ANALYSIS OF "AIR-MOVING ON SCHEDULE" BIG DATA BASED ON CRISP-DM METHODOLOGY

Man-Seok Ha, Jung-II Namgung and Soo-Hyun Park
School of Business IT, Kookmin University, Seoul, South Korea

ABSTRACT

Punctuality of air traffic is one of the most important criteria for choosing an air service. In this paper, we would like to develop and implement an experimental model based on the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology applied to Big Data Mining. In this case, we choose the data from the ASA (The American Statistical Association) air traffic data for the experiment and then analyzed the data by using the Hadoop Distributed File System, Hive and R studio. The using the analysis, the arrival delay can be proposed for optimal airports. In fact there was a way to take advantage of the leverage results, so we got the best results when applying ANN (Artificial Neural Network) model.

Keywords: big data, hadoop, CRISP-DM, data mining.

1. INTRODUCTION

The importance of the data is growing synchronously with the development of the IT technology. Recently, in many areas, the utilization of the Big Data has been increasing steadily.

In this paper, we propose Big Data analysis model based on the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology.

This paper is organized as follows. Firstly, we describe the research background, target of the research scope methods. Then we have a look on the theoretical background related to Big Data analytics. Next we propose a Big Data for model analysis. And we compare this study with the existing system. Finally, we describe the future research.

2. THEORETICAL BACKGROUND

2.1. Big data

Big Data is a large amount of data and it is difficult to handle using conventional methods so it takes great deal of processing time unlike the conventional data. Big Data is a concept containing data and its informal structure. That means a huge amount of data exceeds the limit that can be processed and analyzed by a general database management system [1].

2.2. Hadoop

Hadoop is made based on the idea of a Google announced Google File System and Map-Reduce. Hadoop was developed in 2005 by Doug Cutting. Hadoop stores the data in the distributed file system HDFS (Hadoop Distributed File System), and processes the data using a Map-Reduce. Before we continue to upgrade the system in order to analyze the mass of the data. But Hadoop is handled by utilizing a distributed processing system and a mass of data.

Therefore, there is an advantage that can rapidly process large quantities of data over a cluster consisting of a system of the generic specification [2].

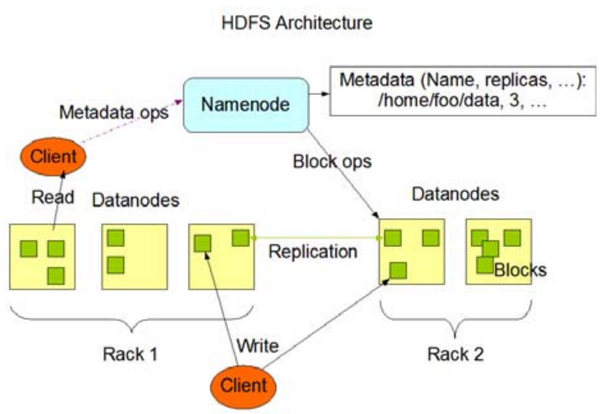

Figure-1. HDFS architecture [3].

2.3. CRISP-DM methodology

Data Mining discovers statistically meaningful rules and techniques that automatically find patterns in large amounts of data [4]. There are several ways of performing Data Mining.

In this paper, we follow the most commonly used CRISP-DM methodology. CRISP-DM methodology is a standard methodology for performing standard Data Mining process, which does not depend on a particular industry or a particular tool. There are various techniques of Data Mining, such as decision trees, clustering, regression analysis, and neural network analysis [5].

VOL. 10, NO. 5, MARCH 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

Figure-2. Procedure of CRISP-DM methodology [6].

2.4. Air-moving on schedule

Air-moving on schedule is a very important element of the evaluation criteria for selecting air services. Most airports and airlines are utilized to calculate the punctuality statistics data based on the flight schedule by 15 minutes [7, 8, 9].

Mayer (2002) analyzed the statistical techniques utilized in various aspects of the data related to the flight time [10].
Coy (2005) suggested an analysis related to the timeliness and punctuality about before and after September 11, 2001 [11].
Cho (2011) visualized the data mart to analyze a variety of factors affecting the flight time [7].
Baik (2013) utilized to analyze the routes using the Flight Department/Arrival data [12].
So far, there have been many attempts to present a pattern or reliable information by analyzing data from the Flight punctuality.

