Analysis of Data mining Applications adopting Information Retrieval in the Cloud

Sateesh Nagavarapu
Computer Science and Engineering
OPJS University, Churu, Rajasthan, India
sateeshnagavarapu@gmail.com

Dr. Arvind K Sharma
Department of Computer Science
OPJS University, Churu, Rajasthan, India

Abstract:- This analysis was allotted to investigate whether or not IR will improve security while not negatively poignant performance. During this analysis, data processing application was enforced underneath cloud atmosphere. A IR protocol was conjointly applied to the info mining application to enhance security. The interval of IR and whole data processing application over multiple datasets with completely different sizes were recorded. The results were analyzed victimization t-test and statistical regression so as to investigate the relationships among dataset size, interval of IR and whole data processing applications.

Keywords-component: Data mining Algorithms, information retrieval protocol.

I. INTRODUCTION

Data mining is associate more and more necessary field of applied science. Its goal is to collect data and extract patterns and information from great amount of knowledge. Data processing is often utilized in a very wide selection of areas like games, business, human rights, medical, science and engineering. However, data processing applications and hardware needed are often barriers certainly sorts of organizations. Not each organization that's inquisitive about data processing will afford these 2 aspects because the price of knowledge storage, maintenance and data processing applications are often on the far side the scope of bound organizations, particularly little organizations.

Cloud computing is a perfect platform for data processing; an outsized proportion of expenditure has been lined by the cloud trafficker once information mining technologies are adopted within the cloud surroundings. Cloud vendors supply data processing applications, infrastructure and information storage. The client will select the categories of services they need and there's no have to be compelled to purchase the functions that they are doing not use. to boot, customers share the infrastructure and storage, more decreasing the expenditure. Existing issues of knowledge mining are security and privacy. Data processing in some cases will raise questions on ethics, lawfulness and privacy. Data processing within the cloud surroundings poses more privacy problems. The info mineworker, World Health Organization has the correct to access the info, additionally has the responsibility to ensure that the info and therefore the results of knowledge mining are each secure and not visible to the cloud service supplier. Whereas Cloud computing will solve security problems to associate extent [5] it additionally brings regarding the inner security issue. Cloud vendors don't offer strategies to ensure that user data can't be seen from server facet. For instance, information analyst or information connected employees has the flexibility to access information then, client or business data might not be entirely secure.

B. Motivation and Research Objective

Data mining in cloud computing looks to be a brand new trend within the data processing space specialize in making certain confidentiality of outsourced information; they illustrate that albeit data is outsourced to a 3rd party, the information worth shouldn't be discernible to the cloud service supplier. IR protocol that is meant to shield user info from server facet may be a appropriate cryptography protocol to use in such a state of affairs. There's additionally some analysis into the employment of IR to secure data processing results [8]. However, to date, there's no analysis that has combined the information mining technologies, cloud computing and cryptography, above all IR, along to analyze performance. IR and cloud computing each contain options which will be exploited to profit data processing technologies. Though cloud computing with its resources is ready to accelerate the calculation processes of knowledge mining applications, and IR protocols are capable of securing the data.

Data mining application needs massive an outsized an oversized} quantity of calculation to analyze large datasets, which suggests the cryptography ways that are wont to secure the information might presumably prolong the interval. Therefore it's necessary to gauge performance once operating in a setting that mixes these 3, i.e. data processing technologies, cloud computing and cryptography, so as to make a decision whether or not IR may be a valid possibility for safeguarding information values from third parties in such AN setting.

II.REVIEW OF LITERATURE

A. Data Mining

Data mining, that is employed in information Discovery in databases, is that the method of analyzing information and generating helpful patterns and relationships. The 3 steps area unit created clear once you think about classification of information. Model learning happens beneath 2 circumstances: once Associate in nursing rule that's utilized in data processing project has learned from the information within the coaching set, or Associate in Nursing rule is applied to information so as to provide a classifier. In model analysis, the classifier that is made in model learning step is checked with a test dataset with illustrious attributes to seek out the accuracy of the model. Once the model reaches expected accuracy, it may be applied to classify new information.

