Volume: 12 Issue: 05 Pages: 4725-4731(2021) ISSN: 0975-0290

Short Survey of Data Mining and Web Mining using Cloud Computing

Aye Pwint Phyu
Department of Information Technology Supporting and Maintenance, Computer University, Myanmar
Email: dawayepwintphyu.app@gmail.com
Ei Ei Thu
Faculty of Computer Science, Computer University, Myanmar
Email: eieithueet7@gmail.com

----------**ABSTRACT**----------
In this paper we present the data mining and web mining using Cloud Computing Technology.
Data Mining is used for extracting potentially useful information from raw data. The integration of data mining techniques into normal day-to-day activities has become common place. Web mining includes how to extract the useful information from the web and gain knowledge using data mining techniques. Web mining techniques (specially web usage mining techniques) and applications are much needed in cloud computing. Web mining refers to the application of data mining techniques to the World Wide Web. Data mining techniques and applications are very much needed in the cloud computing paradigm. The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage.

Keywords - **Cloud Computing, Data mining, Web mining, Clustering, Classification, World Wide Web**

Date of Submission: Mar 26, 2021 Date of Acceptance: Apr 26, 2021

1. INTRODUCTION

The World Wide Web is daily used by millions of people. The data are added, edited and read on the web. It is the reason why the World Wide Web can be viewed as biggest database in the world. This dynamically changing database is good subject for data mining research. The data mining is basically discovering unknown patterns in large amount of data. If data mining techniques are used on web data, we are calling it web data mining or web mining. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. (Mining means extracting something useful or valuable from a baser substance, such as mining gold from the earth.)- Basically data mining technique are used in web mining. Web mining is extended version of data mining. Data mining is work upon Off-Line whereas Web mining is work upon On-Line. In data mining data stored in (database) data warehouse and in web mining data stored in server database & web log. The main component of Web Mining Technology has been under development for decades, in research area such as internet, artificial intelligence, and machine learning. The use of cloud computing is gaining popularity due to its mobility, huge availability and low cost. On the other hand, it brings more threats to the security of the company's data and information. At an equally significant extent in recent years, data mining techniques have evolved and became more used them.

2. WHAT IS CLOUD COMPUTING

The term "cloud" is used as a metaphor for the Internet, based on the cloud drawing used in the past to represent the telephone network. The actual term "cloud" borrows from telephony in that telecommunications companies, who until the 1990s offered primarily dedicated point-to-point data circuits, began offering Virtual Private Network (VPN) services with comparable quality of service but at a much lower cost. In early 2008, Eucalyptus became the first open-source, AWS API-compatible platform for deploying private clouds. In early 2008, Open Nebula, enhanced in the RESERVOIR European Commission-funded project, became the first open-source software for deploying private and hybrid clouds, and for the federation of clouds. Cloud computing really is accessing resources and services needed to perform functions with dynamically changing needs. An application or service developer requests access from the cloud rather than a specific endpoint or named resources.

2.1 Aspects of Cloud Computing

Cloud computing represents both the software and the hardware delivered as services over the Internet. Cloud Computing is a new concept that defines the use of computing as a utility, that has recently attracted significant attention.
In Figure 1 below it is illustrated the computing paradigm shift on the last half century through six distinct phases:
• Phase 1: people used terminals to connect to powerful mainframes shared by many users.
• Phase 2: stand-alone personal computers became powerful enough to satisfy user's daily work.
• Phase 3: computer networks allowed multiple computers to connect to each other.
• Phase 4: local networks could connect to other local networks to establish a more global network.

