International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 07 | July 2018 www.irjet.net p-ISSN: 2395-0072

MRI- A Map Reduce Interpretation Framework for Uncertainty Reduction using SOM in Big Data

S.M.Lakshmanan., M.B.A.,(M.Phil)[1], P.Karthikeyan.,B.Sc(CS).,M.C.A.,M.Phil., (Ph.D)[2],

[1]Research Scholar, Department of Computer Science, Prist Deemed to be University.
[2]Research Advisor and Assistant Professor, Department of Computer Science, Prist Deemed to be University, Thanjavur.

Abstract - *Big data include capturing data, data storage, dataanalysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. Big data can be described by the following characteristics: Volume, Variety, Velocity, Variability and Veracity. Applications are in Government, International development, Manufacturing, Healthcare, Education, Media, Information Technology . In this paper, we propose a new frame work to support uncertainty in the analytics process through a self-organizing map algorithm running in Map Reduce framework for parallel computations on massive amounts of data. This framework uses an interactive data mining module, uncertainty modeling and knowledge representation that supports insertion of the user's experience and knowledge for uncertainty reduction in the big data.*

Key Words: **Uncertainty, preprocessing, SOM, Map Reduce, Big data**

1. INTRODUCTION

The massive amounts of data that collect over time which difficult to analyze using common database management tools. Big data includes activity logs (machine generated data) which consist of unstructured format capture from web. The storage industry is continuously challenged as Big data increases exponentially where security is one of the challenging and harmful concern. To handle Big data technology takes cardinal part in analysis.

1.1 Uncertainty

Uncertainty is widely spread in real-world data. A data can be considered un- certain, vague or imprecise where some things are not either entirely true nor entirely false. To model uncertainty, numerous techniques have been proposed, including probabilistic measures, Bayesian networks, belief functions, interval sets and fuzzy sets theory. There has been a lot of research in the application of fuzzy sets theory to model uncertainty. The Fuzzy set (FS) theory is a more flexible approach than classical set theory where objects belong to sets (clusters) with certain degree of membership. Using fuzzy sets theory as a mean to measure and quantify uncertainty.

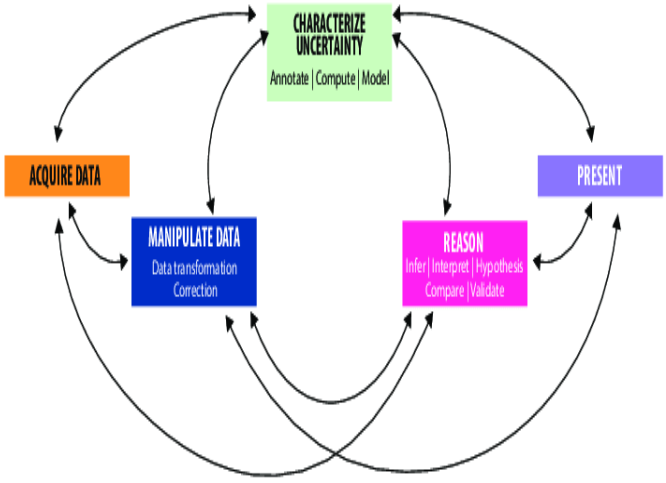

Fig. 1. Uncertainty modeling

1.2 Big Data

As everyday data are being collected from applications, networks, social media and other sources Big Data is emerging. Studies have shown that by 2020 the world will have increased 50 times the amount of data it had in 2011, which was currently 1.8 zettabytes or 1.8 trillion gigabytes of data . The basic reason for the sharp increase in data being stored over the years simply comes down to cost of storage. The IT industry has made the cost of storage so cheap that applications are capable of saving data at exponential rates. This brings the challenge of having existing network infrastructure learn how to manage and process this big data so that it can be utilized into useful information. Many big data applications work in real-time. Hence, these applications need to create, store and process large amount of information which produces a great deal of volume and demand on the network. When looking at data from a networking perspective, many different areas are needed to be explored These include network topology optimization, parallel structures and big data processing algorithms, data retrieval, security, and privacy issues. The topic of big data is still a new exciting area of research among the IT community and will be requiring much attention for the years to come. A typical organization has a limited network infrastructure and resources capable of handling these volumes of traffic flows which cause regular

services (e.g., Email, Web browsing, video streaming) to become strained. This can reduce network performance affecting bandwidth and exposing hardware limitations of devices such as firewall processing being overwhelmed. Providing security and privacy has also become a major concern in Big Data as many critical and real-time applications are developed based on Big Data paradigm.

