An Approach for Detecting and Correcting Errors of Big Sensor Data

Neela S[1], Dr. Chandramouli H[2],Dr. B.R.Prasad Babu[3],Mr.Pramod[4]

MTech Student, Department of CSE ,R&D Centre, East Point College of Engg & Technology[1]

Professor, Department of CSE,East Point College of Engg & Technology[2]

Professor & Head of Department of CSE,East Point College of Engg & Technology[3]

Associate Professor of CSE,East Point College of Engg & Technology[4]

[1]neelashivkumar@gmail.com, [2]hcmcool123@gmail.com ,[3]brprasadbabu@gmail.com,[4]pramod741230@gmail.comAbstract

Sensor data in scientific analysis applications with high volume, velocity, is difficult to handle using database management tools. For instant data error detection and correction, we develop novel data error detection approach which exploits computation power of hadoop-big data platform, network feature of WSN that is temporial-spatial data blocks.

Keywords – Sensor networks, Error detection, Map Reduce, Cloud

I. Introduction

Of late, new era of data explosion which brings about new challenges for big data handling . In common, big data [i], [ii] is a collection of data sets so large and complex that it becomes very difficult to deal with on hand database management systems or traditional data processing applications. It represents the progress of the human cognitive processes, usually includes data sets with sizes beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time [i], [ii], [iii], [vii], [v], [viii], [iv], [vi], [xv], [xvi], [xvii], [xviii]. Big data basically has characteristics of five 'V's, Volume, Variety, Velocity, Veracity and has Value. Big data sets come from many areas, including meteorology, complex physics simulations, genomics, biological study, gene analysis and environmental research [i], [ii]. According to literature [i], [ii], since 1980s, initiated data doubles its size in every 40 months all over the world. Hence, processing the big data has become an elementary and critical challenge of present evolving society. Big data-Hadoop provides a convincing platform for big data processing with powerful handling capability, storage, scalability, resource reuse and low cost.

One of the important source for scientific big data is the data sets collected by wireless sensor networks (WSN). Wireless sensor networks are more efficient in increasing people's ability to monitor and interact with their physical environment. Big data set from sensors is often prone to corruption and losses due to wireless medium of communication and presence of erroneousness hardware in the nodes.WSN can be categorized as a form of complex networks systems [ix]. These complex network systems[ix], [x], [xiii], such as WSN and social network, will be prone to data abnormality and error become an bothersome issue for the real network applications [xiv], [xi], [xii]. Therefore, the issue of how to find data errors in complex network systems for improving and debugging the network has attracted the interests of researchers.

WSN big data error detection requires powerful real-time processing and storing of the huge sensor data additionally dealing with in the context of using inherently complex error models to spot and locate events of abnormalities. In this paper, aim is to develop a novel error detection approach by exploiting the massive storage, scalability and computation power of Big data Hadoop to detect errors in big data sets from sensor networks.

The proposed error detection approach in this paper will be based on the classification of error types. Specifically numerical data errors are set down and introduced in bigdata Hadoop error detection approach. The defined error model will trigger the error detection process, compares to previous error detection of sensor network systems. Hadoop will be designed and developed by utilizing the massive data processing capability. In addition, the architecture feature of complex networks will also be analyzed to combine with the Hadoop parallel computing with a more efficient way. Sensor network is a kind of scale-free complex network system which matches Hadoop scalability features.

II Material And Methodology

A. Classification Of Error Types In WSN Data Sets

In this paper, we focus on error detection for numeric big data sets from complex networks.. Considering specific feature of numeric data errors[x], there are several unusual data scenarios demonstrated in Fig. 1.

The "flat line faults", a time series of a node in a network system keeps unchanged for unacceptable long time duration.

The "out of data bounds faults" indicates impossible data values are observed based on some domain knowledge.

The "data lost fault" means there are missing data values in a time series during the data generation or communication.

The "spike faults" indicates in a time series data items which are totally out of the prediction and normal changing trend. The above four errors happen both at data generation and

doi : 10.17950/ijer/v5i4/033

exchange stages, hence errors are categorized into node and edge side.

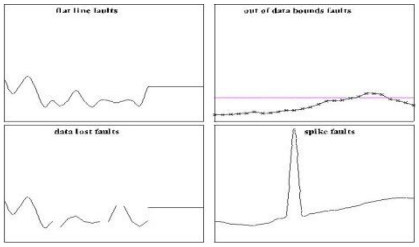


Fig. 1. Error scenarios from sensor systems.

B. Error Definition

We present classification, definition of error type to guide error detection algorithm. Suppose a data record from node is denoted as r(n, t, f(n, t), g(n, l)), where n is ID of node in network systems. t represents window length of time series. f(n, t) is numerical values collected within window t from node n. g(n, l) is a location function which records the cluster, the data source node and partition situation related to node n.