Alas, there is not much research and exploration to examine and apply Data Mining methodology that attempts to analyze the timeliness of data under the Big Data environments.

3. MODEL DEFINITION AND DATA EXTRACTION

In this paper, based on analysis of the literature discussed above, we proposed a model that adopts Data Mining under the Big Data analytics environment to predict the timeliness of the airport.

Moreover, we use the public Airline on-time performance data provided by the ASA (The American Statistical Association), as we want to recommend reliable airports to the customer.

3.1. Business understanding

Customers want convenient, safe and guaranteed flights to arrive on time. We will recommend best airlines and airports to its customers.

3.2. Data understanding

The data used in this paper belongs to from American airlines, so airline flight information data is collected for 22 years from 1987 to 2008.The size of the data is 11GB. And the total numbers of records are 123, 534, 969 [13].

3.3. Data preparation

We have downloaded the data as a csv file from the ASA website.

Then we have uploaded the entire data on the HDFS and analyzed the basic statistics by Hive [14].

And the data was stored in MySQL and SQLite for a performance comparison of the HDFS.

We analyzed each day of the week and the monthly average arrival delay time using the Hive.

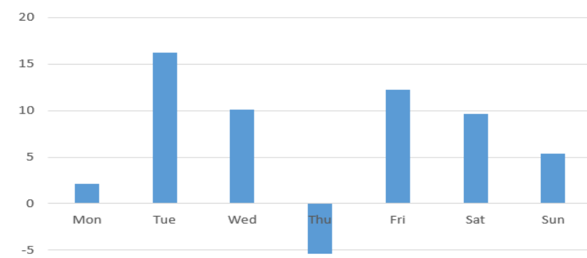

Figure-3. Weekday average arrival delay.

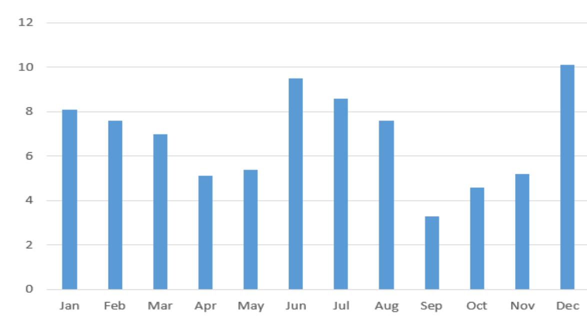

Figure-4. Monthly average arrival delay.

4. MODELLING

4.1. Data sampling

Examining the huge amount of data of 22 years is time-consuming, that's why the samples were only intended for quick analysis of the 2008 data.

A total of five data sets of data from 500 pairs in 2008 was 7, 009, 728 more uniformly random sampling. We used a self-developed Java program to sample randomly. The ratio of the dataset for training and testing datasets are 7:3.

4.2. Data analysis algorithms

We selected three independent variables through correlation analysis (Departure delay, delay arrival time, and flight distance). Then we classified the degree of airports based on the arrival delay time as A, B, and C. Rating of the airports was classified as in Table-1.

Table-1. Rating of airports.

Grade	Airports number	The average arrival delay
A	144	0 ~ 7 min
B	130	7 min ~ 15 min
C	16	15min ~

We used LOGIT (Logistic regression), decision trees, and ANN (Artificial Neural Network) methodologies.

We found out that selecting the best performance from them is a good way to develop a model. Early experiment is the dataset for training. And also experiments to test the dataset are used in order to verify the model. As a result, we can classify good or bad group airports. We used R Studio for data analysis.

4.3. Model evaluation

The results are shown in Table-2. The best results can be predicted using ANN. You can refer to the analysis of the code in Table-3.

Decision tree analysis suggests the performance of the lowest. The reason for this seems to be due to a number data type. Decision tree analysis is more suitable for the data type of category.

Table-2. Airline ratings forecasting performance of the model comparison.