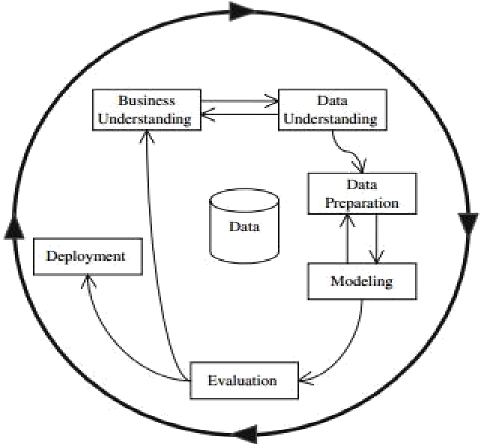

Figure 1: The CRISP-DM Knowledge Discovery Processes Model

The CRISP-DM KDP model contains six steps. The primary step is termed business understanding, during this step; most efforts specialize in understanding the wants and objectives from industrial perspective, and changing this information into data processing downside definition. Once determinative the info mining goals, information understanding starts. Many tasks like information assortment, identification of information quality and outline of information are going to be conducted. Information preparation covers all the mandatory activities to create the ultimate dataset. This step is split into 5 parts: information choice, information cleansing, information construction, information integration and data format. In modeling half, totally different modeling techniques area unit applied. Once models are engineered, analysis is going to be dead to review the model performance and verify following step. Preparation is that the last step of the CRISP-DM data discovery method model. This step is as difficult as applying a repeatable KDP, or as straightforward as writing a report.

B. Cloud Computing

Cloud computing has many characteristics like flexibility and value savings that attract differing kinds of organizations. It guarantees the flexibility to re-provision technological infrastructure resources, price reduction is another advantage; Cloud computing provides huge price savings; [13]explain that the cloud revolution brings an answer to the rising price of IT and also the constant demand for capital investments. It lowers system complexness and also the want for specialists for support and maintenance. [20]. Turner, in his analysis additionally points out that cloud computing might decrease prices since it allows firms to concentrate on the work they are doing and source technologies to vendors' specialists.

Deliver progressive development activities from their analysis on cloud computing and believe that cloud computing is raising new problems in implementation, style and design. They notice there area unit 5 aspects that area unit being centered on by researchers. These 5 aspects area unit, namely, routing knowledge center techniques, virtual-networking within the cloud atmosphere, and challenges of resource allocation in cloud, energy efficient cloud networking and resource allocation for distributed cloud. These area unit the problems and challenges in current cloud networking which require to be investigated. However, the sole security they discuss is that of the distributed denial of service (DDoS) attacks to cloud suppliers. They did not contemplate the interior security problems as a possible threat.

C. Information Retrieval

By using IR query generation algorithm, user can retrieve an element of index i from the target database. The database combines its record with the IR query using a IR reply generation algorithm and produces a result to send back to user. Then the user decodes the results through the reply decoding algorithm.

- Data that is stored in the database does not need pre-processing, storage of additional information or coordination between several different users. Hence, it does not require privacy and has a lower communication complexity.

- Instead of multi-round protocols, the scheme uses a single-round query-answer protocol. This protocol is the common communication pattern in the database environment.

- The scheme is based on the one-way function, which is a function that can be efficiently computed. However, this function cannot be modified in polynomial time.

III.BRIEF SURVEY OF EARLIER WORK

In this part, the latest progress on data mining in the cloud environment, IR and data mining applications and algorithms in the cloud environment will be reviewed. Also, vendors of cloud service and data mining will be discussed to find out the trends in these two areas and the requirements arising from them.

A. Survey of the State of the Art

Use of the Cloud computing paradigm in data processing application and techniques are required by corporations and enterprises [21]. a lot of and a lot of scientific computing and businesses are concerned in cloud computing with data processing changing into a big space to be targeted on. Cloud computing provides services that consider cloud servers to method tasks APIs or Application Programming Interfaces are one more reason that data processing will currently simply access cloud services. Discusses however cloud computing additionally advantages from the employment of genus APIs. Genus APIs is wont to ease the work of programming. Cloud computing permits terminals to- act with cloud computing platform in an exceedingly means that the services and knowledge is deployed across multiple cloud computing vendors. The failure of a cloud service seller won't jeopardize all the opposite copies of client knowledge .Both cloud computing suppliers and data processing applications currently supply genus apis .Developers will use these genus apis to customize their own applications supported what they actually need, instead of purchase services containing elements that they are doing not need.

