IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Plagiarism Checker And Classification Of Files On Cloud Using Smart Cloud

[1]Dr.Balaji K, [2]Divya .M, [3] Krupa.B, [4]Gouri.G, [5] Arpita.P

[1]Professor, Department of MCA, Cambridge Institute of Technology CITech, Bengaluru, India, [2,3,4] Student, Department of MCA, CITech, Bengaluru, India

ABSTRACT

This study suggests a method to address the cloud duplication problem. The restricted amount of data storage that these service providers can offer is only partially utilized by the abundance of redundant and useless data on the internet. Inadequate cloud efficiency might result in far greater costs for service providers than anticipated. The fact that the data is kept on a server that may be located kilometres distant from the user presents another problem. This raises the price and efficiency even further. A smart cloud can assist with resolving these problems. By calculating the ratio between the proportions of two vectors, we can assign a categorization rating to the document.

1. INTRODUCTION

The way that hardware and software are incorporated has altered due to the revolutionary process known as cloud computing. Users of cloud computing benefit from a number of advantages, including easy access to their data, reduced capital costs for upgrades and hardware, enhanced security and compliance, and more. Cloud computing is helping everyone, from tiny businesses to major corporations, transition to the digital age. These days, most smart devices even use the cloud to store the personal data of their users. The reason these gadgets' function is directly related to cloud computing. The cloud servers are also home to a number of applications. Therefore, it is plausible to claim that there is a large amount of similar and identical data on the cloud**.** It makes it challenging for users to locate specific data on the cloud. This problem can be explained by the information grouping, which can make it easier to find files in the cloud. This solution then raises another issue, which is the cloud's efficiency. The volume of data kept in the cloud has suddenly increased recently. By 2020, there will likely be 40 zettabytes of digital data stored, compared to the approximate 3.5 zettabytes that were recorded in 2013. One of the biggest issues is the capacity discrepancy between production and demand for data storage. Producing massive volumes of capacity to match the volume of data stored on it is more difficult.

2. LITERATURE REVIEWS

Numerous algorithms and techniques are employed to enhance the functionality and efficacy of the cloud. There are a few ways to accomplish this, including using two-phase frameworks with 3D security applied. The first stage encrypts all of the data, making it unreadable without the right key. After that, it is kept on the cloud. The availability, integrity, and secrecy of the user are taken into consideration when encrypting data. Following that, a critical rating is assigned based on these characteristics, and protection is then assigned based on these ratings. The user must go through an authentication process in order to access the data in the second step. To enhance the management of the files stored in the cloud, a cloud file management system that supports document clustering is employed. A mix of k-means and top-k frequent item sets would make up the system. The par TF-IDF algorithm uses the top-k frequent item sets, which are a component of SHDC algorithms.

3. METHODOLOGY

3.1. TF-IDF Algorithm

TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a widely used weighting descriptive mechanism for the documents . Term frequency (TF) is the number of times a term appears in a document. It is calculated as: $number\ of\ occurence\ of\ t/e\ key2ord\ 4n\ t/at\ part4cular\ document\ total\ number\ of key2ords\ 4n\ t/e\ document$

(1) Inverse Document Frequency (IDF) measures the uncommonness of a term in the whole archive. It measures the rarity of a term. idf = log $N\ d\ f$

(2) where N is the aggregate number of reports in a collection and df is the document frequency. The concept of term frequency and inverse document frequency are consolidated, to deliver a composite weight for each term in each corpus. tf-idf = tf * idf.

4 RESULT:

This article suggests a solution for the problem of duplicate content clogging cloud storage. As we already know, everyone uses cloud services: individuals as well as major enterprises. The fact that the cloud contains a large number of duplicate files is a major issue. It is possible to submit the same file under a different name to the cloud, and these redundant documents take up a lot of cloud storage space.

For instance, the cloud depicted in the picture contains five files, the majority of which are duplicates of other files in the cloud. Presume that the files listed below have nearly identical contents. All of these files combined would total 1773 kb in size. The amount of space in the cloud would have been 200 KB rather than 1773 KB if duplicate papers could have been prevented. This duplication squanders storage space and important resources. This drives up the cost of cloud storage and upkeep as well.

