Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 4, April -2017

Duplicate File Searcher and Remover

Remover Of Duplicate Files

Ekta Thorat[1], Lekha Sonawane[2], Dhanshree Wadile[3], Prof.Manisha Sonawane[4]

[1]*Department of Computer Engineering, Shivajirao S. Jondhale College Of Engineering,Dombivali,India*

[2]*Department of Computer Engineering, Shivajirao S. Jondhale College Of Engineering,Dombivali,India*

[3]*Department of Computer Engineering, Shivajirao S. Jondhale College Of Engineering,Dombivali,India*

[4]*Department of Computer Engineering, Shivajirao S. Jondhale College Of Engineering,Dombivali,India*

Abstract – *In computer hard drive is one of the core component. There's a lot of possibility to have same files on a same or different directory, searching for the same file in each directory is very difficult and take a long time. Duplicate File Searcher and Remover application is able to resolve this problem. It will able to find out the same file that is located in a different directory in hard drive. Duplicate File Searcher and Remover application be able to compute the hash value of each file that can find the duplicate of same file. Duplicate File Searcher application uses MD5 (Message Digest 5) hashing Algorithm to compute hash value of each file. This Application will not only find a duplicate file but also remove the duplicate file which is not of any use to the user. Duplicate File Searcher application is designed using Java programming language.*

Keywords— *MD5(Message Digest), Hash value, SHA, Duplicate file, Data Analysis*

I. INTRODUCTION

While managing and performing file operations on computer or on other storage devices, many duplicate files with a considerable size may be gathered there. Accumulation of these digital junk levels can be a primary cause for shortage of storage space and decrease in computer performance. Therefore, you need to search and erase duplicate files from computer hard drive. But, would it be an easy task for anybody to make search for all duplicate files? What if you do not have any information about such files those have replica also. Will you search duplicate of each file one by one? It seems troublesome and definitely, it is a time taking process. If talk especially about computer users they should know how the access of duplicate files can affect their job. We all are aware of importance of RAM (Random Access Memory) in a computer. When you perform any operation on a file, it comes on RAM memory from where OS reads data from that file**.** If duplicates of a requested file are present on your computer, all will take place in RAM hence it may cause your system performance slowdown. Due to presence of duplicate files, OS has to give an extra bit of time to search and identify particular file .Hence to solve this issue we are designing an application which will not only find duplicate files on your hard drive but also deletes the duplicate files that are not useful to user **.**We use MD5 algorithm to work on such files.

II. LITERATURE SURVEY

In Redundant file finder, remover in mobile environment through sha-3 algorithm writer says that ,Mobile environment provides storage as a main service. Data storage is a desired property when users outsource their data to be stored in a place irrespective of the locations. File systems are designed to control how files are stored and retrieved. Without knowing the context and semantics of file contents, file systems often contain duplicate copies and result in redundant consumptions of storage space and network bandwidth. It has been a complex and challenging issue for enterprises to seek de-duplication technologies to reduce cost and increase the storage efficiency. To solve such problem, Hash values for files has been computed. The hash function competition to design a new cryptographic hash standard `SHA-3' is currently one of the well-known topics in cryptographic research, its outcome heavily depends on the public evaluation. Testing the finalists in the competition for a new SHA-3 standard shows generally fast, secure hashing algorithms with few collisions. Focus of computation is performed for duplicate knowledge removal. Hash computation is done by the method of comparing files initially and followed by SHA3 signature comparison. It helps to reclaim valuable disk space and improve data efficiency in mobile environment,[1]

In Distributed Duplicate Detection in Post-Process Data De-duplication, Data De-duplication is essentially a data compression technique for elimination of coarse-grained redundant data. Since the advent of de-duplication the conventional approach has been to scale-up de-duplication at a storage controller by using more of the controller resources. This approach has led to several bottlenecks including the most evident one of hogging controller resources, in-turn leading to limiting the number of concurrent de-duplication threads running on the controller, finally ending up with poor de-duplication performance. Going by the rate at which we are experiencing data explosion, with data becoming the core entity separating one organization from other, high performing scalable de-duplication is one challenge organizations are already starting to face. Through the current effort, we propose a scalable design of a distributed de-duplication system which leverages clusters of commodity nodes to scale-out suitable tasks of a typical de-duplication system. We explain our distributed duplicate detection workflow, implemented in Hadoop's map-reduce programming abstraction. We also discuss the performance statistics we obtained with the scale-out de-duplication model,[2]

In Redundant Data, Duplicat , we come up with this application will use SHA-3 Algorithm to provides a good security for the data using the authentication format by generating hash code. Also the whole design has whole design has a simple hardware structure and fast running speed. But It is more time consuming. The cost of hashing-based methods goes up sharply as the number of collisions pairs of inputs that are mapped to the same hash value increases. [3]

In Connection: To Enhance File Search, It combines traditional content-based search with Context information gathered from user activity. we conclude that is To identify and store the relationships, Connections adds two new components: the tracer and the relation-graph. Connections demonstrates the benefits and flexibility of combining content and context in file search. Identifying connections are complicated. It takes more efforts to find contextual relationships between files. [4]

III. PROPOSED SYSTEM

We propose in this project is a type of procedure that finds and deletes duplicate files on your computer. You may have many duplicate text files on your computer, after numerous downloads from the Internet, or scattered over your home or corporate network.

Duplicate files are in most cases redundant and unnecessary, so keeping them is merely a waste of hard disk space. Your hard drives may be full of extra copies of documents waiting to be removed. Duplicate File searcher and Remover will help you reclaim valuable disk space and improve data efficiency. Deleting duplicates will help to speed up indexing and reduces back up time and size.

