International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V69I5P106

Volume 69 Issue 5, 45-49, May 2021

Original Article

A Survey on Cloud Data Fetching Techniques and Feature Sets

Amit Kumar Jha[1], Megha Kamble[2]

[1]*Ph.D. Scholar, Department of CSE, LNCT University, Bhopal, India.*
[2]*Professor, Department of CSE, LNCT University, Bhopal, India.*

Received Date: 09 April 2021
Revised Date: 16 May 2021
Accepted date: 20 May 2021

Abstract *- Data storage and fetching algorithm are very useful in a cloud environment to generate a unique pattern from the raw dataset. A number of researchers have proposed different techniques for the processing of raw datasets to extract information. In these algorithms, input is user data output is feature set result in the form of a pattern, cluster, class, etc. This paper introduces some algorithms developed by the researcher to drive information from data. The paper throw light on implementation area of data storage techniques, type of features and there requirement in a different dataset. Evaluation parameter for the analysis of prediction, classification, clustering algorithms as well.*

Keywords *- Cloud computing, Data Fetching, Data Storage, Feature Extraction, Genetic algorithm, Neural Network.*

I. INTRODUCTION

Cloud computing is a newly emerging technology for the future with its roots based on the rapidly increasing demands on data centers that need to be catered to. Cloud computing is defined as the use of computing resources to access data over the internet. It is a means or a mechanism to enhance the existing capabilities of Information technology by many folds [1]. The cloud term comes from the way that the information isn't put away on your work area or your gadget yet is situated far away like a cloud in exacting terms; however, regardless of it being away from its inside your reach, you can get to it independent of your geological area utilizing a registering gadget by means of a web. Distributed computing is an innovation for the future and will change the whole situation of the IT business, being a proficient expense methodology, with the decreased exigency of purchasing the product or the equipment assets. It is an on-request type of utility registering for the individuals who approach the cloud [2]. Recent web search patterns have shown a change in perspective in people groups' interest towards the cloud.

Storage of data in the cloud enables users with a facility to access data as per their requirements irrespective of their time and location. In the cloud, data is stored at data centers in the form of clusters of raw data servers which enables retrieval of this raw data [3]. Cloud users find it very convenient to move their data to the cloud because they no longer have to worry about making any huge investments for hardware infrastructure and their maintenance and deployments [4]. Storage services in the cloud involve the delivery of data storage as a service. The resources are often provided through a utility computing-like basis, i.e., the part of the cloud resources a user is using is the only part he is charged for; it can be for a month, a week, or just a day. In [5], an optimal cloud storage management system is introduced. The biggest advantage of cloud storage is the geographical independence that it offers, i.e., the data in the cloud can be accessed from any location at any time using any computing device.

Since the data in the cloud is stored at some external location away from the users, therefore, this data is prone to several attacks by external sources like some unauthorized person or by some internal sources, which can be due to some untrustworthy service provider, etc. A storage cloud with enhanced metadata scalability. In order to ensure proper hosting of data in the cloud, an efficient data storage auditing protocol for the cloud is introduced by Yang and Jia [6]. Cloud Zone is a cloud data storage architecture based on multi-agent system architecture, and consists of two layers cloud resource layer and MAS layer. Several protocols are present to ensure access to data present in the cloud. A survey of these protocols has been analyzed by Priyadharshini and Parvathi [7]. This can get an overview of the state of the art for various cloud storage approaches. In a cloud setup, the data which is available is huge, and in an unstructured format, this kind of data is difficult to be stored and managed by fixed and structured data models, which are made available by a Relational database Management System. This has led to the development of No SQL system or Key-Value Store systems and will provide the advantage of simplifying rollouts.

The rest of the paper is arranged in a few sections where the second section has summarized different methods adopted by the researcher to extract information as a literature survey. Third gives a brief explanation of different features used in a different type of dataset, while fourth is the collection of data storage techniques. In the fifth section, different evaluation parameters were explained with their formulas.

II. LITERATURE SURVEY

A. Data Storage

V. Anitha et al. [8] In this system, at the first stage self-organizing map trains the features that were extracted from the different wavelet transformation and mix wavelets. Then the obtained filter vector is trained by the nearest neighbor, and then finally, the verification process was done in two stages. This proposed two-level classification was found more effective than the traditional training process.

E. Jadon et al. [9] gave classification of multiclass documents for text documents. To solve the text classification problem, Naïve Bayes classifier was used. The experiment was done first with first linear and second in a hierarchical way to get effective results. Finally, there was a conclusion that the hierarchical way is more effective than the linear approach, and it improved the efficiency and accuracy of such a classifier.