Data set	LOGIT		Decision trees		ANN	
	Train	Test	Train	Test	Train	Test
#1	54.47%	50.98%	52.49%	46.54%	57.14%	47.33%
#2	50.44%	49.04%	49.57%	50.32%	54.29%	54.00%
#3	56.01%	53.73%	51.70%	45.50%	59.71%	51.33%
#4	50.43%	52.32%	45.96%	58.16%	52.86%	51.33%
#5	57.55%	57.05%	47.31%	50.34%	63.14%	51.33%
Average	53.78%	52.62%	49.40%	50.17%	57.42%	51.06%

Table-3. Code for airline ratings forecasting performance of the model comparison.

```
mysample=read.csv("sample.csv", col.names=c('Level',
'DepDelay','ArrDelay','Distance'))
mysample.scale <- cbind(mysample[1], scale(mysample[-1]))
data.size <- nrow(mysample.scale)
samp <- c(sample(1:data.size, data.size * 0.7))
data.tr <- mysample.scale[samp, ]
data.test <- mysample.scale[-samp, ]
model.nnet <- nnet(Level ~ ., data = data.tr, size = 4, decay = 5e-04, maxit = 200)
# Dataset for training
predicted <- predict(model.nnet, data.tr, type = "class")
actual <- data.tr$Level
model.confusion.matrix <- table(actual, predicted)
confusion.matrix.rate = prop.table(model.confusion.matrix) * 100
round(confusion.matrix.rate, digit = 2)
diag.index <- cbind(1:3, 1:3)
error.overall = sum(confusion.matrix.rate) - sum(confusion.matrix.rate[diag.index])
paste("Error Rate =", round(error.overall, digit = 2), "%")
# Dataset for testing
predicted <- predict(model.nnet, data.test, type = "class")
actual <- data.test$Level
model.confusion.matrix <- table(actual, predicted)
confusion.matrix.rate = prop.table(model.confusion.matrix) * 100
round(confusion.matrix.rate, digit = 2)
diag.index <- cbind(1:3, 1:3)
error.overall = sum(confusion.matrix.rate) - sum(confusion.matrix.rate[diag.index])
paste("Error Rate =", round(error.overall, digit = 2), "%")
```

5. CONCLUSIONS

In this paper, we suggested developing and implementing an experimental model based on CRISP-DM methodology applied to Big Data Mining. We choose the air-time flight data with experimental data set and analyzed it by using the Hadoop distributed file system. We want to offer the best carriers to a specific date.

We use three methods of LOGIT (Logistic regression), decision trees, and ANN (Artificial Neural Network) analysis.

We found that ANN method is the best among the three methods. The significance of this paper is to research being applied to Big Data analysis methodology for the evaluation of the air traffic punctuality.

Our future research will be as follows.

The data used in the experiments in this paper is a small amount of data as the Big Data. Further increasing the amount of data may be the data node expansion in order to solve this problem.

We need more data in the various points of view in order to predict the flight punctuality.

So we need to develop a model that can be an objective assessment of the timeliness of operations.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A2012461).

REFERENCES

[1] James Manyika. 2011. Big Data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute Report, May.

[2] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler. 2010. The Hadoop Distributed File System, Mass Storage Systems and Technologies (MSST). 2010 IEEE 26th Symposium on, May. pp. 1-10.

[3] http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[4] http://en.wikipedia.org/wiki/Data_mining.

[5] Yihua Zhang, Yuan Wang, Chunfang He and TingTing Yang. 2014. Research on Forecast Model and Application of Customer Loyalty under the Background of Big Data. International Journal of Multimedia and Ubiquitous Engineering. Vol. 9, No. 10, pp. 209-222.

[6] http://crisp-dm.eu/.

[7] Jae-Hee Cho, De-Kui Li. 2011. Airport Punctuality Analysis Using Multi-Dimensional Visual Analysis Method. Journal of Korea IT Service.

[8] Michael J. Mazzeo. 2003. Competition and Service Quality in the U.S. Airline Industry, Review of Industrial Organization. 22, pp. 275-296.

[9] Jeffrey T. Prince, Daniel H. Simon. 2009. Multimarket Contact and Service Qulity: Evidence From On-Time Performance in the U.S. Airline Industry, Academy of Management Journal. Vol.52, No.2, pp. 336-354.

[10] Christopher Mayer, Toidd Sinai. 2002. Network Effects, Congestion externalities, and air traffic delays: or why all delays are not evil, Nber Working paper series.

[11] Steven Coy. 2005. Management of airline arrival performance before and after September 11, 2001 in US domestic markets. Journal of air transport management. 11, pp. 209-230.

[12] Euiyoung Baik, Jaehee Cho. 2003. Analysis of Airline Network using Incheon and Narita Passenger Flight Origin-Destination Data. Journal of Information Technology Applications and Management.

[13] http://stat-computing.org/dataexpo/2009.

[14] http://hive.apache.org.