REAL-TIME ANALYSIS OF MOBILE INTERNET DATA

Priyanka Pendharkar¹, P S Metkewar²
¹Student, MBA-IT, SICSR, Affiliated to Symbiosis International University(SIU), Pune-16
²Associate Professor, SICSR, Affiliated to Symbiosis International University(SIU), Pune-16

Abstract - Mobile Internet Data traffic is increasing on daily basis with the increase in popularity and usage of mobile Internet. Analysis of this big amount of data produced daily i.e. in terabytes or petabytes is an issue. Real-time analysis of data is required at the operational data sources, to pass a reduced subset of data to the data warehouse for further prediction. The load on data warehouse will be reduced. This would be useful for quicker analysis of people involved in using mobile Internet, which can be used by the state for security and e-commerce enablement. This paper emphasizes on resolving the issue of analysis of data before storage by enabling real-time analysis.

Keywords: — Mobile Internet data, data Warehouse, integration, real-time analysis, operational data store.

I. INTRODUCTION

Internet Users prefer using mobile Internet due to the ease of usage and access. Mobile data traffic in 2013 had crossed 18000 PB. An authoritative prediction shows that growth of the mobile data traffic will be a CAGR of 61% in the period of 2013 to 2018. Effective management of this “explosion” of Internet traffic is required.

The amount of data that would potentially be transferred into the data warehouse is huge. Reducing the data that enters into the warehouse by enabling only useful data is a need. This reduction can be achieved by enabling 2 levels of analysis through machine learning. One level can be achieved by doing predictive analysis on mobile Internet data, before it is transferred into the data warehouse. This enables to send only data about outliers detected, to be sent to the warehouse. Second level of predictive analysis will be on only this subset of data (containing all outliers) in the data warehouse. This method enables and achieves a few things: (1) Capacity management of data warehouse. (2) Increased speed of detection. (3) More precision of the results obtained.

1.1 Machine Learning

Machine learning deals with study and construction of algorithms that learn from data. They do not completely rely on scientific calculations. They rather have a power to observe the inputs and their respective outputs, and make predictions for any type of data imposed on them.

Machine leaning is linked to artificial intelligence and finds application in spam filtering, search engines, detection of patterns and frauds. Machine Learning tasks are classified as Supervised Learning, Unsupervised Learning and Reinforced Learning.

1.1.1 Supervised Learning

Supervised learning is a specific task of machine learning. It uses the method of inferring a function from labeled training data. A set of training examples are contained in the training data. Every example is a pair that consists of an input object (which is a vector) and a desired output value. A supervised learning algorithm produces an inferred function by analyzing the training data, which can be used for mapping new examples. A scenario which is optimal will allow for the algorithm to correctly determine the class
labels. The class labels may also be determined for unseen situations or instances. By learning the way different datasets behave, the algorithm can detect potential diversion of normal behavior, which may be fraudulent behavior present in the dataset.

Approaches to Supervised learning include Classification methods.

Classification method consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a set of training data called as training set, containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to unveil and discover relationships between the attributes that would make it possible to predict the outcome. In the next step, the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm then analyses the input and produces a prediction. Some common classification techniques include: Neural Networks, Support Vector Machines, Bayesian Classifiers and Decision Trees.

II. PROBLEM STATEMENT

This paper throws light on the application of real-time analysis of mobile Internet data using statistical techniques, before the data is entered into the data warehouse. The reason why this data goes through real-time analysis is that, the data warehouse has a limited storage; moreover, the amount of mobile Internet traffic produced daily is huge. Hence, storing all the mobile traffic in one warehouse is an issue. Real-time analysis of such data will help in reducing the amount of data that gets entered into the database and helps in increasing the efficiency of results obtained through analysis from the warehouse.

III. LITERATURE REVIEW

3.1 Research Papers
3.1.1 Breaking into Social Nervous System: Architecture for Reality Mining
Deepak R. Vashisth from the Department of IT, Rohtak, India has gathered and analyzed the communication information shared between the people to study their behavioral patterns. For this he has used Reality Mining that collects the digital traces that act as leftovers by the people while their communication. This enabled analysis of behavior between known groups(intragroup) as well as unknown groups(intergroup).

3.1.2 Characterizing User Behavior in Mobile Internet
Jie Yang, Yuanyuan Qiao, Xinyu Zhang, Haiyang He, Fang Liu from China and Gang Cheng from USA have collected mobile traffic data from 2G and 3G networks from the metropolitan area of China to determine the usage of mobile Internet. The study focused on the behavior of users based on their data usage, application usage and mobility patterns. They analyzed that high mobility triggered more usage. The study allows to improve the access to mobile Internet for mobile users.

3.1.3 Traffic Prediction in 3G Mobile Networks Based on Multifractal Exploration
Yanhua Yu, Meina Song, Yu Fu, and Junde Song have investigated the forecasting of traffic for 3G data services in mobile networks. They investigated that internet traffic can be modeled using Fractional AutoRegressive Integrated Moving Average (FARIMA) process. Their studies showed that FARIMA process failed sometimes. They came up with an integrated solution AutoRegressive Moving Average (ARMA) and FARIMA. The results proved the effectiveness of the solution. Hence, effective solution for traffic prediction for network planning and optimization was proposed.

