A Survey on Relevant Text Data Searching Techniques and Features in Cloud

Rajesh Kumar Nigam[1], *Alesh Kumar Sharma*[2], *Pradeep Kumar Pandey*[3]

[1]Department of Computer Science & Engineering, School of Computer Science and Technology, SAM Global University, Bhopal, M.P., India.
[2]Department of Computer Science & Engineering, LNCTE, Bhopal, M.P., India.
[3]Department of Computer Science & Engineering, School of Computer Science and Technology, SAM Global University, Bhopal, M.P., India.

E-mail(s): rajeshrewa37@gmail.com, aleshks@lnct.ac.in, pradeep65656@gmail.com

Abstract: Internet access increases the volume of data for storage, analysis, fetching, etc. Out of different type of data text is most bulky and unorganized in nature. Many of researchers have proposed different models for data management and retrieval. This paper is a deep survey of cloud text data fetching and storage. Many of cloud application use encryption model for the stored data security. So a detailed survey of various authors work was summarized in the paper with type of data and techniques adopt. Features used in the text mining were also brief in the paper for the analysis of impact of type of text data application. Paper has brief some of evaluation parameters that needs for comparing of relevant data fetching models.

Keywords: Cloud computing, Data Fetching, Data Storage, Feature Extraction, Machine Learning.

I. INTRODUCTION

In recent years, the nature and volume of data have been impacted by technological advances, posing a serious difficulty for data management and retrieval approaches. Almost every element of our life has been transformed by information communication. Data, once thought to be an impossible dream, has now come true, allowing computers to understand and communicate with humans while processing their thoughts.

Most text databases store semi structured data, which is material that is neither totally unstructured nor completely structured. A document, for example, might have a few structured fields like title, authors, publication date, and category, but also some entirely unstructured text components like abstract and contents [1]. In contemporary database research, there have been numerous studies on the modeling and implementation of semi structured data. In addition, strategies for retrieving information, such as text indexing approaches, have been developed to deal with unstructured documents. For the ever-increasing volumes of text data, traditional information retrieval approaches are becoming inadequate.

The organizing and retrieval of information from large database collections is concerned with information retrieval [2], [3].

It deals with information retrieval, as well as the representation, storage, and organization of knowledge. Information retrieval is concerned with search operations in which a user must identify a subset of information among a big amount of knowledge that is relevant to his information demand. "Finding relevant information or a document that satisfies user information needs" is the main purpose of an information retrieval system (IRS). [4] IRSs often use processes like indexing, filtering, and searching to achieve this purpose. The three processes listed above are the fundamentals of information retrieval. In indexing, documents are described in a summary fashion.

Because cloud storage divides data into fixed-size parts, there is a requirement for a reliable and efficient fault tolerance solution that can ensure that even if a plurality of slices is lost, relying on the remaining slices may also restore file integrity [5]. Figures 1 and 2 depict a data security schematic diagram based on cloud storage. Because the essential technology of data security in cloud storage is a broad topic with numerous facets, it is vital to explain the paper's main content. There has been a lot of research into data security in cloud storage, cloud storage in access control, and so on.

II. Literature Review

Perform document clustering utilizing supplemental information in addition to the material in [7] paper to get clusters with improved purity. This research also identifies the usage of such supplemental data for clustering in applications involving different file formats such as audio, image, video, and so on. If the extra information linked with pure content is noisy, the clustering performance may suffer. Taking this into account, this research employs a partitioning-based clustering technique as well as a probabilistic model.

Cluster-based Retrieval with Pattern Mining is a revolutionary cluster-based information retrieval approach described in [8]. (CRPM). Various clustering and pattern mining methods are combined in this strategy. To begin, it creates clusters of things that are similar in nature. To reduce the number of shared terms across clusters of items, three clustering techniques based on k-means, DBSCAN (Density-based spatial clustering of applications with noise), and Spectral are proposed. Second, each cluster is subjected to frequent and high-utility pattern mining methods in order to extract the pattern bases. Finally, for each query, the clusters of items are ranked. Two ranking strategies are proposed in this context: I Score Pattern Computing (SPC), which computes a score representing the similarity between a user query and a cluster; and ii) Weighted Terms in Clusters (WTC), which computes a weight for each term and uses the relevant terms to compute the score between a user query and each cluster. Unexpected user inquiries are also dealt with using irrelevant information gathered from the pattern bases.

In [9] authors provides text document categorization utilizing two clustering algorithms, K-means and K-means++, with a comparison to determine which approach is optimal for categorising text documents. Pre-processing is also introduced in this project, which comprises tokenization, stop-word elimination, and stemming. It also entails calculating Tf-Idf. The impact of the three distance/similarity measurements (Cosine Similarity, Jaccard coefficient, and Euclidean distance) on the outcomes of both clustering algorithms (K-means and K-means++) is also assessed. The evaluation dataset comprises of 600 text pieces from three different categories in India: festivals, sports, and tourism. Our findings suggest that employing the K-Means++ clustering algorithm with the Cosine Similarity measure to categorise text articles produces better results.