S.NO	NAME	SIZE (KB)
1.	Name.pdf	200
2.	DuplicateName.pdf	243
3.	Codes.doc	341
4.	CodeDesign.doc	456
5.	DupliName2.pdf	533

Table. 1 Example of list of files on the cloud.

The quantity of files that customers have downloaded throughout the time that the file has been posted to the cloud is shown in Table 2. Days indicates how long a file has been on the cloud, while download indicates how many files clients have submitted and downloaded. The fact that the documents in the list are not categorized makes it difficult for the client to find the necessary document, which is another major issue.

S. NO	File	Download	Days
1.	Name.pdf	55	10
2.	DuplicateName.pdf	65	8
3.	DupliName2.pdf	53	7
4.	Codes.doc	23	16

The approximate distance between the servers and the client's location is displayed in Table 3. When the customer accesses the files, it shows how effective cloud services could be. The cloud assigns a server at random to a client when the client attempts to access a file that has been uploaded to the cloud server. For instance, the network's efficiency and download speed would decline if a client from Delhi was assigned to a server that might be located in Canada and was far from the client.

S.No	Server	Distance from Connaught Place, Delhi (km)
1.	Melbourne, Australia	10,133.6
2.	London, United Kingdom	6374.11
3.	New York, USA	19,605,2
4.	Singapore	5897.22
5.	Toronto, Canada	20361.85
6.	Moscow, Russia	5221.91

Table. 3 Distance of the servers from the client.

5.1 SUGGESTED METHOD:

5.1.1 Content Delicacy Checking

An technique known as the "checker's algorithm" is used to determine whether the content of a document that a user uploads is duplicated across all other documents stored on the cloud.

5.1.1.1 Enumeration

It is the process of eliminating the corpus's white and empty gaps. The characters with whitespace are. For instance, line breaks, punctuation, and spaces.

5.1.1.2. Climbing

Reducing derivative terms to their stem word is the process involved. It assists in lowering the plagiarism checker's overhead. Typically, a basic stemmer consults a lookup table to find the inflected form. This method can handle unique situations with ease and is highly flexible.

5.1.1.3. Stop word removal

The document has several stop words that prevent it from being compared; therefore, in order to make it a valid document for content comparison, the stop words must be removed.

1. Word w in the processing document d where d $ D is the 5.1.2 TF-IDF Algorithm.

2) Log(|d|/fw,D) * fw,d

3) To determine Wd for each word in d, repeat the previous procedures for each word.

4) Words with high Wd indicate that they are significant words in document d. In this example, "D" represents the collection of cloud-based documents, "d" represents the document that is being processed algorithmically, "W" represents the word in "d" that is being processed algorithmically, "Wd" represents the weight of the word in "d," "Fw,d" represents the number of words "w" in document "d," and "Fw,D" represents the number of words "w" in the set of documents D.

6 CONCLUSION

In this research, we suggested a method that uses the TF-IDF algorithm to reduce the number of duplicate documents uploaded to the cloud. If this strategy is carried out correctly, it will make the cloud free of plagiarism and enable the system to be more robust. Faster and more effective results would be possible with the use of a more unique and superior algorithmic technique. If the new document contains more than a predetermined percentage of content from an already-existing cloud document, it will also block the user's document.

REFERENCES

[1] Jayalakshmi, M. B., and S. H. Krishnaveni. "A study of data storage security issues in cloud computing." Indian Journal of Science and Technology 8.24 (2015).

[2] http://www.techradar.com/news/internet/data-centre/world-could-run-out-of-storage-capacity-within-2-years- warns-seagate-vp-1278040/2

[3] https://en.wikipedia.org/wiki/Plagiarism_detection

[4] Ngnie Sighom, Jean Raphael, Pin Zhang, and Lin You. "Security Enhancement for Data Migration in the Cloud." Future Internet 9.3 (
2017): 23.