Volume: 12 Issue: 05 Pages: 4725-4731(2021) ISSN: 0975-0290

• Phase 5: the electronic grid facilitated shared computing power and storage resources
• Phase 6: Cloud Computing allows the exploitation of all available resources on the Internet in a scalable and simple way

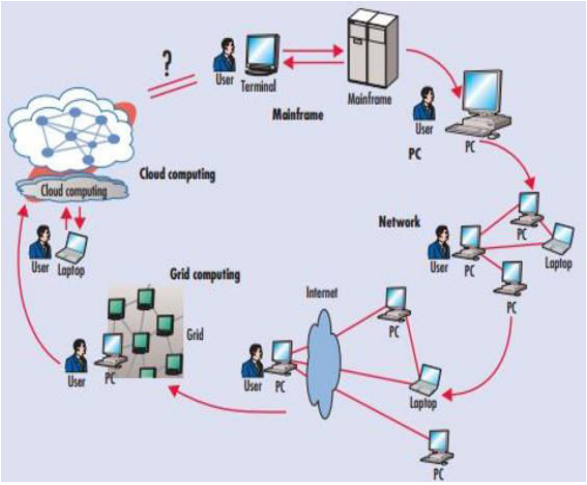

Figure 1: Computing paradigm shift of the last half century

As it is defined by the National Institute of Standards and Technology, Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment mode.

3. OVERVIEW OF DATA MINING

The development of Information Technology has generated large amounts of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

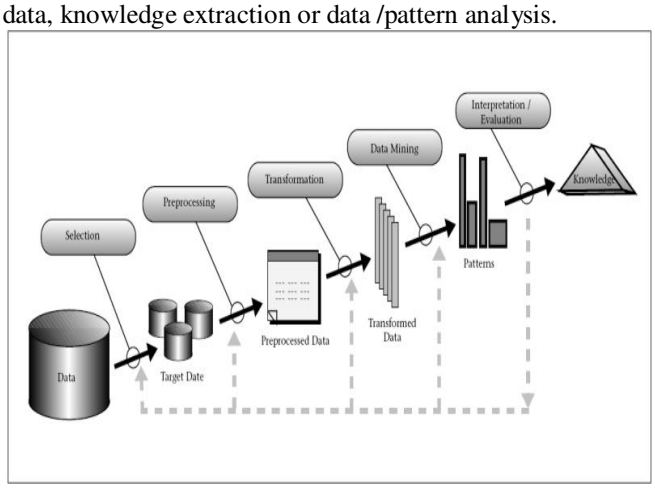

Figure 2: Knowledge discovery Process

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Three steps involved are
- Exploration
- Pattern identification
- Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.

Patterns Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.
Deployment: Patterns are deployed for desired outcome.

3.1 Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

3.1.1 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre classified examples to develop a model that can classify the population of records at large. Fraud detection and credit-risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis.
The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:
- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

Volume: 12 Issue: 05 Pages: 4725-4731(2021) ISSN: 0975-0290

3.1.2 Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques, we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

- Partitioning
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

3.1.3 Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict.

Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response, variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

3.1.4 Association rule

Association and correlation are usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However, the number of possible association rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

3.1.5 Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it.

During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example, handwritten character reorganization, for training a computer to pronounce English text and many real- world business problems and have already been successfully applied in many industries.

Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Types of neural networks

- Back Propagation

4. DATA MINING THROUGH CLOUD COMPUTING

The Microsoft suite of cloud-based services includes a new technical preview of Data Mining in the Cloud “DM Cloud”. DM Cloud allows you to perform some basic data mining tasks leveraging a cloud-based Analysis Services connection.

DM Cloud is valuable capability for IWs that would like to begin considering SQL Server Data Mining without the added burden of needing a technology professional to first install Analysis Services. Additionally, IWs can use the DM Cloud services no matter where they may physically be located as long as they have an Internet connection! The data mining tasks you can perform with DM Cloud are the same Table Analysis Tools found in the traditional Excel Data Mining add-in. These data mining tasks include:

Analyze Key Influencers

- Detect Categories
- Fill -From Example
- Forecast
- Highlight Exceptions
- Scenario Analysis
- Prediction Calculator
- Shopping Basket Analysis

Volume: 12 Issue: 05 Pages: 4725-4731(2021) ISSN: 0975-0290

As Cloud computing refers to software and hardware delivered as services over the Internet, in Cloud computing data mining software is also provided in this way.
The main effects of data mining tools being delivered by the Cloud are:
-the customer only pays for the data mining tools that he needs – that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive;
-the customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing.
Using data mining through Cloud computing reduces the barriers that keep small companies from benefiting of the data mining instruments.
"Cloud Computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources.
The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users."
The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage.

