Evaluation of Processing Time and Dataset Size with Using Data Mining Application in Cloude

Dr.Arvind Kumar Sharma[1]
Associate professor
Department of Computer Science
OPJS University, Churu, Rajastan, India

Sateesh Nagavarapu[2]
Research Scholar
Computer Science and Engineering
OPJS University, Churu, Rajastan, India

Dr.N.Sathish Kumar[3]
Associate professor
Computer Science and Engineering
S.V.S Group of Institutions, Warangal, India

***Abstract:* In this chapter, the experiment results collected from data mining system are evaluated and analyzed. The IR technology used in this project is able to retrieve the wanted information correctly. The formulas that illustrate the relationships have also been discussed and used. The relationship among IR, data mining system and dataset size has been found out. In summary, IR technology is not suitable for encrypting big dataset.**

***Key words:* Cloud computing, C4.5, C5.0, K-Means**

I.INTRODUCTION

1. RELATIONSHIP BETWEEN PROCESSING TIME AND DATASET SIZE

Linear regression was used to investigate the mostly increasing time taken with IR with increasing dataset size. To begin with, the focus is on the relationships among IR processing time, entire data mining system processing time and dataset size. In this evaluation, the processing time of both IR and entire data mining system with different datasets, which range from 1000 to 10000 in increments of 1000, are involved to find out the relationships.

1.1. Relationship between IR processing time and dataset size in C4.5

In this section, simple linear regression is used to identify the relationship between IR processing time and dataset size. First, data including IR processing time and relevant dataset size is stored in R environment, then two lines of code are executed to use simple linear regression analyzing the data and generate report by R:

```
IRSize <- lm(IR~Size, data=sumIRresults)
Summary(PIRSize)
```

Call: lm(formula = IR ~ Size, data = sumIRresults) Residuals:				
Min	1Q	Medium	3Q	Max
-49850	-7239	-1820	8285	81224
Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 2.973e+03 1.238e+02 24.02 <2e-16 ***, Size 8.549e+00 3.003e-01 28.47 <2e-16 *** Signif. codes: 0 ****0.001 ***0.01 **0.05**0.1 **1				
Residual standard error: 14940 on 298 degrees of freedom Multiple R-squared: 0.7312, Adjusted R-squared: 0.7303 , F-statistic: 810.7 on 1 and 298 DF, p-value: < 2.2e-16				

Table 1: Simple Linear Regression Results

F-statistic: 810.7 on 1 and 298 DF, p-value: < 2.2e-16

According to the report, the increase rate of IR is 8.549, and the formula is:

$$IR= 2.973e^3 + Data\ size * 8.549.$$

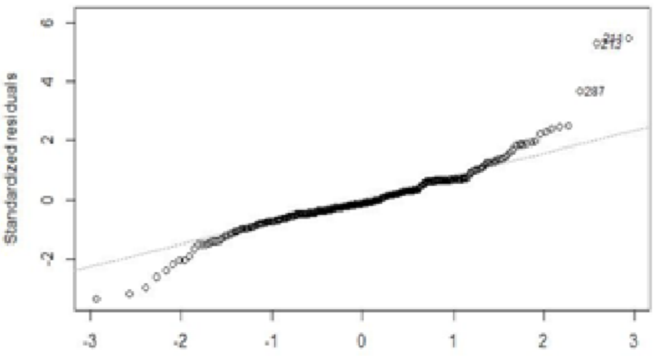

Figure 1: IR - Dataset Size Normal Q-Q Plot

The normal Q-Q plot shows that the points lie on a straight line which means these two variables are correlated. Therefore this model is successfully identified the relationship between IR processing time and dataset size.

In order to compare IR and entire data mining system processing time and find out which one grows faster, the relationship between entire data mining system processing time and dataset size is required. So the next step is to use the same method to analyze the data and identify the relationship.

1.2. Relationship between IR processing time and dataset size in C5.0

Although the data mining system processing time is vary from the machines running the system and algorithms, it is still necessary to identify the system processing time growth rate and compare with IR processing time growth rate in this project.