B.Current Vendors in the Cloud Environment

As mentioned on top of, several firms have began to give differing types of cloud services, and firms like Amazon, Microsoft, Google and Open Stack alter the approach info technologies square measure consumed[11]. Every of their product have options that focus on and attract completely different styles of customers. Cloud suppliers attempt to differentiate their services, targeting their specific customers and specializing in the aspects that they need determined to supply. Since cloud suppliers don't produce equal cloud services and applications, many factors square measure concerned like once evaluating the cloud vendors in order that customers will comprehend those companies' services square measure most fitted for them. These factors square measure mentioned within the following paragraphs.

Performance: Achieving high speed delivery of applications is that the most significant facet of cloud computing performance. Customers expect cloud vendors to deliver high speed services within the cloud. However, to attain this varied challenge, associate degree finish-to end read of the appliance request-response path is needed. Some problems like network performance inside and in-and-out of the cloud, input/output (I/O) access speed between information store tiers and reckon layer, and geographical proximity of the system to the shoppers would have an effect on performance of cloud services.

Technology stack: Technology stack is that the stacks of package services that cloud computing vendors give to customers. Some cloud suppliers emphasize their services on a particular package stack, particularly those who try toremodeltheirservicesfromIaaStoPaaS.

SLAs associate degreed reliability: associate degree SLA is an agreement between 2 or additional parties wherever service is formally outlined. Aspects like responsibilities, quality and scope square measure united between the service user and therefore the service supplier. SLAs square measure an honest indicator of the implications of service failure. Though associate degree SLA might specify a provider's level of commitment,

APIs: API may be a set of protocols, tools and routines for building package applications. Arthropod genus square measure a important issue of cloud supplier choice. Associate degree API that is supported by multiple cloud merchandisers helps cut back vendor lock-in by simplifying migration from one cloud supplier to a different. Also, a well supported API has a whole scheme around it of complementary capabilities and services.

Cost: price may be a easy thanks to compare cloud vendors, the matter is that it's tough to live as a result of there's no consistency among cloud vendors relating to the resources that users truly retrieve and acquire. The virtual machine (VM) that a cloud merchandiser provides varies wide in electronic equipment clock speed, memory capability and alternative options.

Security and compliance: Security, during this case, isn't security threats however inability to achieve compliance with security-related standards like within the payment card trade. Security and compliance is also the most important barrier that forestalls firms and enterprises from adopting cloud computing. Presently there's no protocol or policy to safeguard the confidentiality of knowledge in Cloud server.

C. Data Mining Algorithms

Data mining algorithms are vital to knowledge discovery in databases. As a particular step, the data mining process extracts information patterns from data.

C4.5

C4.5 may be a call tree formula developed from the algorithms CLS and ID3 that were its predecessors. It will tackle categorical and continuous attributes to predict classification. C4.5 handles continuous attributes by cacophonous the info values into 2 components that is predicated on the chosen threshold. It may also manage missing values, and has comparatively sensible performance with each nominal and numerical knowledge. Usually C4.5 is delineated and wont to learn call trees. However, it may also construct classifiers in an exceedingly kind that are a lot of accessible like rule set classifiers

.

C5.0

The C5.0 may be a industrial system developed from C4.5 with benefits over its forerunner. Both C5.0 and C4.5 contain call trees and rulesets, but C4.5's strategies area unit slower and want additional memory. The C5.0 ruleset has lower error rates for forest cowl kind datasets and sleep stage rating datasets. It's extremely optimized: therefore it will use totally different algorithms and performs a lot of quicker than C4.5. Also C5.0 uses less memory than ruleset construction. C4.5 and C5.0 have similar accuracies within the call trees made. However, C5.0 has considerably quicker computation times and smaller tree sizes than C4.5. Different new options of C5.0 embrace ability to handle additional knowledge sorts and a simplified program (Wu et al., 2008).

The k-means algorithm

k-means is the most widely used partitioning method in clustering which was proposed by MacQueen. The k-means algorithm is an iterative method designed to partition a dataset into a certain number of clusters, k. This algorithm has two separate phases—the assignment step and the update step.