The security and energy consumption of medical electronic health record (EHR) data transmission and storage between cloud server and IoT device users are the subject of the [10] study. By integrating a safe energy-saving communication scheme and encryption algorithm to the existing medical cloud model, this work creates a secure energy-saving communication and encrypted storage model. This paper proposes the MedGreen communication authentication technique, which is based on an elliptic curve and a bilinear pair. The approach allows the two communication parties to complete key establishment and identity authentication in only one conversation, successfully balancing the key centre and user resource overhead and preventing the Man-in-the-Middle attack. This paper presents MedSecrecy, a secure data storage algorithm based on Huffman compression and RC4 that aims to address the features of large repetition and high sensitivity of medical data.

Describe the system architecture, data distribution technique, and retrieval system that this effort has created in [11]. For effective retrieval and indexing of data for crawling, a convolutional neural network (CNN) is used to classify text

documents. An API-based micro-service architecture is used to disseminate and retrieve information depending on the identifying key. The system provides a platform for extracting knowledge and channelling data for use by the company, as well as allowing support centres to provide on-demand services.

II. Features of Data Searching

Term Frequency: The TF is the count of category-of-words of every category in each document. So the documents term frequency for a category is the occurrence of the words in 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 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 was having 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.

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

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

III. Relevant Data Class Storage Techniques

Decision Tree Algorithms

It is a tree-like structure that gives many hopeful solutions to a problem which 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.

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 dividing hyper plane or a decision type plane. In other words, this algorithm makes an perfect hyperplane by using input data or data of training into categories. Examples are SVM which can perform both like linearity and non-linearity classification results.

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 is 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.

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.

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 solution was done to reduce the execution time by using 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.

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.

K-MEANS the CLUSTERING

IT divides the data into set-off disjoint group. 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.

Random forest Tree

It constructs a collection of random independent and non-independent identical tresss of 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 decision based from 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 a isolate node in the decision tree perfectly.

IV. 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 to be 1.

Accuracy

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

$$Accuracy = \frac{Number\_of\_Relevant\_Images}{Total\_Number\_of\_Retreive\_Images}$$

Execution Time

This parameter evaluates execution time of the algorithm that is 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.

V. Conclusion

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

References

1. Jiaohua Qin, Ha0 Li, Xuyu Xiang, Yun Tan, Wenyan Pan, Wenta0 Ma1, And Neal N. Xi0ng. “An Encrypted Image Retrieval Meth0d Based 0n Harris Corner optimizati0n And Lsh In Cl0ud Computing”. Ieee Access March 7, 2019.

2. Y. Sang, Q. Zhang, K. Zhang, D. Wang, Q. Yuan and H. Liu, "Fuzzy Keywords-Driven Public Sports Resource Allocation Strategies Retrieval With Privacy-Preservation," in IEEE Access, vol. 8, pp. 195980-195988, 2020.

3. M. Oppermann, R. Kincaid and T. Munzner, "VizCommender: Computing Text-Based Similarity in Visualization Repositories for Content-Based Recommendations," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 495-505, Feb. 2021

4. Yannan Li, Yong Yu, Bo Yang, Geyong Min, Huai Wu, Privacy preserving cloud data auditing with efficient key update, Future Generation Computer Systems, Volume 78, Part 2, 2018.

5. Jing Han, Yanping Li, Weifeng Chen. "A Lightweight And privacy-preserving public cloud auditing scheme without bilinear pairings in smart cities". Computer Standards & Interfaces, Volume 62, 2019.

6. 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.
7. Chitra Kalyanasundaram, Snehal Ahire, Gaurav Jain, Swapnil Jain. "Text Clustering for Information Retrieval System Using Supplementary Information". International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015.
8. Djenouri, Y., Belhadi, A., Djenouri, D. et al. Cluster-based information retrieval using pattern mining. Appl Intell 51, 2021.
9. A. A. S. S. F. “Text Categorization of Documents Using K-Means and K-Means++ Clustering Algorithm”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, no. 6, June 2016
10. J. Zhang, H. Liu and L. Ni, "A Secure Energy-Saving Communication and Encrypted Storage Model Based on RC4 for EHR," in *IEEE Access*, vol. 8, pp. 38995-39012, 2020.
11. H. Chiranjeevi and K. S. Manjula, "An Text Document Retrieval System for University Support Service on a High Performance Distributed Information System," 2019 IEEE 4th International

Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2019.

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 March 7, 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 Volume 7, May 30, 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 Volume 7, Feb 6, 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”. accepted March 9, 2018, date of publication March 15, 2018, date of current version May 9, 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) Vol. 13 (No.5) Sept. 2016.
20. Niti Arora and Mahesh Motwani. "A Distance Based Clustering Algorithm" International Journal of Computer Engineering & Technology (IJCET) Vol.5(No.5) May 2014.