The data mining processing time experiments are same as IR's, two simple lines of code are run to analyze the data:

TotalSize <- lm(Total~Size, data=sumIRresults), Summary(TotalSize)

lm(formula = IR ~ Size, data = sumIRresults), Residuals:				
Min	1Q	Medium	3Q	Max
-51055	-7904	-1919	7394	85455
Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 1.860e+03 1.630e+02 11.4e <2e-16 *** , Size 8.6407 0.3224 28.47 <2e-16 *** Signif. codes: 0 ****0.001 ***0.01 **0.05**0.1 **1				
Residual standard error: 14940 on 298 degrees of freedom Multiple R-squared: 0.7312, Adjusted R-squared: 0.7303, F-statistic: 810.7 on 1 and 298 DF, p-value: < 2.2e-16				

Table 2: Simple Linear Regression

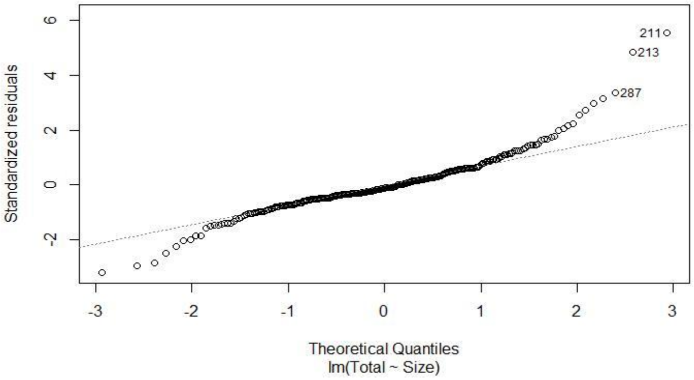

Figure 2: Data Mining Processing Time - Dataset Size Normal Q-Q Plot

The normal Q-Q plot shows that the points lie on a straight line which means that the data mining processing time and dataset size are correlated.

By comparing the two equations, it can be seen that the IR and entire data mining system have similar processing time increase rate while entire data mining system is slightly faster than IR's.However, the Residual standard error in IR and entire data mining system results show that the residual standard error are 14940 on 298 degrees of freedom and 16040 on 298 degrees of freedom respectively, which means that the equations may not accurately predict the growth rate of data mining system. Thus, the relationship between IR and data mining system processing time is needed.

1.3. Relationship between IR processing time and dataset size in K-Means

The growth rate of IR and data mining system have now been found and discussed. However, in order to compare these two aspects in a more detailed manner, a more direct view is needed to be created to recognize the relationship between them.

Therefore, the next step is to identify the relationship between IR and data mining system. Like the last two evaluations, this experiment will focus on linear regression to analyze the data and recognize the relationship.

TotalIR <- lm(Total~IR, data=sumIRresults), Summary(TotalIR)

lm(formula = IR ~ Size, data = sumIRresults), Residuals:				
Min	1Q	Medium	3Q	Max
-41257	-6258	-1687	7994	82585
Coefficients: Estimate Std. Error t valuePr(>|t|) (Intercept) 1.860e+03 1.630e+02 11.4e <2e-16 ***, Size 8.6407 0.3224 28.47 <2e-16 *** Signif. codes: 0 ****0.001 ***0.01 **0.05**0.1 **1				
Residual standard error: 14940 on 298 degrees of freedom Multiple R-squared: 0.7312, Adjusted R-squared: 0.7303 F-statistic: 810.7 on 1 and 298 DF, p-value: < 2.2e-16				

Table 3: Linear Regression Result

The formula of the relationship between total and IR processing time is:

$$\mathrm{Total} = 1.446e^{4} + 1.017 * \mathrm{IR}$$

This report gives the same results. Entire data mining system has higher processing time increase rate than that of IR.

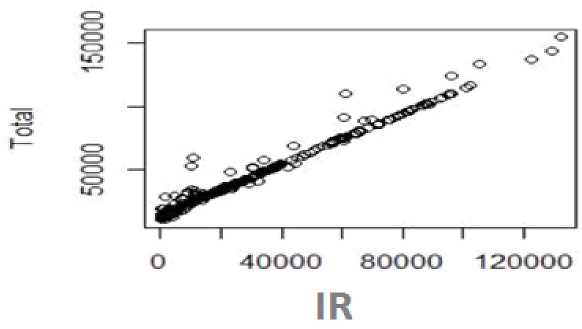

Figure 3: Linear Regression Model of Total and IR Processing Time.