2. MRI- MAP REDUCE INTERPRETATION FRAMEWORK

2.1 Theoretical Foundation

The Map Reduce framework first splits an input data file into G pieces of fixed size, typically being 16 megabytes to 64 megabytes (MB). These G pieces are then passed on to the participating machines in the cluster. Usually, 3 copies of each piece are generated for fault tolerance. It then starts up the user program on the nodes of the cluster. One of the nodes in the cluster is special the master. The rest are workers that are assigned work by the master. There are M map tasks and R reduces tasks to assign. M and R is either decided by the configuration specified by the user program, or by the cluster wide default configuration. The master picks idle workers and assigns them map tasks. Once map tasks have generated intermediate outputs, the master then assigns reduces tasks to idle workers. Note that all map tasks have to finish before any reduce task can begin. This is because a reduce task needs to take output from every map task of the job. A worker who is assigned a map task reads the content of the corresponding input split. It parses key/value pairs out of the input data chunk and passes each pair to an instance of the user defined map function. The intermediate key/value pairs produced by the map function are buffered in memory at the corresponding machines that are executing them. The buffered pairs are periodically written to a local disk and partitioned into R regions by the partitioning function. The framework provides a default partitioning function but the user is allowed to override this function by a custom partitioning. The locations of these buffered pairs on the local disk are passed back to the master. The master then forwards these locations to the reduce workers. When a reduce worker is notified by the master about these locations, it uses remote procedure calls to read the buffered data from the local disks of map workers. When a reduce worker has read all intermediate data, it sorts it by the intermediate key so that all occurrences of the same key are grouped together.

The sorting is needed because typically many different keys are handled by a reduce task. If the amount of intermediate data is too large to fit in memory, an external sort is used. Once again, the user is allowed to override the default sorting and grouping behaviors of the framework. Next, the reduce worker iterates over the sorted intermediate data and for each unique intermediate key encountered, it passes the key and the corresponding set of intermediate values to the reduce function. The output of the reduce function is appended to a final output file for this reduce partition. When all map tasks and reduce tasks have completed, the master wakes up the user program. At this point, the Map Reduce call in the user program returns back to the user code.

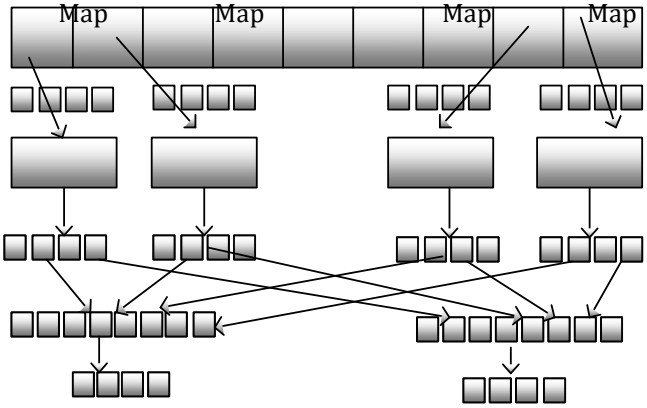

Fig. 2 MapReduce paradigm

There are several tools like Wireshark, tshark etc for displaying the packets. But when it comes to large number of packets say petabytes and terabytes, these tools don't contribute much. Map reduce is the best framework for doing such work. Clients can write suitable map and reduce function for the particular task. As mentioned earlier Map Reduce framework requires a key value pair. In this work key is the source ip ,destination ip address and the protocol and the value is count. Mapper part gets the keys and the value. Reducer part shuffles, sorts and merges and gets the output as the number of packets of specific type from a particular source to destination. It can also be used to find the number of packets send to specific ports.

