Comparative Study on Accurate Recovery of Internet Traffic Data

Jithya K V
Department of CSE
SJEC
Mangaluru, Karnataka, India

Ms Supreetha R
Department of CSE
SJEC
Mangaluru, Karnataka, India

***Abstract*— Data Traffic is the amount of data passing through the network in a given point of time. The data may get lost due to the traffic or because of other reasons. Main goal of this paper is to recover from the data loss in the network traffic. There are different methods used to recover data lost in network. . Existing methods cannot provide 100 percentage of accurate recovery of data traffic. Also the other methods cannot find out the hidden structure for the data recovery. This paper describes the Sequential Tensor Completion method to recover the lost data efficiently. In order to reduce the complexity a 3 way tensor method is incorporated.**

I. INTRODUCTION

The information passing through the network during certain period of time is known as Network Data Traffic. There is a chance of missing of data or data loss due to the traffic. The main reason for the traffic is the inference of malware function, videoconferencing, video streaming etc. There are different methods used to recover from the data loss in the network traffic. In this paper represents a survey on different methods used for accurate recovery of data loss in network traffic. The existing methods are using different algorithm for the recovery. But all the methods cannot efficiently find out 100 percentage recovery of data and the complexity of the entire algorithm is very high. This survey discovers a 3 way traffic tensor method to recover from the data loss in the network traffic. This method gives 100 percentage recoveries of data. The set of input data is converted into a 3 dimensional matrix format to find out where the loss of data has occurred. To reduce the complexity, the Sequential Tensor Completion (STC) algorithm is used [7]. To more accurately recover the data exploiting the feature of the dynamic data, Dynamic Sequential Tensor Completion algorithm (DSTC) based on STC is used.

Tensors are the higher-order generalization of vectors and matrices. Tensor-based multi linear data analysis has shown that tensor models can take full advantage of the multi linear structures to provide better data understanding and information precision. Tensor-based analytical tools have seen applications for web graphs, knowledge bases, chemo metrics, signal processing, computer vision, anomaly detection etc.

Two dimensional matrix based algorithm also used to recover from the data loss in the network traffic. But by using this algorithm it cannot find out all the hidden structure for the data.

In this paper the proposed algorithm consider both test data and train data for comparing with each other. The data set are the input of the algorithm. The data set is the collection of huge volume of data which is generally depicted in table form. Column symbolizes the key elements of the traffic data and Row defines the value of the element.

A. Train and Test Data

Test data is the part of input data which is used to compare with the trained data. Test data contains a part of input data. Train data is the complete data. The test data is taken from the input data based on critical value which is used to split. The data mining techniques are mainly used in this method for all the collection data process.

B. Data Mining

Data mining is the operation of taking out the desired information from the collection of row of information. There are different steps are included in this activity.

- Data Collection: It is the procedure of collecting information from the different field. Mainly it checks different matters and chances, based on that the information’s are collected.
- Data Preprocessing: Cleaning of information is also known as the preprocessing. The data is converts into another format which is easy to understand.
- Data Selection: Procedure of selecting data for the data extraction. Based on the selection procedure the data extraction is taking place.
- Data Extraction: It is the action to get the data from the large data source for further process.

C. Network Traffic Data

The information is passing through the network in a period of time is known as Network Data Traffic. There is a chance of missing of data due to the traffic. The reason for the data loss in the network traffic is described below:

- File Sharing: One of the major reasons for the traffic is the file sharing. While sharing the file the information is transmit into the node for the users for the connection. So due to the connection there is a chance to slow the sharing of information.
- Streaming Media: In YouTube and other media it provides the audio and video for the users. So the data is in the low bit rate. In big companies there are thousands of users accessing the data in low mode. So there is a chance of traffic.
- Videoconferencing: In videoconferencing there are multiple users connecting each other in a network for the video. So it lowers the bandwidth and the congestion occurs.

- Malware: There are various kinds of malwares are present to lower the performance of the internet. So in order to protect from the viruses software of antivirus is install in the system.
- Network dis-connectivity: During transmission of data the connectivity between the source and destination may get loss. Due to that the loss of data or traffic may occur.
- User Movement: The user travels from one location to another location. So while travelling there is a chance of loss of data may occur.

II. LITERATURE WORK

Matrix-completion-based algorithms are proposed to recover the missing traffic data by exploiting both spatial and temporal information [1]. This algorithm is mainly works in the 2 dimensional matrix forms. But it cannot find out the hidden structure for the data. The main disadvantage of this method is performance is very low when the missing data is very large. The advantage of this method is that it finds out both temporal and spatial information and also the performance is very high when missing data is low. The analysis of spatial and time features tells that every time users visit the same cite of internet in same time. So there is a chance of calculating the measurement in same time of different days may alike. The disadvantage of this method is difficult to align the matrices of different days.