II. CONCLUSION

This experiment results show that the processing time of IR has a higher growth rate than data mining system. First and second report illustrates that the processing time of IR increase slower than the processing time of data mining system by comparing the increase rate with growth of dataset. Since the Residual standard error of first and second experiment is high, we used the third experiment to investigate the relationship between the processing time of IR and data mining system. According to the third experiment results, the formula Total = $1.446e^4$ + 1.017 * IR, the ratio of processing time between IR and data mining system is 1: 0.9619. Therefore, with the increasing of dataset, the IR processing time keeps rising and will eventually occupy most of the entire data mining system processing time.

Based on these findings, the IR is an encryption technology that is only suitable for certain circumstances. The IR should be used on small datasets, since the processing time will rise with the growth of dataset size.

Therefore, it is concluded that IR performance is greatly influenced by dataset size. The IR is efficient when the target dataset is small and simple. The formula shows that given big enough dataset, the processing time of IR will eventually cost over ninety percent of the time spend by data mining system. Due to the reason illustrated above, the PIR is not efficient while dealing large datasets.

III. FUTURE WORK

Due to the limitations of time and hardware, the experiment has several aspects that can be improved as mentioned in the limitations. In future research, these aspects need to be considered to optimize the experiment in order to provide more thorough research and experiment result.

There is no research currently about using other encryption method on the data mining system under cloud environment. Therefore comparisons between IR performance and other encryption methods cannot be made. It would be worth investigating to see whether other encryption methods have better performance such as faster processing time or higher security level than IR. Thus other cryptographic algorithms need to be implemented in the data mining system and evaluated the performance in future.

IV. REFERENCES

[1]. Adapa, S., Srinivas, M. K., & Varma, A. H. V. (2013). A study on cloud computing data mining. *International Journal of Innovative Research in Computer and Communication Engineering, 1*(5), 1232-1237.

[2]. Agrawal, G. L., & Gupta, H. (2013). Optimization of C4. 5 Decision Tree Algorithm for Data Mining Application. *International Journal of Emerging Technology and Advanced Engineering, 3*(3), 341-345.

[3].Ambainis, A. (1997). Upper bound on the communication complexity of private information retrieval *Automata, Languages and Programming* (pp. 401-407): Springer.

[4].Blakeway, S. (2012). Security Practices in Cloud Computing and the Implications to SMEs: ISBN.

[5].Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005). Clustering with Bregman divergences. *The Journal of Machine Learning Research, 6*, 1705-1749.

[6].Barkol, O., Ishai, Y., & Weinreb, E. (2007). On locally decodable codes, self-correctable codes, and t-private PIR *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (pp. 311-325): Springer.

[7].Beimel, A., & Ishai, Y. (2001). Information-theoretic private information retrieval: A unifiedconstruction *Automata, Languages and Programming* (pp. 912-926): Springer.

[8]. Ishai, Y., & Kushilevitz, E. (2005). General constructions for information-theoretic private information retrieval. *Journal of Computer and System Sciences, 71*(2), 213-247.

[9].Beimel, A., Ishai, Y., Kushilevitz, E., & Raymond, J.-F. (2002). *Breaking the O (n 1 (2k-1)/) barrier for information-theoretic Private Information Retrieval.* Paper presented at the Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on.Bojanova, I., Zhang, J., & Voas, J. (2013). Cloud computing. *IT Professional, 15*(2), 12-14. Bond-Graham, D. (2013).

[10].Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. *Hadoop Project Website, 11*, 21.

Dr.Arvind Kumar Sharma', Associate professor, Department of Computer Science , OPJS University, Churu, Rajastan, India.
Mr. Sateesh Nagavarapu, research scholar, Department of Computer Science and Engineering ,OPJS University, Churu, Rajastan, India .
Dr.N.Sathish Kumar, Associate professor, Department of Computer Science and Engineering, S.V.S Group of Institutions, Warangal, India