Exercise set 1

Hou, Jue

014695647

* 1. Wikipedia describes big data as a term that refers to the data is so large and complex that conventional data process application can hardly deal with them. And Gartner describes big data as 3Vs, which means volume, velocity and variety [1]. Volume means the quantity of big data is of a great size. Velocity means that big data is available all the time and keeps updating all the time. Variety means that big data not only contains text data, but also images, audios and data in many other formats.

“Big” can obviously refer to the quantity of data. I think the main reason for the emergence of big data is that with the technology developed, people just simply have more tools or sensors to observe things. Keeping tracking everything will of course generate data of big size. “Big” can also refer to the big potential of data. The insights and patterns that lie under data are of great value for the world.

* + 1. Tour Recommendation: Help people to plan a better tour route.
    2. User Behavior Analysis: Analyze
    3. Text Categorization: Classify a paragraph or a book to

1. K-Means algorithm is one of algorithms I learned from Introduction to Machine Learning. It is also one unsupervised machine learning algorithm and is used for clustering. It will get input K from user to decide the number of clusters. Starting with initial mean vectors, which can be randomly selected, the algorithm will assign data vectors to the mean vector with smallest Euclidean distance to form clusters. After clusters are formed, new mean vectors of each cluster will be generated. Again, the algorithm will assign data to new closest mean vector. Loop until the mean vectors stabilized. Now, all the data have been exactly assigned to K clusters.

I used to implement it to cluster images of hand-writing numerical figures. If I have to implement this algorithm in the distributed environment, one problem could be how to extract valid mean vector from data. Compared to storing all the data in a single node, algorithm will have to traversal data across all of the nodes, which potentially require for a higher reliability for data storage and transmission. Also, if the amount of data is enormous, algorithm will spend a lot of time for clustering. One possible solution is to pre-cluster with some reasonably selected sample data.

* 1. The problem for enormous amount of data could be that people have to spend a long time to analysis these data. Besides that, some intermediate result will also cost a huge amount of resource because of the size of raw data. One possible solution is to optimize the algorithm for analysis so that the size of intermediate result can be minimized. Or transform raw original data to some kind of hyper parameters whose meaning are equivalent to the original data.
  2. It is quite obviously that a higher computation power and reliability is needed for such rapid transmission rate. First of all, the system has to have the ability to keep up with the transmission rate and be able to handle with it as well. Once it happens to have a break down or a failure, people will inevitably suffer huge loss of information. One possible solution is to have a distributed system with hierarchy structure. A master assigns and managers all the work, while slaves do the computation and return results. MapReduce model of Google [2] is a good example for such system.
  3. RDD is a memory abstraction for distributed cluster computation.
  4. Advantage of RDD is mainly due to its limitation on “creating” RDD blocks. This limitation benefits lineage mechanism as fault-tolerance strategy, backup task implementation, and the management for memory and task scheduling.
  5. DAG is a directed graph with finite number of verticals and edges but no directed circle.
  6. Checkpointing is a mechanism that system will take a snapshot for certain data at regular intervals in case of node failure.
  7. Fault-tolerance means that distributed system should have some mechanism to detect and handle all kinds of failure which will eventually lead to failure and should be able to recover from failure in a short time with acceptable loss.
  8. IndexedRDD is a key-value store based on RDD to improve the function of the programming model for searching and manipulating.

**References**

1. Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011. Retrieved 13 July 2011.
2. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.