RESEARCH ARTICLE OPEN ACCESS

Data Copy Sweeper

S.Sambasivam MCA., M.Phil.*, J.Senthil Kumar**
*Professor, Department of MCA, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.
Email: sammy2173@gmail.com
** Final MCA, Department of MCA, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.
Email: senthilkumarjayanjs@gmail.com

****

Abstract:

Growth of physical science in past number of years resulted in increase in computation powers, storage that demand to cope in numerous computational algorithms. With this there's almost ton of cupboard space provided in computers and there's no worry for the user to paste Brobdingnag Ian quantity of files in pc hard drives. But this advantage has one drawback during which persistently user isn't aware about range of files square measure traced by the users square measure duplicated. During this paper we've got focused on the techniques by that we will find duplicate files gift on the pc drives or the other removable device. We propose file detection systems to present associate degree indication to users concerning what number duplicate files square measure gift which can have either similar name or similar contents. The key idea of this method is to watch files in a very given directory or drive sporadically or at user's discretion and check similar files within the storage system therefore on save the house by proving choice to the user to delete the duplicate files. During this work, different attributes like file name, file size, file checksum, file content etc. are thought-about while looking duplicate files. Content based mostly file looking can offer correct results. Contents of files are compared computer memory unit by byte or by having check on samples of chunk of bytes within the files at completely different locations. The later methodology is time economical because it won't scan the entire file ensuing savage of your time.

*Keywords —**Checksum, duplicate.***

****

I. INTRODUCTION

We all recognize the price of disk drive storage space. Disk drive of a laptop is that the first storage area used to store info and if there are duplicate files keep it up your exhausting drive, then you're positively wasting the precious area. Because of such kind of duplicate files storing on disk drive, you may run out of memory area that makes a superb problem for you and can your system gets slowdown in its performance. Hence it is necessary to find and delete all those files that are having the same contents. Finding duplicate files manually in your disk drive isn't an easy task as it takes an excellent deal of some time and energy, and conjointly it could be risky as you'll delete a really important file by mistake. However with the help of machine-driven tools, you will be ready to simply realize all the duplicate files that are abode on your disk drive terribly simply. Remo tons of tool is best software package that's perfectly programmed to identify duplicate files in disk drive of a laptop. Typically it's found that on your laptop there exist tons of than one file having same name and same Content. Such files need to be known and deleted from disk drive to free memory area occupied by duplicate file. Tons of suites are one perfect software package that's developed with several

options to look out the duplicate files. It scans the selected folder or drive and if there are 2 files with same content, then it'll assist you to delete one of them and retains one in one click. It has got a special formula that performs rigorous scanning of folders and compares the content of two files although they have same names to make sure whether or not they are having completely different information or same.

With the help of Remo tons of software package, you can set searching criteria for searching the files. It is capable of searching numerous sorts of files including Words, Access, surpass so on. This tool produces a detail report when quick scanning and allow you to create alternative that file you would like to delete which to skip. It is a user friendly interface that creates it straightforward to use. It has flexible search criteria that helps user to customize search and helps in quick scanning. This software package will determine duplicate files altogether in Kinds of storage devices alongside USB drive, flash drive, iPod and camera. Additionally, duplicate file finder conjointly helps you to spice up your system performance by deleting unnecessary files and releasing memory area for further use. You'll no longer need to confuse for locating right files on the disk drive. Tons of suites are altogether safe and secure and there is no risk in victimization it. This is why; it is the primary alternative for several of the users for finding the duplicate file. If you often realize that you simply produce duplicate files as you rename or use files in another context, or just backup over you had supposed, Duplicate File Finder assist you realize and/or delete these duplicates and regain the cabinet space they antecedent occupied. Judgment from variety of the recent comments, it'd seem that some cautions need to be noted before victimization this type of software package. One can also note that with the cost per GB quickly decreasing, redundancy is no longer such a nasty factor. Don't ever simply blindly run the program, find all the duplicates so click "remove". This can be fateful within the sense that albeit filenames are identical, the content is usually completely different. As an example, you will have multiple icons with identical name, however of differing sizes; or you may need a photograph that has been modified once placed in an exceedingly completely different folder however retains identical file name.