5. OVERVIEW OF WEB MINING

Based on the different emphasis and different ways to obtain information, web mining can be divided into three major parts: web Content mining, Web Structure mining and Web Usage mining. Web contents mining can be described as the automatic search and retrieval of information and resources available from millions of sites and online databases through search engines/ web spiders. Web structure mining operates on the Web's hyperlink structure.

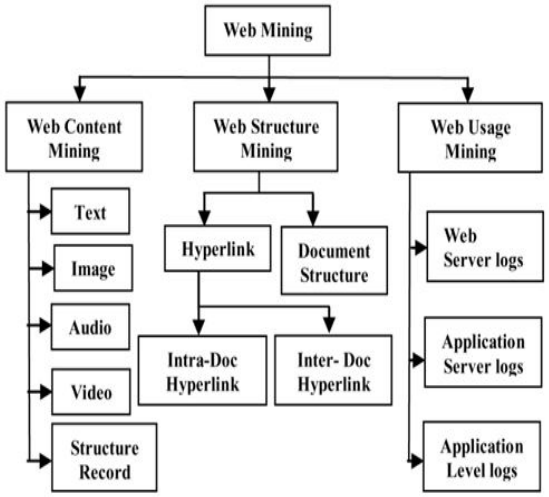

Figure 3: Web mining taxonomy

Web mining has become very vital for effective web site personalization and management. It is crucial for network traffic flow analysis, creating business services, business support, etc. Web mining can be classified into three different types, which are Web content mining, Web structure mining and Web usage mining.

5.1 Web content mining: Web Content Mining is the process of extracting useful information from the contents of Web documents. It is the application of one of the data mining techniques to content published on the internet, usually as semi structured, un-structured, documents. The most widely studied research topics of Web content mining is structured data extraction. Structured data on the Web are often very important as they represent their host pages essential information. The extraction of such data allows us to provide different value-added services like meta-search and shopping. The studies made by researchers in AI and data mining and database reveals the problem that in contrast to unstructured texts, structured data is also easier to extract.

5.2 Web structure mining: Web Structure Mining can be is the process of discovering structure information from the Web. Identifying interesting graph patterns or pre-processing the whole web graph to come up with metrics such as Page Rank. Web's hyperlink structure is operated by the Web structure mining. The illustration of the information about pages ranking or authoritativeness and enhance search results through filtering is provided by this graph structure. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level. The research at the hyperlink level is also called Hyperlink Analysis.

5.3 Web usage mining: User interaction with a web server, including web logs, click streams, and database transactions results at a website or a group of a related sites is analyzed by Web usage mining. Analyzing the web usage log data web mining systems can discover knowledge about users' interest and systems usage characteristics. Personalization and collaboration in decision support, website design, website evaluation, marketing and web-based systems are the various applications of such knowledge. For effective web site management, creating adaptive web sites, business and support services, personalization and network traffic flow analysis web usage mining has become very crucial.

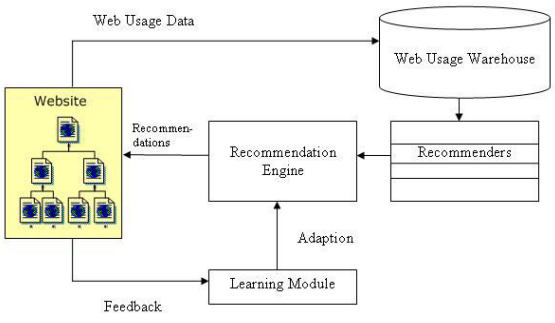

Figure 4: Web Usage Mining

Volume: 12 Issue: 05 Pages: 4725-4731(2021) ISSN: 0975-0290

Web Usage Mining is a part of web mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data, Web Usage mining involves mining the usage characteristics of the users of Web Applications and this extracted information can then be used in a different- ways such as checking of fraudulent elements and improvement of the application.
This paper provides a survey and analysis of current Web usage mining technologies and systems. A Web usage mining system must be able to perform five major functions:

i. Data Gathering,
ii. Data Preparation,
iii. Pattern discovery,
iv. Pattern analysis, and
v. Pattern Applications.