Definition 1 (Node/Edge side flat line error). Let ri(ni,ti), f(ni,ti), g(ni,l) be a time series recorded from node ni, where i is a time stamp. If any element x ≡ δi,where δi is an effective constant during time window t, x € f(ni,ti) and g(ni,l) =0/ g(ni,l) !=0, ni is the data source node,there is an node side/edge side flat line error.

Definition 2 (Node/Edge side data lost error). Let ri (ni,ti) , f(ni,ti), g(ni,l) be a time series recorded from node ni, where i is a time stamp.If f(ni,ti) = null && t $_i$ > ι ,ι is the time duration from outside application requirement, and if g(ni,l) = 0/ g(ni,l) != 0, ni is the data source node, the error is a node/edge side data lost error.

Definition 3 (Node/Edge side out of bounds error). Let r_i (n_i,t_i), f(n_i,t_i), g(n, l) be a time series record from node n_i, where i is a time stamp. If any element x > θ , x € f(n_i,t_i), θ is a threshold defined from the application requirement, and if g(n_i,l) = 0/ g(ni,l) != 0; n_i is the data source node, the error is a node/edge side out of bound error.

Definition 4 (Node/Edge side spike error). Let r_i (n_i,t_i), f(n_i,t_i), g(n, l) be a time series record from node n_i, where i is a time stamp. If | f(n_i,t_i) – f^p(n_i,t_i)| / t_i > $_{\Psi}$,$_{\Psi}$ is the acceptable changing trend, f^p(n_i,t_i) is the predicted time series with an adopted prediction model, and if g(n_i,l) = 0/ g(ni,l) != 0, n_i is the data source node, the error is a node/edge side spike error.

III Results and tables

A. Model Based Error Detection for Sensor Network Data.

During the filtering of big data sets, whenever an anomalous data is encountered, detection algorithm have to perform two tasks. They are illustrated as two functions here. “ f_d(n/e,t) ” is decision making function which determines whether detected anomalous data is a true error. In other words, f_d(n/e,t) has two outputs, “false negative” for detecting a true error and “false positive” for selecting a non-error data. “f_l(n/e,t)” is a function for tracking and returning original error source. As above two functions results with error detection the process gets successfully completed.

As shown in Fig. 3, there is a complex network and cloud platform which runs map-reduce technique for running error detecting algorithms. Without any consideration of network features and data characteristics, the error detection algorithm needs to filter whole big data set from the network. Whenever, an anomalous is encountered, algorithm going to call f_d(n/e,t) and f_l(n/e,t) to go accross the whole network big data set for the final decision making and error source location.

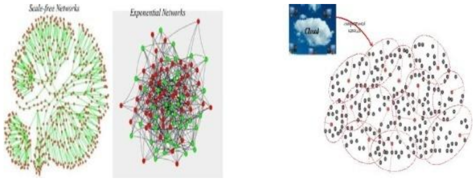


Fig. 3. scale-free and non scale-free networks.

However, we know that scale- free network systems networks contains a clustering and hierarchical topology. Only little nodes in the whole network have more sets of links to other nodes. So, with these nodes, the whole networks can be splitted into a group of clusters (red circles). If there is certain anomalous data occurs for a certain node k, the high opportunity is that most of the related data for f_d(n/e,t) and f_l(n/e,t)” will be located in the clusters where node k locates. As a result, f_d(n/e,t) and f_l(n/e,t) only need to move across the related clusters for error detection result. This is because of fact that except for a few central nodes, most of nodes only have limited links within themselves in their clusters.

B. ALGORITHMS

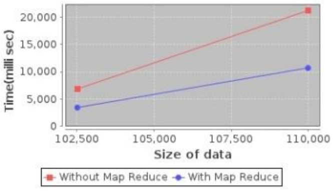
To deploy the proposed error detection model to locate the error, the algorithm can be splitted into two parts, detection and location. In this section, we introduce the big data error detection/location algorithm, and its combination strategy with Hadoop's Map reduce technique.

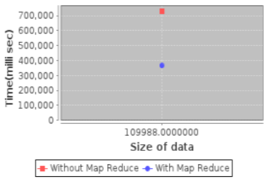
B.1 Error Detection

We have a two-phase approach to conduct the computation required in the whole process of error detection and localization. On the phase of error detection, there are three inputs for the error detection algorithm. The first is the graph of network. The second is the total collected data set D and the third is the defined error patterns p. The output of the error detection algorithm is the error set D'.

5.2 Error Localization

After error pattern matching and error detection, it is important to locate the position and source of the detected error in the original WSN graph G(V, E). The input of the Algorithm 2 is original graph of a scale-free network G (V, E), and an error data D from Algorithm 1. The output of the algorithm 2 is G’(V’, E’) which is the subset of G to indicate the error location and source.