This product is easy to handle and works best for the standalone machine. In this project, the software does not interact with any other systems. It can quickly and safely find the unwanted duplicate files from the system and then delete or move the duplicate files to separate folder, according to the user requirement. The duplicates will be removed from your system.

3.1 ARCHITECTURAL DESIGN

The word project architecture intuitively denotes the high level structures of a software system. It can be defined as the set of structures needed to reason about the software system, which comprise the software elements, the relations between them, and the properties of both elements and relations.

The term software architecture also denotes the set of practices used to select, define or design software architecture.

Finally, the term often denotes the documentation of a system's "project architecture". Documenting software architecture facilitates communication between stakeholders, captures early decisions about the high-level design, and allows reuse of design components between projects.

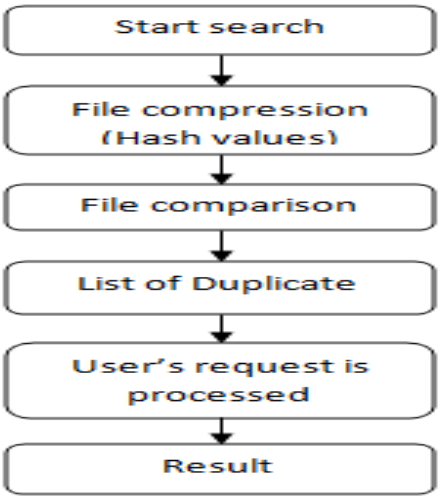

Figure 3.1 Architecture Design

3.2 MD5

MD5 processes a variable length message into a fixed-length output of 128 bits. The input message is broken up into chunks of 512-bit blocks; the message is padded so that its length is divisible by 512. The padding works as follows: first a single bit, 1, is appended to the end of the message. This is followed by as many zeros as are required to bring the length of the message up to 64 bits fewer than a multiple of 512. The remaining bits are filled up with a 64-bit integer representing the length of the original message.

The main MD5 algorithm operates on a 128-bit state, divided into four 32-bit words, denoted *A*, *B*, *C* and *D*. These are initialised to certain fixed constants. The main algorithm then operates on each 512-bit message block in turn, each block modifying the state. The processing of a message block consists of four similar stages, termed *rounds*; each round is composed of 16 similar operations based on a non-linear function *F*, modular addition, and left rotation. There are four possible functions *F*, a different one is used in each round.

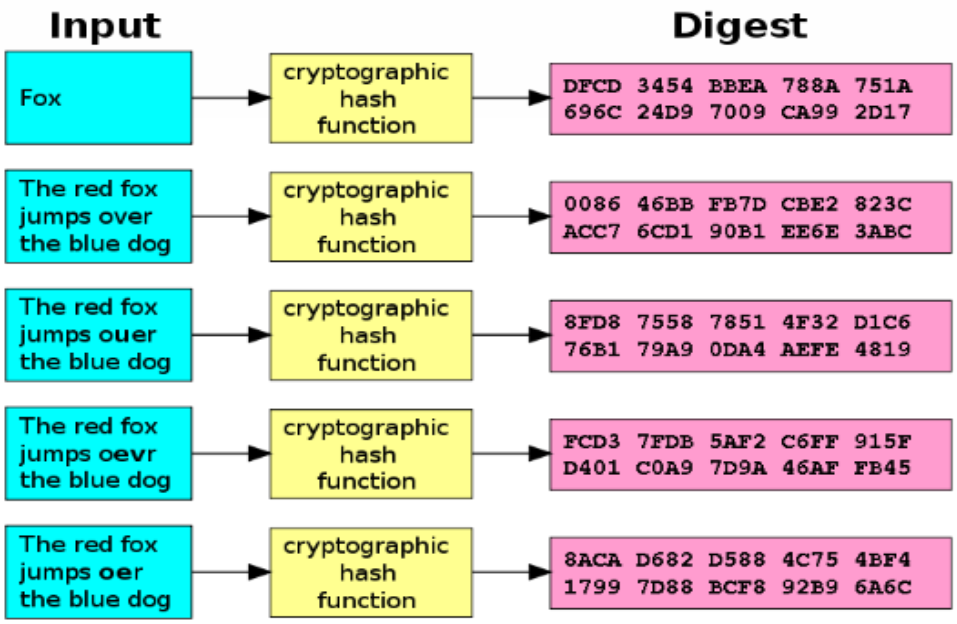

Figure3.2 Working of MD5 Hashing Algorithm

IV. EXPERIMENTAL RESULT

It can be concluded by saying that The Duplicate File Searcher and Remover is a powerful and reliable tool that makes dealing with duplicate files an easy and accurate job. It employs a powerful MD5 Message Digest hashing algorithm that ensures the precise detection of the duplicate files, based on their contents and not superficial aspects like their names. MD5 algorithm provides higher security and generates unique hash values for each unique input. This project helps you examine each file and decide which one to keep. It works best for a standalone machine.

Figure4.1 Home page

Figure4.2 Computer window

Figure4.3 Deletion Box

Figure4.4 Confirmation Box

V. APPLICATIONS

This software can be developed with client-server architecture. It can be used for following:

- Program Management
- Plagiarism Detection
- Program reuse and re-engineering

- Uninstallers
- Compression
- Clustering

REFERENCES

Redundant file finder, remover in mobile environment through sha-3 algorithm by meera.k, krishna sankar.p and sriram kumar.k [1]

Distributed Duplicate Detection in Post- Process Data De-duplication by Atishkathpal, Matthew John and Gauravmakkar[2]

Duplicut: Redundant File Searcher And Remover by Sumita Chandak Abhishek Kadam, Bhagyshree Gawade, jidgnesh Sanke.[3]

Connection: Using Context to Enhance file Search by craig A. N. soules, Gregory R. Ganger[4]