II. OBJECTIVES

In duplicate file finder process most of technique involves some standard hashing algorithm. Hashing algorithm act as one way encryption process that generate signature of document on which it is applied i.e. each document will generate a unique hash. For duplication detection hash values of files are compared between. If both files are with same hash it indicates that file content are same thus duplicate. Also combined with strong light weight hashing function we are going to use all the attributes that define the current duplicate file search. These attributes include the current filter criteria (such as filename, masks, etc.), search paths, exclusion folders, file matching methods, visible duplicate result report columns, and more. In this paper we propose following methods to find duplicate files in the directory or folder.

A. CHECKSUM SEARCH

Search the selected drive or folder for duplicates that match by file name, extension, size and 128 bit content checksum. It is quick with a good degree of accuracy.

B. MATCHING OF FILES BASED ON CONTENTS WITH DIFFERENT EXTENSION.

Here, search is performed in selected drive(s) or folder(s) for duplicate files that match files by their contents as a whole byte by byte having different extensions. In this method we compare byte by byte and see if file match occurs. There will be increased accuracy to find out duplicate file with different extensions.

C. CONTENT BASED SEARCH

It will be done byte by byte and by taking samples of chunks of bytes at different predetermined locations in the files. Byte by byte content matching is slower as it needs to scan the complete file. However second method of

completion is time efficient as few bytes will be compared. The overall objectives of this paper are

- •Find all types of duplicate files
- •Accurate search of files and deletes them
- •Delete and rename of duplicate files
- •Compare files by their content and lets you to preview any files
- •Manage your documents, photos, songs, movies

III. ANALYSIS & SYSTEM ARCHITECTURE

Duplicate File Finder is a powerful tool to search for file duplicates on your computer. It can find duplicates of any files: text, binary, music or your system performance by deleting unnecessary files and freeing memory space for further use. And as a computer user you will no longer need to confuse for finding right files on the hard drive. More suites are totally safe and secure and there is no risk in using it. This is why; it is the first choice for most of the users for finding the duplicate file.
If you often find that you create duplicate files as you rename or use files in another context, or just backup more than you had intended, Duplicate File Finder help you find and/or delete these duplicates and regain the storage space they previously occupied. Judging from some of the recent comments, it would appear that some cautions should be noted before using this type of software. One might also note that with the cost per gigabyte rapidly decreasing, redundancy is no longer such a bad thing.

Don't ever just blindly run the program, find all the duplicates and then click "remove". This can be disastrous in the sense that even though filenames are the same, the content is sometimes different. For example, you may have multiple icons with the same name, but of differing sizes; or you may have a photograph that has been modified when placed in a different folder but retains the same filename. Don’t scan too much at once. A complete scan of drive C: would not be smart. However, scanning a few related folders at one time is much more efficient and the results less overwhelming. Despite the best software, human intelligence may still have to be used. It is best, like eating an elephant, to take one bite or one small scan at a time.

Algorithms used in the program allow for quickly analyzing the content of small and large files. The following criteria can be used search for duplicates: filename, file size, or file content. For multimedia files (MP3, OGG, WMA), the content of the following tags can be also analyzed: "Artist", "Album", "Title" and "Comment".

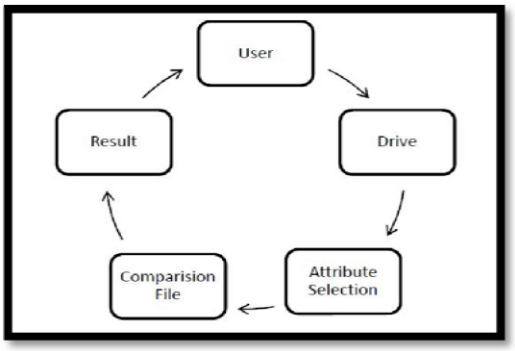

Figure 3.1: System Architecture

If you store hundreds or even thousands of documents, pictures, photos, music and video files, we often feel in doubt: it seems that we already have these file, but where?. Duplicate File Finder is here to help us. This program is a powerful yet easy-to-use tool for searching for file duplicates. It avoids wastage of our time browsing through directories to find duplicates, just leave this to this program, it will do it for us. However it is our duty to decide what to do with the found duplicates (copy, delete, move etc). Duplicate file finder analyzes the contents of our hard drive and we are able to see how many identical files we have. We can free disk space by deleting duplicates, but we need to be careful. This paper is based on tool that lets •Find files with same contents, same name.
•Find duplicate pictures, videos, songs etc.
•Works with removable media devices.
•Search local PC and over network.
•Identify and recover wasted disk space.
•Increases free space on laptops and memory disks.
•Reduce files searching time.