Vehicles are mainly using sensors to find out the road condition. So the data collected by the sensors in the vehicle is aggregated and the whole data is passes to another vehicle. The CS novel Scheme is used in this method. This method is proposed by K Xie, W Luo [2]. The vehicle which is passing through the same road gets the data of the other vehicle which is already passing through that road. The main advantage of this scheme is data missing is very low but the network congestion is very high.

The network applications are increases day by day. Mainly these applications are used for chatting, video conferencing etc. So the delay in network also begins to be a problem. The olden days mainly used sampled measurement using matrix factorization. B Liu, et al. proposed Network Coordinate System (NCS) to find out the latency [3]. The Network Coordinate System assigns each host into the coordinate space for predict the delay in network by calculating the distance between the host. The drawback of this method is the prediction between every three hosts may have triangle inequality and also the geographical distances between hosts that dictate propagation. So there is a chance of missing useful information.

To find out the data in the network the network connectivity can be used. So for determining the source destination connectivity basically graphs are used. Luoyi Fu proposed combinatorial optimization problem [4] to find out the connectivity between source and destination. In this method the computational complexity is represented in graphs. This problem cannot be solved in polynomial time. This method does not work in dynamic programming algorithm with exponential time complexity.

There are different methods to find out the network delay between the source and destination. All methods are mainly used to find out the latency between the source and destination when it is static condition. It is very difficult to find out the latency when the source and destination dynamically changes i.e., while the laptop and other system are moving from one location to another location the latency prediction is little bit difficult. Zhu, et al. [5] proposed dynamic latency prediction method to find out the latency in dynamic situation. The delay of the new frame is checks by comparing with the previous frames. The information set sampled in multiple time periods so that the accuracy of prediction is increases. The drawback of this method is the current frame is not effective enough to capture the changing latencies.

Kun Xie et al. proposed Reshape-Align scheme to form tensor for the accurate recovery of internet traffic data [6]. The temporal features and user features also considered for the tensor completion. This method has efficient recovery and also better performance in terms of several matrices. Main disadvantage of this method is whenever the frequency changes it is difficult to predict the accuracy.

To find out the missing of data in the network and recover from the data loss Kun Xie, et al proposed 3 way traffic tensor methods [7]. 3way traffic tensor catches all the features of network data traffic. Mainly this method uses to algorithm which is known as The Sequential Tensor Completion (STC). The algorithm are mainly used because to speed up the recovery process. The computation cost of this algorithm is also less. It also finds out the bandwidth, packet transmitted, and delays of the data set.

III. PROPOSED WORK

To find out the missing data and recover from the data loss the 3 way tensor method is used. 3way traffic tensor is mainly used because it finds out the hidden structure for the data. The two algorithms are mainly used in this method Sequential Tensor Completion (STC) and Dynamic Sequential Tensor Completion (DSTC) algorithm. Both algorithms reduce the complexity. The main objective of this scheme is given below

- Uploading the data set.
- Preprocessing.
- Feature Extraction.

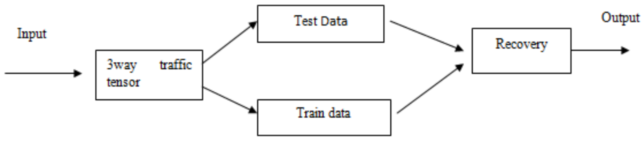

Fig 1.Architecture of recovery of traffic data.

The above figure shows the architecture of accurate recovery of traffic data by using STC method. The input data is converted into orthogonal matrix by using sequential tensor algorithm and the data again split into test data and train data. Based on that it checks any missing or loss age of data occurs. By using DSTC algorithm the recovery process occurs for dynamic changes. The main advantage of this method is efficiently finds out the missing data recovery and also the find out the hidden structure for the data traffic.

The recovery process by using these two algorithm increases the speed of recovery and also find all the hidden features of the network traffic data.