5.3.1 Concept of web usage mining
Data accumulation:
Data accumulation is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality.

Data preprocessing:
Some databases are insufficient, inconsistent. The data pre-treatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion.

Data Cleaning
The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining.
According to the purposes of different mining applications , irrelevant records in web access log will be eliminated during data cleaning. 1. The records of graphics, videos and the format information. The records have filename suffixes of GIF, JPEG, CSS, and so on, which can be found in the URI field of every record. 2.
The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or fewer than 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

5.3.2 User and Session Identification
The task of user and session identification is finding out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions

5.3.3 Path completion
Another critical step in data pre-processing is path completion. There are some reasons that result in path's in completion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time.

5.3.4 Knowledge Discovery Use statistical method to carry on the analysis and mined the pretreated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Pattern analysis Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model store house; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

6. WEB MINING THROUGH CLOUD COMPUTING

Cloud Computing is clearly one of today's most seductive technology areas due at least in part to its cost efficiency and flexibility. However, despite increased activity and interest, there are significant, persistent concerns about cloud computing that are impeding momentum and will eventually compromise the vision of cloud computing as a new IT procurement model. The term _cloud is a symbol for the Internet, an abstraction of the Internet is underlying infrastructure, used to mark the point at which responsibility moves from the user to an external provider. Basically, Cloud Mining is new approach to faced search interface for your data. S a S (Software-as-a-Service) is used for reducing the cost of web mining and try to provide security that become with cloud mining technique. Now a day we are ready to modify the framework of web mining for demand cloud computing. In terms of —mining clouds, the Hadoop and Map Reduce communities who have developed a powerful framework for doing predictive analytics against complex distributed information sources.

6.1 Online Web Usage Mining in Cloud System
Web based recommender systems are very helpful in directing the users to the target pages in particular web sites. Moreover, Web usage mining cloud model systems have been proposed to predict user's intention and their

Volume: 12 Issue: 05 Pages: 4725-4731(2021) ISSN: 0975-0290

navigation behaviors. In the following, we review some of the most significant WUM systems and architecture that Cloud model for navigation pattern mining through Web usage mining to predict user future movements. The approach is based on the graph partitioning clustering algorithm to model user navigation patterns for the navigation patterns mining phase.

7. CONCLUSION

Cloud Computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users. Here we explore the how the data mining tools like SAS, PAS and IaaS are used in cloud computing to extract the information. A cloud provider for a data mining and natural language processing system. Leading cloud computing providers Amazon Web Services, Windows Azure, OpenStack. People use this feature to build information listing, get information about different topics by searching in forums etc. Companies use this service to see what kind of information is floating in the world wide web for their products or services and take actions based on the data presented. The information retrieval practical model through the multi-agent system with data mining in a cloud computing environment has been proposed. It is however, recommended that users should ensure that the request made to the IaaS is within the scope of integrated data warehouse and is clear and simple. Thus, making the work for the multi-agent system easier through application of the data mining algorithms to retrieve meaningful information from the data warehouse cloud computing allows the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage.

We provide a survey about the research in the area of Web mining's today structure and tomorrow view. We point some confusion between data mining and web mining. Web data is growing at a significant rate. Web Mining is fertile area of research. Many Successful applications exist. We also suggest the subtask of web mining & future of web mining. Now we also work for the process mining and try to combine usage mining with structure mining. We also go for the mining from cloud. Whenever we work on mining over cloud computing that -time we hesitate for the cost but that come very less by cloud mining. So, we can say that cloud mining can be seen as future of web mining.

REFERENCES

[1] Jeffrey Voas and Jia Zhang, ―Cloud Computing: New Wine or Just a New Bottle.‖ IEEE Internet Computing Magazine,2009.
http://www.cmlab.csie.ntu.edu.tw/~jimmychad/CN2011/Readings/CloudComputingNewWine.pdf.