W. Zhang et al. in [10] paper proposed a dispersed data management and calculation kNN (D-kNN) algorithm. The D-kNN calculation has the accompanying benefits: First, the idea of k-closest neighbor limits are proposed, and the k-closest neighbor search inside the k-closest neighbors limits can adequately lessen the time intricacy of kNN. Second, in view of the k-neighbor limit, enormous informational collections past the fundamental extra room are put away on appropriated capacity hubs. Third, the calculation performs k-closest neighbor looking effectively by performing disseminated computations at every capacity hub. At last, a progression of examinations was performed to confirm the adequacy of the D-kNN calculation.

B. Data Searching

Alan Díaz-Manríquez et al. in [11] proposed a model to raise the issue of programmed order of logical writings as an advancement issue, which will permit acquiring bunches from an informational collection. The utilization of developmental calculations to tackle classification issues has been a repetitive methodology. Nonetheless, there are a couple of approaches wherein classification issues are tackled, where the information credits to be classified are text-type. Along these lines, it is proposed to utilize the relationship for figuring hardware scientific classification to get the similitude between records, where each archive comprises a bunch of catchphrases.

Jia0hua Qin et al. in [12] authors primarily enhanced the Harris algorithm utilized to pull out the image traits. Subsequently, the Speeded-Up Robust characteristics algorithm and the Bag of Words model are applied to produce the characteristic vectors of every image. After that, the Local Sensitive Hash algorithm is applied to build the seekable guide for the characteristic vectors. The disordered encryption method is used to guard images and indexes safety. So utilize of chaotic boosts the retrieval time. While the combination of visual Harris and SURF feature reduce the accuracy of image retrieval.

J Li. et al. in [13] proposed a productive and protection safeguarding Multi-catchphrase Ranked Search conspire with Fine-grained admittance control (MRSF). MRSF can understand exceptionally precise ciphertext recovery by joining coordinate coordinating with Term Frequency-Inverse Document Frequency (TF-IDF) and improving the protected kNN strategy. Plus, it can viably refine clients' pursuit advantages by using the polynomial-based admittance system. Formal security examination shows that MRSF is secure as far as the secrecy of rethought information and the protection of files and tokens.

III. FEATURES OF DATA SEARCHING

As per the type of data, text, image, number features were extract, so based on three types of data, features are fetched for learning.

A. Image Features Set

In the transform domain, the host image is segmented into multiple frequency chains using several transformations such as DWT(Discrete Wavelet Transform) or DCT (Discrete Cosine transform) and many more [12]. Then, the inverse transform is applied to obtain the watermarked image. DWT and DCT are the most widely used transform for image watermarking. The frequency transformation in the DCT domain divides the image into different frequency bands, so they facilitate embedding the watermarking information in a specific frequency band [13]. It has been found that the middle-frequency bands are most suitable for embedding the watermark because the low-frequency band carries the most visual essential parts of the image. At the same time, the high-frequency band is exposed to removal through compression and noise attacks on the image. Therefore, embedding the watermark in the middle frequency band neither affects the visible essential parts of the image (low frequency) nor overexposes them to removal through attacks when high-frequency components are targeted [14]. The DWT transform divides into four different parts, namely: LL, LH, HL, and HH sub-bands. Majorly, the LL sub-band is utilized for watermark because the LL sub-band contains Low-frequency components that attain resistance against different attacks**.**

B. Text Feature

Term Frequency: The TF is the count of category-of-words of every category in each document. So the document's term frequency for a category is the occurrence of the words in a single document or article [15].

Document Term Frequency: Gives the number of documents that contain any particular term.

IDF**:** Inverse Document Frequency shows the ability to provide information of words in a document by categorizing it as common or rare. It is the value of a logarithmically inverse fraction of the total documents that contain any word.

$$IDF(t) = log\left(\frac{N}{n}\right)$$

In which n= total number of documents that contain in dataset and n is the number of times that term t appears in the document.

TF-IDF: TF-IDF [16] (Term Frequency-inverse Document Frequency), weight the terms based on inverse document frequency. It simply means the more the term is common in all the documents, the less that term is important and so will be weighted less.

$$TFIDF(t) = TF_t * log\left(\frac{N}{n_t}\right)$$

TF-IDF-CF: TF-IDF had some shortcomings, and so this new parameter was introduced to determine class characteristics, and this class was called frequency by authors, and it calculates the term frequency in documents belonging to a particular class.

$$TFIDFCF(t) = \log(TF_t + 1) * log\left(\frac{N+1}{n_t}\right) * \frac{n_{c,t}}{N_c}$$

The number of documents where term t appears within the same class c document. Nc represents the number of documents within the same class c document.

C. Numeric Feature Extraction

Markov Model: The Kth order Markov models were developed from a series of numbers. These are patterns obtain from the numeric dataset like the weblog page visiting sequence [17].

Regression: As per requirement, different types of regression (linear/logistic) features were extracted from the numeric data [18]. Finding a feature from temporal data is done by this regression.