- Assignment step. Select the k value from dataset (k is the desired number of clusters). The data objects in dataset which are the most similar are assigned to a cluster, based on distance between cluster mean and data objects.
- Update step. Compute the new mean of each cluster and update the new centroid. Then repeat this process. The algorithm is finished when centroids no longer change.

k-means clustering is easy to apply and implement on large datasets. Additionally, this algorithm suits various topics of data including geostatistics, computer vision, agriculture, market segmentation and astronomy

D. Current IR Progress

Single info IR has begun to emerge because the well-liked coding protocol of selection since analysis has created breakthroughs in computationally personal single info IR at the side of the invention of economical solutions that were mentioned in previous sections. Erasure coded systems that have gained increasing quality, currently conjointly would like IR to secure information [14]. Erasure codes cipher and store knowledge in multiple nodes. Solely a little a part of the first knowledge is needed to be keep in every node, that will increase the provision and responsibility. Meanwhile, erasure codes greatly decrease the whole storage needs. IR on the opposite hand might give privacy primitives in erasure based mostly systems. supported that demand, sovereign et al designed a precise IR algorithmic program and erasure code that resolved the issues like what number property, query-size, and transfer area unit needed by IR.

IV.PLANE OF WORK AND METHODOLOGY

A. Methodology

Experiment research and design thus has several features such as high level of control and level of replication; low level of difficulty to control; low cost of replication; results can be statistically analyzed which means less argument; experiment can be easily replicated; variable can be easily manipulated. Experimental research in the computer science area has been employed for many purposes. For example, experimental research can be applied for system design to find inputs which result in optimal system.

B. Research Questions

Application of IR protocol encrypts the whole database. It can be expected that processing speed of IR will increase with increased size of datasets, as encrypting and decrypting of larger amounts of data is involved. It can be expected that overall data mining system processing time will also increase, not only due to the encryption done by IR protocol, but also due to additional work done by the data mining algorithm. Therefore, the research questions investigated are:

Research question 1: What is the relationship between processing time taken by IR protocol and time taken by overall data mining system?

Research question 2: Can we predict processing time for larger datasets based on the results?

This research study includes designing the data mining system with IR technologies and gathering the time taken to perform data mining process while using IR protocol. In the next section the research questions will be operationalized. The research results will depend on the evaluation measurement and may be different in other environment or settings. The design of the various components of the system is also explained. Following this, the experiment will be conducted. The experimental design will represent the elements, conditions and consequences. The experimental design is divided into six steps:

- Identify and control non experimental factors
  - Select system components, including data mining algorithm, tools, IR protocol and Cloud environment
- Construct and validate system to measure outcomes
- Conduct pilot study
- Determine physical device and time of the experiment.
- Process raw data and collect results which will include retrieval accuracy rate and processing time of both IR protocol and entire data mining system.
- Identify appropriate evaluation method

There is one experimental factor in this investigation. This research sets out to identify whether the IR protocol stays efficient while the dataset size increases. Dataset size is therefore an experimental factor.

Next we go on to recognize the non-experimental factors and find solutions for controlling these. Two types of non-experimental factors are involved in this experiment, the hardware and software, both of which along with the control methods are discussed in the next section. We also consider the other components that are used in the data mining system. Description of each component, comparison with similar products and the reason for choosing these components will then be covered.

V.OBJECTIVE

The aim of 1st analysis is to spot whether or not IR will accurately retrieve the knowledge. The aim of next 2 evaluations is to acknowledge the relationships among the dataset size, interval of IR and whole data processing system. My proposals square measure the primary analysis shows that the IR is in a position to extract data from the datasets that contain one thousand to ten thousand records. in keeping with the second analysis, the interval increase of information mining system and therefore the interval increase of IR square measure similar. The third analysis shows that the statistical regression model anticipated that the interval of IR can eventually represent ninety % of overall interval. Therefore, supported the analysis results, though the IR is in a position to retrieve the knowledge properly, this IR protocol isn't economical for the information mining system and dataset employed in the atmosphere got wind of for this analysis. Thus, the cryptography performed victimization IR is way a lot of sophisticated than data processing victimization K-means. The IR Protocols can be solely appropriate beneath sure circumstances. For instance, IR protocols employed in this thesis failed to perform well since it price an excessive amount of time compared to the time price by data processing algorithmic program. This thesis suggests that cSIR employed in this analysis may not suit the atmosphere like data processing algorithmic program that was chosen within the experiment. To secure the information mining results, IR protocol is needed to be more custom specified the time price is reduced or alternative cryptography technologies square measure required to shield the knowledge.