A. Sequential Tensor Completion algorithm

Sequential Tensor Completion (STC) is mainly used for the better performance. The algorithm mainly used to split the data set into two test data and train data. The dataset is converted as matrix format and test data and train data compare with each other. After comparing both data missing of data can be identifies. So the missing data is adds to the remaining data for the full recovery. The algorithm [1] is given below:

```
1. Consider the input matrix
2. For each column i= 1 to n
    3. Update each column in the orthogonal matrix by using the equation
       $(U\ (_{t+1})) = (U\ (_{t+1})) + ((\cos\ (\sigma\varphi) - 1))\ P/|P|\ +\sin\ ((\sigma\varphi))l/|l|\qquad w^{tr}/|w|$
4. End for
5. Return traffic tensor
6. Stop
```

B. Dynamic Sequential Tensor Completion Algorithm

Dynamic Sequential Tensor Completion (DSTC) algorithm is derived based on STC. To find out the more feature of the dynamic data DSTC algorithm is proposed. In DSTC it takes more times to check each column in the matrix by testing. The least square minimization problem is introduced in this algorithm. In DSTC multiple rounds are used to measure the data .It increases the accuracy of this algorithm. Also these two algorithms find out the recovery, Error during transmission, Recovery computation time, Loss etc.

C. Performance Analysis

The accuracy of these two algorithms is calculated by using F-Measure method. It is harmonic combination of precision value in addition to recall value used for information extraction. “Precision is the fraction of retrieved instances that are relevant” and “recall is the fraction of relevant instances that are retrieved”. High recall rate indicates an algorithm returned most of the relevant results. High precision rate indicates an algorithm returned more relevant results than irrelevant.

The equation of F- measure method is given below:

$$\text{F1 Measure} = \frac{\text{True Positive + True Negative}}{\text{False positive + False Negative}}$$

IV. COMPARISON

Table 1. Comparison on Various Approaches

Author(s)&Ref	Title	Efficiency
Kun Xie, et al. [1]	Matrix-completion-based algorithm.	Efficiency is 100 percentage when the missing ration is very low.
K Xie, et al.[2]	A novel CS-Sharing scheme.	Efficiency is very less.
B Liu, et al.[3]	Network Coordinate System.	To predict the latency efficiency is high but for recovery efficiency is low.
L Fu, et al.[4]	Combinatorial optimization problem.	Efficiency is high to predict the latency in statistically and dynamically. But efficiency for recovery is less.
Zhu, et al.[5]	Approximate tensor completion scheme.	Efficiency is high.
Kun Xie, et al[6]	Sequential tensor completion	100 percentage of efficiency in recovery procedure.
Kun Xie, et al[7]	Reshape-Align scheme	Efficiency is high.

In table 1 shows the efficiency of the different method. The efficiency of each method is different and also the value cannot be calculated.

IV. CONCLUSION

Missing of data in the network is very high due to the traffic. So this paper shows the comparison work of different methods to recover from the data loss. The best method is found by using Sequential Tensor Completion algorithm. Each method has advantages and disadvantages also. But comparing to other method Sequential Tensor Method is the best. The method provides efficient accurate recovery of missing data. Also it finds out the hidden structure for the data traffic.

REFERENCES

[1] K. Xie, C. Peng, W. Gang and G. Xie, "Accurate recovery of internet traffic data under dynamic measurements.", In IEEE INFOCOM 2017-IEEE Conference on Computer Communications 2017 May 1, pp. 1-9.

[2] K .Xie, W. Luo, X. Wang, D. Xie and J Cao, "Decentralized context sharing in vehicular delay tolerant networks with compressive sensing.", In2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS) 2016 Jun 27 , pp 169-178.

[3] B. Liu, D .Niu, Z. Li and HV. Zhao, "Network latency prediction for personal devices: Distance-feature decomposition from 3D sampling”, In 2015 IEEE Conference on Computer Communications (INFOCOM) 2015 Apr 26 pp. 307-315.

[4] L. Fu, X. Fu, Z. Xu, Q. Peng and X .Wang, "Determining source–destination connectivity in uncertain networks: Modeling and solutions”, IEEE/ACM Transactions on Networking. 2017 Dec; 25 pp.3237-3252.

[5] R. Zhu, B. Liu, D. Niu and Z. Li, "Network latency estimation for personal devices: A matrix completion approach. IEEE/ACM Transactions on Networking.” IEEE/ACM Transactions on Networking 25, no. 2 (2017), pp 724-737.

[6] K. Xie , L. Wang, X. Wang, G.Xie, J. Wen, G.Zhang, J.Cao, and Dafang Zhang , "Accurate recovery of internet traffic data: A sequential tensor completion approach.", IEEE/ACM Transactions on Networking (TON). 2018 Apr 1; 26, pp 793-806.

[7] Xie, K., Peng, C., Wang, X., Xie, G., and Wen, J. "Accurate recovery of internet traffic data under dynamic measurements." In IEEE INFOCOM 2017 May-IEEE Conference on Computer Communications pp. 1-9.