IV. RELEVANT DATA CLASS STORAGE TECHNIQUES

A. Decision Tree Algorithms

It is a tree-like structure that gives many hopeful solutions to a problem that depends on constraints. The beginning of the tree is from the root and then spreads into several branches and reaches the under the prediction of decision is made. It tends to provide the potential solution to a problem faster and with accuracy than others. Examples of the decision tree are, Classification and Regression Tree or CART, Conditional Decision Trees Decision Stump, Iterative Dichotomiser or ID3, C4.5 and C5.0, Chi-squared Automatic Interaction Detection or CHAID, M5, etc.

B. Support Vector Machine (svm)

It is quite a famous technique that has a group of itself. To demarcate the decision boundaries in the data set with different labels, it uses a hyper dividing plane or a decision type plane. In other words, this algorithm makes a perfect hyperplane by using input data or data of training into categories. Examples are SVM which can perform both linearity and non-linearity classification results.

C. Artificial Neural Network (ann) Algorithms

It is a model which is the exact replica of the neural networks of animals or humans. ANN is considered as a non-linear model that gives the complex relationship between input and output data[8]. But it reduces both cost and time[21] as it compares only the data and not the complete data set. Examples are Hop-related Network, CNN, Recurrent machine learning, Perceptron, Back- Propagation, associative type memory networks, ART, counter propagation networks.

D. Clustering

Clustering was used to reduce the size of the data to manage the large dataset. Here cluster centers were identified, and each cluster tends to select data units from other clusters. It is a tough task to select good bunch centers in large units of data. Many researchers used algorithms such as K-means, Clara, divisive, k-medoid, FCM, etc. [22] related to clustering of data. Out of which some were considered unsupervised while some were partially supervised, in which steps were taken to improve the accuracy of cluster selection. [35]. Once a cluster identifies other elements present in the group similarity value is obtained.

E. Genetic algorithm

It needs plenty of time because the combination increases exponentially with the increase of the sets of data, and a solution was needed for this. So, random choosing of the solution was done to reduce the execution time by using a genetic type of algorithms [23]. These algorithms worked on the concept of environmental and biological activity of the surroundings. Research implemented this concept to solve many problems like clustering, load balancing, shortest path identification, feature selection, classification, etc. [24]. Butterfly, Bee colony, Ant Colony, PSO, etc., are some of the well-known genetic algorithms. It depends on the nature of the problem that which genetic algorithm needs to apply.

F. K-NN

Also called as K nearest neighbors [34, 35] and is a supervised learning machine used to solve regression and classification problems. Based on resemblance like Euclidian distance, it gives new data points. So this algorithm is used to classify data points like Euclidian distance; it differentiates the data points that are similar.

G. K-MEANS the CLUSTERING

IT divides the data into set-off, disjoint groups. Each item can be a member of the group if it's similar. K-mean [34,35] is commonly used in the partitioned clustering algorithm. It clusters the n number of data points into k groups. It gives k centroids that are randomly selected which is single for each cluster. After this, it sends each point of data in the nearest centroid.

H. Random forest Tree

It constructs a collection of random independent and non-independent identical tresses of a decision based on the randomization method. Each decision tree randomly selects vector parameter, feature samples, and a subset of sample data as its training set [28]. The developing algorithm of the random forest is like k gives the number of trees of a decision based on any random forest, n gives the number of samples in the training data set corresponding to each decision tree, and thus segmentation is carried out on an isolated node in the decision tree perfectly.

V. EVALUATION PARAMETER

To test outcomes of the work following are the evaluation parameter such as Precision, Recall, and F-score.

Precision = TP / (TP+ FP)

Recall = TP / (TP + TN)

F-measure = 2 * Precision * Recall / (Precision + Recall)

Where

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

NDCG (Normalized Discounted Cumulative Gain)

$$NDCG@P = Z_P \sum_{i=1}^{P} \frac{2^{l(i)} - 1}{\log(i+1)}$$

Where P is the considered depth, l(i) is the significance level of the i-th image, and ZP is a normalization constant that is selected to let the optimal ranking's NDCG score be 1.

A. Accuracy

Here image fetches from the dataset are evaluate that how many of them are relevant as compared to the total fetch images. Accuracy can be obtained by the below formula:

$$Accuracy = \frac{Number\_of\_\text{Re}levant\_\text{Im}ages}{Total\_Number\_of\_\text{Re}treive\_\text{Im}ages}$$

B. Execution Time

This parameter evaluates the execution time of the algorithm that is the time taken by the method for fetching the images from the dataset as per user query request. It is expected time required for image retrieval should be less.