2.2 Analytics Process Model

Analytics is defined as analytical reasoning supported by highly interactive visual interfaces that involves information gathering, data pre-processing, knowledge representation, interaction and decision making. A process model of visual analytics is illustrated in Fig. 2. According to Fig. 2, the first step is pre-processing such as data cleaning and data transformation over input data to be able to use it in the desired format for further investigations. After the pre-processing step, visualization methods and automated analysis methods are applied to the data.

Afterward, automated analysis methods using data mining methods are applied to generate models. These models can be evaluated and refined by the user through a modification of initial parameters or selecting other type of analysis algorithms. User interaction with the visualization is needed to reveal information by applying different visualization techniques on the data such as descriptive analysis, graphical representations etc. Based on this

interaction, the user can conduct the model building and refinement in the automatic analysis. Furthermore, knowledge can be gained during mentioned different types of user interaction. Finally, the feedback loop stores this knowledge of insightful analyses in the system and enables the analyst to draw faster and better conclusions in the future.

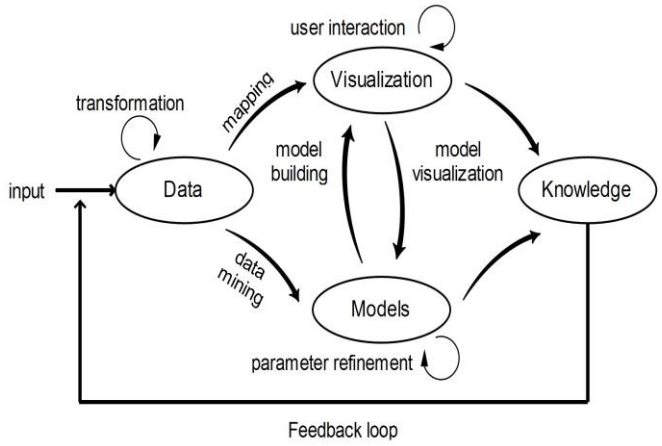

Fig. 3 Analytics Process model

3. PROPOSED SCHEME

Map reduce is the data processing framework. It deals with the implementation for processing and generating large datasets with a distributed algorithm on a cluster . Map reduce is used in big data and Input data is splitted and fed to each node in the map phase. The results generated in this phase are shuffled and sorted then fed to the nodes in the reduce phase. The technique uses the historical information that is being stored in each node and using that information it finds the real slow tasks. Then it maps the slow tasks and reduces the slow tasks.

3.1 Uncertainty aware analytics

Our proposed model is derived from the model of visual analytics presented by in Fig. 4. Input data is collected, transformed and pre- processed, both automatically, through the visualization and the user interaction to be ready in the desired format for the analysis. After pre-processing, one of the main challenges is the selection of an appropriate technique for uncertainty modeling.

The applied technique is based on our previous work in, a self-organizing map for uncertainty visualization in uncertain data sets. We have extended our previous work integrating by MapReduce framework to be able to use the big data for uncertainty modeling and visualization. We add an interactive module in our prototype design that allows refinement of the applied techniques by the user. This prototype also consists of a graphical representation to support uncertainty visualization as well as a descriptive analysis for knowledge representation to draw conclusion.