3.1.4 MobSafe: Cloud Computing Based Forensic Analysis for Massive Mobile Applications Using Data Mining
Jianlin Xu, Yifan Yu, Zhen Chen, Bin Cao, Wenyu Dong, Yu Guo, and Junwei Cao have analyzed the need for security of mobile apps and proposed a methodology suggesting the use of cloud computing and data mining for improved security. They have adopted Android Security Evaluation Framework (ASEF) representing dynamic analysis and Static Android Analysis Framework (SAAF) representing static analysis, to evaluate the Android apps. They estimate the total time needed to evaluate all the apps stored in one mobile app market. The use of cloud computing along with data mining helps to filter all the malware apps from the safe apps on the mobile phone.

3.1.5 Mobile Internet Big Data Platform in China Unicom
Wenliang Huang, Zhen Chen, Wenyu Dong, Hang Li, Bin Cao, and Junwei Cao have studied on the problems related to the ever increasing mobile Internet data and have provided solutions for the same. They have built a Big data storage and analysis platform based on Hadoop Distributed File System (HDFS). This platform has improved the writing speed to 1390000 per sec and record retrieval time to less than 100 ms.

3.2 Review Analysis
The problem of increasing mobile Internet data cannot be solved. But, the problem of storing this huge amount of data has been reduced by using compression techniques and better storage platform. This mobile Internet data has been effectively used for identification of user behavior and typical pattern of usage by users. It has been used for creating better network solutions and securing the apps from potential malwares as well.

The problem of storing can also be solved by making use of machine learning techniques which will pass a reduced subset of data for storage in the data warehouse. The paper will discuss how these techniques can be effectively applied for better storage and analysis.

IV. PROPOSED SOLUTION
4.1 Proposed Concept: Neural Networks
Neural network is a computation system made up of a number of elements which are simple and provide highly interconnected processing. They are used to process information by their nature of dynamically identifying response to external inputs.

Neural networks are organized into a series of layers. These layers consist of a number of interconnected 'nodes' which calculate a function known as 'activation function'. Patterns are presented to the neural network by means of the 'input layer', which communicates to one or more 'hidden layers'. The actual processing is done at the hidden layer via a system of weighted 'connections'. Neural Networks contain a form of 'learning rule' which recalculates the weights of the connections according to the patterns of inputs entered. Hence, Neural Networks learn by example like a child learns to recognize fruits from examples of fruits.

Neural network finds use in recognizing patterns and detecting anomalous behavior.
4.2 Neural Network Architecture

Fig 1. Neural Network Architecture for Real-time Analysis of mobile Internet Data

The above diagram shows the architecture of Neural Network. It mainly consists of nodes (circular structures) and information flows (lines). It has three layers, which are: (i) input layer which is the passive layer i.e. it does not compute anything. This layer only accepts the input from simple single connections. The nodes I₁, I₂, I₃, …, Iₙ take input which might be in the form a single data value, a pixel, an output value from another source, etc. They do not make any modifications to the data, rather just duplicate the inputs to hidden layer. (ii) Hidden layer (H₁…H₄) and output layers (O₁, O₂) are the ones which are ‘active’. Actual processing takes place in them to make the neural network learn different patterns.

In the case of mobile Internet traffic data, different fields like name of the person, the type of data sent/received, the amount of data transfer, security of network connection (insecure/secure), etc. can be sent as inputs to the neural network.

Basic 4 processing elements can be considered for analysis in real time. They are (i) Location of the person. (ii) Timestamp of peak usage. (iii) Bandwidth. (iv) Usage at the release of sentiment news (storing sentiment score).

The sentiment score can be stored from a range of 0 to 5.0 representing minor usage of the Internet on the release of important news, 5 representing major usage of Internet on the release of sentiment news. These elements will be helpful for the data warehouse such that only potentially fraudulent data can be stored in the warehouse for further analysis purposes. The rest however can be rejected, by considering it safe and not of further need to be stored for a long time in the warehouse.

In the above case, sentiment score with a changed value, abrupt location changes, peak usage of Internet above the threshold value and usage of maximum amount of bandwidth can be considered as anomalous behavior and the particular records can be sent for further analysis.

This enables limited records to be stored, hence enabling limited storage in warehouse and enhancing effective usage of the data warehouse space.

The connections between the input to hidden and hidden to output neurons calculate and adjust different weights for different connections during the ‘Training phase’ of the neural networks.
After the training is done, outputs produced are stable after training for a number of days to a few months. This trained neural network is now used for predictive analysis and pattern recognition against the actual inputs in real time.

V. CONCLUSIONS
The problem of ever increasing mobile Internet traffic cannot be brought under control but the problem of storage of this traffic can certainly be solved. The storage and analysis based on the very useful technique of real-time analysis solves the problem. Making use of machine learning as a tool further enhances the chances of storing only important information and making current as well as future prediction for stock market fraud, illegal trading, fraudulent messages and potential terrorists.

VI. FUTURE SCOPE
The scope for real time analysis of data, whether it is data from telecom domain or any other domain is huge. Further the existence of machine learning techniques and artificial intelligence helps to analyze the data in an effective manner. The mobile Internet traffic will be always increasing and never come at a standstill. The limitations of data warehouse can be reduced but not completely eliminated. But further technological developments can help is restricting the amount of important data that enters into the warehouse.

REFERENCES
2. Jie Yang, Yuanyuan Qiao, Xinyu Zhang, Haiyang He, Fang Liu, and Gang Chen “Characterizing User Behavior in Mobile Internet”, Beijing Key Laboratory of Network System Architecture and Convergence, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China Microsoft, Redmond, WA 98052 USA, 2014 IEEE, pp-2168-6750.