VI. CONCLUSION

As a large amount of data available in different platforms, so information extraction depends on machine algorithms. This paper has summarized techniques of raw data feature extraction proposed by the researcher in various fields Web Mining, Text Mining, Image Processing, etc. It was found that searching data needs a structured dataset and this structure directly depends on extracted features. So features of data as per the type of dataset were also detailed in this paper. Evaluation parameters were also shown in the paper for a comparison of the machine learning algorithm. In future, one can develop a generalize technique which works in all set of datasets.

REFERENCES

[1] S. Subashini, V. Kavitha, A survey on security issues in service delivery models of cloud computing, Journal of Network and Computer Applications, 34(1) (2011) 1-11.

[2] F. Hu, M. Qiu, J. Li, T. Grant, D. Tylor, S. McCaleb, L. Butler, and R. Hamner, A review on cloud computing: Design challenges in architecture and security, Journal of Computing and Information Technology, (2011) 25-55.

[3] Z. Zeng, and B. Veeravalli, Do More Replicas of Object Data Improve the Performance of Cloud Data Centers, IEEE Fifth Int'l Conf. Utility and Cloud Computing (UCC), (2012) 39-46.

[4] Q. Liu, G. Wang, and J. Wu, Secure and privacy-preserving keyword searching for cloud storage services, Journal Network Computer Applications, 2011.

[5] J. Spillner, J. Müller, A. Schill, Creating optimal cloud storage systems, Future Generation Computer Systems, 29(4) (2013) 1062-1072.

[6] K. Yang and X. Jia, Data storage auditing service in cloud computing: challenges, methods, and opportunities, World Wide Web, (2012) 409-428.

[7] B. Priyadharshini and P. Parvathi, Data integrity in cloud storage, Int'l Conf. Advances in Engineering, Science and Management (ICAESM), (2012) 261-265.

[8] V. Anitha, S. Murugavalli. Brain tumor classification using two-tier classifier with adaptive segmentation technique. ISSN 1751-9632 24th July 2014.

[9] E Jadon, R Sharma et al. Data Mining: Document Classification using Naive Bayes Classifier, International Journal of Computer Applications 167 (6) (2017).

[10] W. Zhang, X. Chen, Y. Liu, and Q. Xi, A Distributed Storage and Computation k-Nearest Neighbor Algorithm Based Cloud-Edge Computing for Cyber-Physical-Social Systems, in IEEE Access, 8 (2020) 50118-50130.

[11] Alan Díaz-Manríquez, Ana Bertha Ríos-Alvarado, José Hugo Barrón-Zambrano, Tania Yukary Guerrero-Melendez, And Juan Carlos Elizondo-Leal. An Automatic Document Classifier System Based on Genetic Algorithm and Taxonomy. (2018).

[12] Jiaohua Qin, Ha0 Li, Xuyu Xiang, Yun Tan, Wenyan Pan, Wenta0 Ma1, And Neal N. Xi0ng. An Encrypted Image Retrieval Meth0d Based 0n Harris C0rner 0ptimizati0n And Lsh In Cl0ud C0mputing. Ieee Access (2019).

[13] J. Li, J. Ma, Y. Miao, Y. Ruikang, X. Liu and K. -K. R. Choo, Practical Multi-keyword Ranked Search with Access Control over Encrypted Cloud Data, in IEEE Transactions on Cloud Computing, 2020.

[14] Hongyu Yang And Fengyan Wang. Wireless Network Intrusion Detection Based on Improved Convolutional Neural Network. IEEE Access 7, (2019).

[15] Hong Huang, Fanzhi Meng, Shaohua Zhou, Feng Jiang, And Gunasekaran Manogaran. Brain Image Segmentation Based on FCM Clustering Algorithm and Rough Set. Ieee Access, New Trends In Brain Signal Processing And Analysis 7, (2019).

[16] Vinod Sharma, Dr. Shiv Sakti Shrivastava. Document Class Identification Using Fire-Fly Genetic Algorithm and Normalized Text Features. International Journal of Scientific Research & Engineering Trends, IJSRET Volume 6 Issue 1 Jan 2020.

[17] Alan Díaz-Manríquez, Ana Bertha Ríos-Alvarado, José Hugo Barrón-Zambrano, Tania Yukary Guerrero-Melendez, And Juan Carlos Elizondo-Leal. An Automatic Document Classifier System Based on Genetic Algorithm and Taxonomy (2018).

[18] Gerard Biau. Analysis of a Random Forests Model. Journal of Machine Learning Research 13 (2012) 1063-1095.

[19] Gourav Rahangdale, Manish Ahirwar, and Mahesh Motwani. Application of k-NN and Naive Bayes Algorithm in Banking and Insurance Domain. International Journal of Computer Science Issues (IJCSI) 13 (5) (2016).

[20] Niti Arora and Mahesh Motwani. A Distance-Based Clustering Algorithm International Journal of Computer Engineering & Technology (IJCET) 5(5) 2014.