[2] Virgilio Almeida, Azer Bestavros , Mark Crovella , and Adriana de Oliveira, ― Characterizing reference locality in the WWW ‖ , In IEEE International Conference in Parallel and Distributed Information Systems, Miami Beach, Florida, USA, December 1996. http://www.cs.bu.edu/groups/oceans/papers/ Home.html

[3] Chen, M. S, Han, J. and Yu, P. S. ―Data Mining: An overview from a database perspective‖, IEEE transaction on knowledge and data engineering, Vol. 08, No. 6, pp: 866-883, 1996.

[4] Raymond Kosala, Hendrik Blockeel,‖ Web Mining Research: A Survey‖, In ACM SIGKDD, July 2000.

[5] Peter Mell, and Timothy Grance, ―The NIST Definition of Cloud Computing‖, The National Institute of Standards and Technology, USA, 2011, Link:http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[6] Merriam-Webster Dictionary, ―Definition of data mining‖, Link:http://www.merriamwebster.com/dictionary/data%20mining

[7] Aviral Kumar Singhal and Niraj Singhal, ― Cloud Computing Vs Fog Computing: A Comparative Study, Vol. 12, Issue. 4, DOI :10.35444/IJANA.2021.12403, pp: 4627-4632, https://www.ijana.in/V12-4.php.

[8] Faten Khalil, Jiuyong Li and Hua Wang ―A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses‖ ,Proc. Fifth Australasian Data Mining Conference (AusDM2006), CRPIT Volume 61,177-184.

[9] Wu, K.L. Yu, P. S. Ballman, A. ―A Web usage mining and analysis tool―, IBM Systems Journal, 2010

[10] Jaideep Srivastava et al , "Discovery and Applications of Usage Patterns from Web Data" , Jan 2000 Volume 1, Issue 2 - page 12.

[11] Kaikala Anjani Sravanthi1, "Web Mining Using Cloud Computing" , ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 4, April 2013.

[12] Vivek Raich, Pradeep Sharma, Shivlal Mewada and Makhan Kumbhkar, "Performance Improvement of Software as a Service and Platform as a Service in Cloud Computing Solution", ISROSET-International Journal of Scientific Research in Computer Science and Engineering, Volume-01, Issue-06, Page No (13-16), Nov -Dec 2013

[13] K. Vijayalakshmi, Dr. M . Vinayakamurthy, Dr. V. Anuradha and Pavan Chunduri, A Survey on Mining Algorithms for Predictive Analytics of Asd, Vol. 10, Issue. 5, (March-April 2019), pp: 38-40, https://www.ijana.in/papers/10-NCACTA_181.pdf.

Volume: 12 Issue: 05 Pages: 4725-4731(2021) ISSN: 0975-0290

[14] Bhagyashree Ambulkar and Vaishali Borkar, "Data Mining in Cloud Computing", MPGI National Multi Conference 2012 (MPGINMC-2012), 7-8 April 2012, Link: http://research.ijcaonline.org/ncrtc/number6/mpginmc1047.pdf

[15] Cloud Computing with the Windows Azure Platform By Roger Jennings

[16] Moving To The Cloud: Developing Apps in the New World of Cloud Computing, By Dinkar Sitaram, Geetha Manjunath

[17] The Cloud Computing Handbook - Everything You Need to Know about Cloud Computing, By Todd Arias

[18] Gajendra Sharma and UmeshHengaju, ―Performance Analysis Of Data Mining Classification Algorithm To Predict Diabetes , Vol. 12, Issue. 1, DOI:10.35444/IJANA.2020.12101, pp: 4509-4518, https://www.ijana.in/v12-1.php.

[19] Deepa B G, Dr. Senthil S and Piyush Singh, ― Data Mining on Classifiers Prophecy of Breast Cancer Tissues, Vol. 10, Issue. 5, (March-April 2019), pp: 8-12, https://www.ijana.in/papers/4-NCACTA_142.pdf.

[20] B.V.Sudhakavya, Swathi.V and Dr.S.Senthil, ―Classification Algorithm in Data Mining, Vol. 10, Issue. 5, (March-April 2019), pp: 18-21, https://www.ijana.in/papers/5-NCACTA_154.pdf.