, concurrent accesses to global data structures with low degrees of locality.

2. RELATED WORK:
Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time [2]. In big data like social network graph, graph partitioning is a very helpful method. Graph partitioning in parallel data processing and specially partitioning of unstructured social network graph is very valuable. It is mainly used to partition underlying graph model of computation and communication. Graph partition problem can be represented as $G = (V, E, \omega)$ where V → Vertices, E → Edges and ω → weight of a graph [3]. Here node or vertices represented as an actor and connection between vertices represented as the relationship between different actors.

In balanced graph partitioning problem, we need to partition a graph in equal size while maintaining the minimum number of edges between vertices. Suppose, if we want to parallelize a system, we need to assign data or process evenly between different applications and also we need to maintaining minimum communication between them. For a (k, v) balanced partitioning graph problem, we need to divide k (v represents data or tasks and k represents processes in parallel machine on which processes needs to be executed) pieces of size at most v(n/k) [4].

3. METHODOLOGY:
In this study we proposed a new computing architecture for managing those huge number of big data in social network. Different steps of our methodology are as follows.
A software platform can be built for storing and processing big dataset like Facebook data on distributed hardware. There are 4 parts in that architecture.

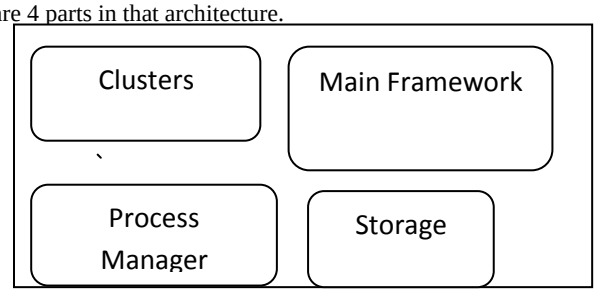

Figure1: Proposed Architecture

Different steps of our proposed architecture are:

Main Framework: This platform is a software framework for handling large distributed processing data on different computer or host machines. This programming model is created for processing data which requires "Data Parallelism", the ability to compute multiple independent operations in any order. This framework performs in parallel if there is any system failure, monitoring and management of the other parts of the architecture.

Process Manager: In this partition, we perform different processing and save these processed data in the storage. Job scheduling task is also done here. Operating system for this framework will be there. It works as a system for data processing. Redundant data is reduced in this layer by running algorithm in different data files and by taking the reduced data from that files are later saved for further storage. All those work are done here.

Storage: The storage infrastructure is specially build to store, manage and retrieve huge amount of data. This storage infrastructure is specially build for large data set so that data can be easily accessed, retrieved from that huge data by applying efficient algorithm to access data as fast as we can.

Cluster: Clusters are the various number of host machines which are connected through a dedicated network to work as a single system.

When an operation is performed it runs on cluster. From cluster it goes to the process manager for processing of the data, here the data required to complete is retrieved from the memory storage and then the output shows through the cluster. Here the framework is regulated and solve any issue in this operation if any error occurs during this operation.

This architecture can handle big scale of data.

4. EXPERIMENTAL SETUP:

Below is a list of minimum hardware configuration for our proposed architecture:

i. RAM: at least 8GB

ii. Disk Space: 50GB

iii. CPU: quad-/hex-/octa-core CPU running at least 2.5GHz

This configuration is enough for learning purpose of our proposed system.

5. FEATURES OF OUR SYSTEM:

i. **Distributed Processing:** Data will be stored in storage in distributed manner.

ii. **Availability:** In this system data are available and we can access data easily after if there is any hardware crash, we can access data using different cluster machine form the storage.

iii. **Scalable:** Our system is highly scalable because we can add new cluster machine in our system whenever we need to add any.

iv. **Cost-effective:** Its cost effective because of the commodity of hardware. We will not need any specialized machine for this system. Thus it saves money.

v. **User Friendly:** Users like Facebook users will not deal with this system framework as it will run in the background.

6. CONCLUSION:

Social computing is a cross-disciplinary research and application field. The data generated from social network usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable time. To facilitate the development of social software, the main issue is the representation of social information and social knowledge. We need special architectures to handle such a huge data generated from social network. In this paper we have tried to solve that issue. In our future work we will try to implement our proposed architecture in real life to test the efficiency of our system.

7. ACKNOWLEDGMENTS

We would like to show our great honor to Dr. Saddam Hossain Mukta who has introduced with this topic and helped us to understand the things we could not understand.

8. REFERENCES

[1] Chaffey, Dr. Dave (2018, March 28). How digitally mature is your business ? World Economic Forum

[2] Snijders, C., Martzat, U., & Reips, U.-D. "Big Data": Big Gaps of Knowledge in the Field of Internet Science, International Journal of Internet Science, 2012

[3] Sanders, Peter and Schulz, Christian, "Think Locally, Act Globally: Highly Balanced Graph Partitioning", International Symposium on Experimental Algorithms, 2013

[4] Andreev, Konstantin and Racke, Harald, "Balanced Graph Partitioning"

[5] George Cybenko, Parallel Computing for Machine Learning in Social Network Analysis