CONCLUSION

In this paper the sensor data values are collected and processed, where we do error detection. For detecting errors, we approach for spatial and temporal correlation models .The error correction will be the future work. It shows a sign of fast Error Recovery as modern technique Map reduce is adopted.

ACKNOWLEDGEMENT

I would like to express my sincere thanks to ***Dr. B M Satish,*** Principal, East Point College of Engineering and Technology, ***Dr. Prasad Babu B R***, HOD, Department of CSE, ***Dr. Chandramouli H,*** Professor and PG Coordinator, Department of CSE for his guidance to complete this paper work on time.

REFERENCE

[i] S. Tsuchiya, Y. Sakamoto, Y. Tsuchimoto, and V. Lee, Big Data Processing in Cloud Environments," FUJITSU Science and Technology J., vol. 48, no. 2, pp. 159-168, 2012.

[ii] "Big Data: Science in the Petabyte Era: Community Cleverness Required," Nature, vol. 455, no. 7209, p. 1, 2008.

[iii] X. Zhang, C. Liu, S. Nepal, and J. Chen, "An Efficient Quasi-Identifier Index Based Approach for Privacy Preservation over Incremental Data Sets on Cloud," J. Computer and System Sciences,vol. 79, pp. 542-555, 2013.

[iv] W. Dou, X. Zhang, J. Liu, and J. Chen, "HireSome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big Data Applications," IEEE Trans. Parallel and Distributed Systems, 2013

[v] X. Zhang, T. Yang, C. Liu, and J. Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using Systems, in MapReduce on Cloud," IEEE Trans. Parallel and Distributed, vol. 25, no. 2, pp. 363-373, Feb. 2014.

[vi] C. Yang, X. Zhang, C. Zhong, C. Liu, J. Pei, K. Kotagiri, And J. Chen, "A spatiotemporal compression based approach forn efficient big data processing on cloud,"

[vii] X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, "A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-effective Privacy Preserving of Intermediate Datasets in Cloud," IEEE Trans. Parallel and Distributed Systems, vol. 24, no. 6, pp. 1192-1202, June 2013.

[viii] C. Liu, J. Chen, T. Yang, X. Zhang, C. Yang, R. Ranjan, and K. Kotagiri, "Authorized public auditing of dynamic big data storage on cloud with efficient verifiable fine grained updates," IEEE Trans. Parallel and Distributed Systems, vol. 25, no. 9, pp. 2234–2244, Sept. 2014.

[ix] R. Albert, H. Jeong, and A. L. Barabasi, "Error and Attack Tolerance of ComplexNetworks," Nature, vol. 406, pp. 378-382, July 2000.

[x] "Big Data Beyond MapReduce: Google's Big Data Papers,"http://architects.dzone.com/articles/big-data-beyond-mapreduce,accessed Mar. 2013.

[xi] K. Ni, N. Ramanathan, M.N.H. Chehade, L. Balzano, S. Nair, S. Zahedi, G. Pottie, M. Hansen, M. Srivastava, and E. Kohler, "Sensor Network Data Fault Types," ACM Trans. Sensor Networks, vol. 5, no. 3, article 25, May 2009.

[xii] S. Slijepcevic, S. Megerian, and M. Potkonjak, "Charaterization of Lacation Error in Wireless Sensor Networks: Analysis and Application," Proc. the Second Int'l Conf. Information Processing in Sensor Networks (IPSN '03), pp. 593-608, 2003.

[xiii]D. Xiong, M. Zhang, and H. Li, "Error Detection for Statistical Machine Translation Using Linguistic Features," Proc. 48th Ann. Meeting of the Association for Computational Linguistics (ACL'10), pp. 604-611, 2010.

[xiv] S. Mukhopadhyay, D. Panigrahi, and S. Dey, "Model Based Error Correction for Wireless Sensor Networks," IEEE Trans. Mobile Computing, vol. 8, no. 4, pp. 528-543, Sept. 2008

[xv] N. Laptev, K. Zeng, and C. Zaniolo, "Very Fast Estimation for Result and Accuracy of Big Data Analytics: The EARL System," Proc. IEEE 29th Int'l Conf. Data Eng. (ICDE), pp. 1296-1299, 2013.

[xvi] X.L. Dong and D. Srivastava, "Big data integration," Proc. IEEE 29th Int'l Conf. Data Eng. (ICDE), pp. 1245-1248, 2013.

[xvii] T. Condie, P. Mineiro, N. Polyzotis, and M. Weimer, "Machine learning on Big Data," Proc. IEEE 29th Int'l Conf. Data Eng. (ICDE), pp. 1242-1244, 2013.

[xviii] A. Aboulnaga and S. Babu,"Workload Management for Big Data Analytics," Proc. IEEE 29th Int'l Conf. Data Eng. (ICDE), p. 1249 2013.