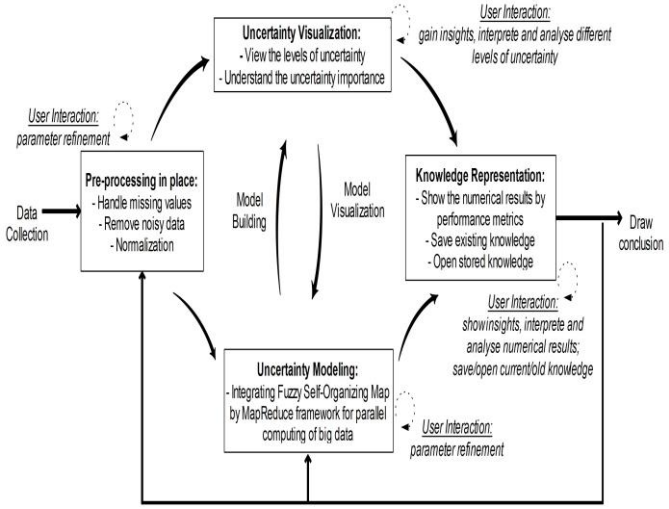

Fig. 4 Uncertainty aware analytics in Bigdata

3.2 Self Organizing Map (SOM) using C-Mean

Our proposed uncertainty modeling is derived from Self-Organizing Map (SOM). We proposed a self- organizing map algorithm using fuzzy c-mean (FCM) to model uncertainties based on a centralized-batch processing framework. SOM works in three phases. In the first phase, FCM technique has been employed to assign a membership degree in clusters' centers in terms of the input data. Then in the second phase, all the clusters' centers cooperate by a Gaussian function with their neighbors in terms of the membership degree. Finally at the third phase, all the centers' positions are updated. These three phases are repeated, until the maximum number of iterations is reached or the changes become smaller than a predefined threshold. First, in this section we present the main design for parallel SOM based on MapReduce framework for a decentralized-batch processing which is depicted. Then we explain how the necessary computations can be formalized as map and reduce operations in detail.

3.2.1. Mapping Function

The input data set is stored in Bigdata. Data in Bigdata is broken down into smaller pieces (called chunks) and distributed throughout the cluster. In this way, the map and reduce functions can be executed on smaller subsets of larger data sets, and this provides the scalability that is needed for the big data processing. MapReduce reads a single chunk of data on the input datastore, then call the map function to work on the chunk. The map function then works on the individual chunk of data and adds one or more key-value pairs to the intermediate KeyValueStore object. MapReduce repeats this process for each of the chunks of data, so that the total number of calls

to the map function is equal to the number of chunks of data. Each mapper runs SOM algorithm. The result of this phase is a KeyValueStore object that contains all of the key-value pairs added by the map function. The key is the cluster centers and the corresponding values are the position of centers in each mapper, the membership degree of each center, and the membership degree of each center for different target classes. After the map phase, MapReduce prepares for the reduce phase by grouping all the values in the KeyValueStore object by unique key in the intermediate phase.

3.2.2. Reduce Function

The reduce function scrolls through the values from the KeyValueStore to perform a summary calculation. We calculate the average of aggregated values to sum up the results. The MapReduce framework is repeated until the clusters' centers do not change any more in the predefined number of iteration

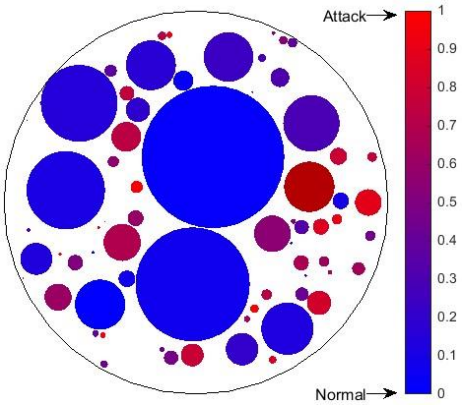

Fig.5 Uncertainty in Bigdata

3.3 Performance

The more uncertain a cluster is, the more impure is its visual representation. For instance, the purple color denotes a 100% uncertainty in a formed cluster (purity = 0.5), neither completely normal nor attack traffic. This is useful for discovering the sources of uncertainty. This visualizes the effect of uncertainty and steers the user's attention towards the most reliable clusters over uncertain data points so that only the most reliable clusters are highlighted to the user. On the other hand, a large size of a node denotes the more uncertain data involved while a small size of a node denotes the less uncertain data involved which can be interpreted as outliers. As a consequence, these small nodes steer the user's attention visually towards the most unreliable nodes as outliers. This prototype design displays a high-level view of entire uncertain big data together with the numerical results. Preliminary results show that the designed prototype produces satisfactory outcomes. Users can steer and control uncertainty based on their own practices or analytic needs in the data preparing step, find outliers visually as well as distinguish visually reliable and unreliable clusters. User evaluations by zooming into sub-regions of clusters and reveal more details (i.e., details on demand) will be carried out in the future.