IV. WORKING OF ALGORITHM

SHA-3 Algorithm

1. Append padding bits

File is "padded" with a 1 and as many 0's as necessary to bring the content length to 64 bits less than an even multiple of 512.

2. Append Length

64 bits are appended to the end of the padded contents. These bits hold the binary format of 64 bits indicating the length of the original file.

3. Prepare Processing Functions

SHA1 requires 80 processing functions defined as:

•f(t;B,C,D) = (B AND C) OR ((NOT B) AND D) (0 <= t <= 19)
•f(t;B,C,D) = B XOR C XOR D (20 <= t <= 39)
•f(t;B,C,D) = (B AND C) OR (B AND D) OR (C AND D) (40 <= t <=59)
•f(t;B,C,D) = B XOR C XOR D (60 <= t <= 79)

4. Main loop

for i from 0 to 79
if 0 ≤ i ≤ 19 then
if = (b and c) or ((not b) and d)
k = 0x5A827999
else if 20 ≤ i ≤ 39
f = b xor c xor d
k = 0x6ED9EBA1
else
if 40 ≤ i ≤ 59
f = (b and c) or (b and d) or (c and d)
k = 0x8F1BBCDC
else if 60 ≤ i ≤ 79
f = b xor c xor d
k = 0xCA62C1D6
temp = (a leftrotate 5) + f + e + k + w[i]
e = d
d = c
c = b leftrotate 30
b = a
a = temp

V. RESULTS

Specify selection lists in section pane. The first list ("Search in folders") must contain the list of folders where you want to search for duplicate files. The second list ("Exclude folders") must contain the list of folders, which you want to exclude from search. The following hotkeys work in both lists: [Ins] add a folder to the list, [Del] delete a folder form the list, [Enter] select another folder. When leaving the program, the content of these lists is preserved and will be restored at the next program startup.
We have selected one directory containing some files present in the form of jpg, text, pdf, mp3, word. Now we can check the duplicate.

Logical Contains

There are certain cases in which message is logically same but physically different. Following is the example of this case. Consider two text files f1.txt and f2.txt. However their contents are not physically same but logically same. In such cases present work scenario is not able to handle this but we can extend this work to show that the files are duplicate. Save the space. Also many a times user may have created several versions of same file which may lead to the confusion. By using this system user gets flexibility to either delete duplicate files, remove, copy, paste & rename like operations. Due to that we can save the time of user so it is reducing the workload.

CONCLUSION:

Duplicate file finder system acts as tool to go looking duplicate files on given drive or directory. Searching duplicate files because the drive or directory is necessary for user perspective as he desires to save the house. Conjointly many another times user could have created many versions of same file that may cause the confusion. By victimization this technique user gets flexibility to either delete duplicate files, remove, copy, paste & rename like operations. Because of that we are able to save the time of user thus it's reducing the workload.

REFERENCES

[1] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, Members, "Duplicate Record Detection", IEEE Computer Society, Jan 2007.

[2] Jaspreet Singh, "Understanding Data Deduplication", on Druva Blog 2009 in Product Deep-Dives.

[3] Sanjay Jain, Puneesh Chaudhry, "Methods and apparatus for content-aware data deduplication", Apr 12, 2011

[4] Thorsten Papenbrock, Arvid Heise, and Felix Naumann, "Progressive Duplicate Detection" in IEEE Transactions on Knowledge and Data Engineering, Volume: 27, Issue: 5, May 1 2015, pp.1316 – 1329

[5] Data Deduplication – "Why, When, Where and How", Evaluator Group. Evaluation Guides, Starter, January 2, 2015.