VI. CONCLUSION

In this research, we presented a potential internal security issue from cloud vendors who provide data mining service. To solve the security issue, an encryption method which is called IR, was proposed. We have implemented IR to secure the data mining results. The IR protocol helps the system to secure the data mining results. However, since original IR protocol requires extracting an entire dataset to prevent the information being seen from the server side, we decided to choose a more efficient IR protocol in this research. Moreover, according to the features of IR, we have selected the corresponding data mining tool, algorithm and cloud framework.

REFERENCES

[1] Jackson, K. (2012). OpenStack Cloud Computing Cookbook: Packt Publishing Ltd.

[2] Joachims, T. (1999). Transductive inference for text classification using support vector machines.Paper presented at the ICML.

[3] Joshi, D. (2011). Polygonal spatial clustering. University of Nebraska.

[4] Kareem, I. A., & Duaimi, M. G. (2014). Improved Accuracy for Decision Tree Algorithm Based on Unsupervised Discretization.

[5] Katsaros, D., Pallis, G., Sivasubramanian, S., & Vakali, A. (2011). Cloud computing [Guest Editorial].Network, IEEE, 25(4), 4-5.

[6] Katz, J., & Trevisan, L. (2000). On the efficiency of local decoding procedures for error-correcting codes. Paper presented at the Proceedings of the thirty-second annual ACM symposium on Theory of computing.

[7] J. P. (2014). Experimental research and design. Retrieved 5/10, 2014, from http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage2.html

[8] Kim, W. (2009). Cloud Computing: Today and Tomorrow. Journal of object technology, 8(1), 65-72.

[9] Kovar, J. F. (2010). Coolest Cloud Storage Vendors. CRN(1293), 32-n/a.

[10] Kushilevitz, E., & Ostrovsky, R. (1997). Replication is not needed: Single database, computationally-private information retrieval. Paper presented at the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science.

[11] Leon, M., & Vadlamudi, P. (1996). Data warehouse vendors do data mining. InfoWorld, 18(24), 39. Li, L., Militzer, M., & Datta, A. (2014). rPIR: Ramp Secret Sharing based Communication Efficient
Private Information Retrieval. IACR Cryptology ePrint Archive, 2014, 44.

[12] Lin, X., Clifton, C., & Zhu, M. (2005). Privacy-preserving clustering with distributed EM mixture modeling. Knowledge and Information Systems, 8(1), 68-81.

[13] Luby, M. G. (1996). Pseudorandomness and cryptographic applications: Princeton University Press.

[14] Luzzi, J. (2014). Experimental Research. Retrieved 6/10, 2014, from http://www.kean.edu/~jluzzi/classes/experim.doc

[15] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Paper presented at the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.

[16] Malek, B. (2005). Efficient private information retrieval: University of Ottawa.

[17] Mayberry, T., Blass, E.-O., & Chan, A. H. (2013). Pirmap: Efficient private information retrieval for mapreduce Financial Cryptography and Data Security (pp. 371-385): Springer.

[18] Melchor, C. A., & Gaborit, P. (2008). A fast private information retrieval protocol. Paper presented at the Information Theory, 2008. ISIT 2008. IEEE International Symposium on.

[19] Mills, E. (2009). Cloud computing security forecast: Clear skies. CNET News

[20] Mittal, P., Olumofin, F. G., Troncoso, C., Borisov, N., & Goldberg, I. (2011). PIR-Tor: Scalable Anonymous Communication Using Private Information Retrieval. Paper presented at the USENIX Security Symposium.

[21]Olumofin, F., & Goldberg, I. (2012). Revisiting the computational practicality of private information retrieval Financial Cryptography and Data Security (pp. 158-172): Springer.

[22]Oracle, V. (2011). VirtualBox user manual. Ostrovsky, R., & Shoup, V. (1997). Private information storage. Paper presented at the Proceedings of the twenty-ninth annual ACM symposium on Theory of computing.