4. CONCLUSION

In this paper, we propose a framework for uncertainty-aware visual analytics in the big data. We integrated a self-organizing map algorithm with MapReduce framework in order to execute a parallel computing on big data. The prototype system includes a set of interactive visual representations that supports the analysis of the uncertain data and user interaction. We believe that this prototype system is useful when the analyst wants to extract a model that explains the behavior of uncertain data, find outliers visually and makes insightful decisions.

5. FUTURE WORK

The future work is needed by more user evaluations: zooming into sub-regions of uncertain clusters and reveal more details.

REFERENCES

1. [1] S.Ezhilarasi, "HHH- A Hyped-up Handling of Hadoop based SAMR-MST for DDOS Attacks in Cloud", International Research Journal of Engineering and Technology (IRJET) , Vol 05 Issue 03, March 2018.

2. [2] Bendler, J., Wagner, S., Brandt, T., Neumann, D.: Taming uncertainty in big data. Business & Information Systems Engineering 6(5), 279–288, 2014

3. [3] Grolinger, K., Hayes, M., Higashino, W.A., L'Heureux, A., Allison, D.S., Capretz, M.: Challenges for mapreduce in big data. In: IEEE World Congress on Services (SERVICES). pp. 182–189, 2014

4. [4] Karami, A., Guerrero-Zapata, M.: Mining and visualizing uncertain data objects and network traffics by fuzzy self-organizing map. In: Proceedings of the AIC workshop on Artificial Intelligence and Cognition. pp. 156–163, 2014

5. [5] Karami, A., Guerrero-Zapata, M.: An anfis-based cache replacement method for mitigating cache pollution attacks in named data networking. Computer Networks80, 51–65 2015

6. [6] Karami, A., Guerrero-Zapata, M.: A fuzzy anomaly detection system based on hybrid pso-

kmeans algorithm in content-centric networks. Neurocomputing 149, Part C, 1253–1269, 2015

7. [7] Keim, D.A., Bak, P., Bertini, E., Oelke, D., Spretke, D., Ziegler, H.: Advanced visual analytics interfaces. In: Proceedings of the International Conference on Ad- vanced Visual Interfaces. pp. 3–10, 2010

8. [8] LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big data, analytics and the path from insights to value. MIT sloan management review 21, 2013.

9. [9] Lee, K.H., Lee, Y.J., Choi, H., Chung, Y.D., Moon, B.: Parallel data processing with mapreduce: a survey. In: AcM sIGMoD Record 40. pp. 11–20, 2012

10. [10] Qian, H.: Pivotalr: A package for machine learning on big data. R Foundation for Statistical Computing 6(1), 57–67, 2014

11. [11] Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., Keim, D.: Visual analytics for the big data era – a comparative review of state-of-the-art commercial systems. In: IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 173–182, 2012

12. [12] Keman Huang, Jianqiang Li, and MengChu Zhou, Jan- 2015," An Incremental and Distributed Inference Method for Large-Scale Ontologies Based on MapReduce Paradigm".

13. [13] Dean J and Ghemawat S, 2008-"MapReduce: Simplified data processing on large clusters," Commun. ACM, vol. 51, no. 1, pp. 107–113,2008.

14. [14] Correa, C.D., Chan, Y.H., Ma, K.L.: A framework for uncertainty-aware visual an- alytics. In: IEEE Symposium on Visual Analytics Science and Technology (VAST). pp. 51–58. 2009

15. [15] Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual an- alytics: Scope and challenges. Springer Berlin Heidelberg, 2008.