ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences &Management Research

ANALYSIS OF OPEN SOURCE DUPLICATE FILE DETECTION TOOL

Mrs. R. Janani*[1] and Dr. S. Vijayarani[2]
[1]Ph.D. Research Scholar, Bharathiar University, Coimbatore.
[2]Assistant Professor, Bharathiar University, Coimbatore.

Keywords: *Text Mining, Information Retrieval, duplicate file detection, Open source tools, Performance Analysis.*

ABSTRACT
In this papers basically the analysis and modeling of the effects of dust and shading on the performance of the solar panel has been performed. The solar photovoltaic (PV) system has been simulated on Matlab/ Simulink simulation environment. The results are prepared using the experimental data and simulation model that has been prepared on Matlab/Simulink environment. After the preparation of the required graphs these graphs are compared. This research paper is based on general behavioral model forPV cell modeling & solar radiance for the conversion of solar intensity to electrical power.

INTRODUCTION
Text mining is a technique, which extracts information from unstructured data and finds patterns, which is novel and unknown earlier. It is also known as knowledge discovery from text (KDT), as it deals with the machine supported analysis of text. Text documents are in semi-structured or unstructured format datasets such as emails, full-text documents, HTML files, etc. The main problem of Knowledge Discovery from Text (KDT) is to mine explicit and implicit concepts and its semantic relations between concepts [1]. This aims to get insights into large quantities of text data. Text mining is a knowledge domain that utilizes techniques from the common field of data mining and also it combines methodologies from various areas like Information Extraction, Information Retrieval, Linguistics, Categorization, Clustering, Summarization, Topic Tracking and Concept Linkage.

An information retrieval (IR) system discovers information that is significant to a user's query. This system normally searches in collections of unstructured or semi-structured documents [2]. The need for an information retrieval system occurs when a collection reaches a size where customary cataloguing techniques can no longer survive. The general applications of information retrieval systems are digital libraries, media search, search engine like desktop search, mobile search, and web search etc.,

In most situations, users might also download the files which can be already downloaded and saved on their desktop [3]. Then, there is a chance of multiple copies of the files which are already saved in distinctive drives and folders on the system, which in turn reduces the overall performance of the machine and these files occupy quite a few memory areas. Analyzing the contents of the file and finding their similarity is one of the main troubles in textual content mining and records retrieval [4].

This paper is organized as follows, Section II discusses the various duplicate detector tools, Section III gives the analysis of these tools. The conclusion given in Section IV.

II. DUPLICATE FILE DETECTOR TOOL

The main objective of this research work is to analyze the performance of open source duplicate file detector tools. In this paper we have taken the following tools for analysis.

1. Duplicate Filter
2. Fast Duplicate File Finder
3. Duplicate File Eraser
4. Exact Duplicate Finder
5. Duplicate Commander
6. CloneSpy
7. Duplicate File Finder
8. AllDup

ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences &Management Research

9. Duplicate File Hunter
10. Slim Cleaner

2.1 DUPLICATE FILTER

Duplicate Filter is a duplicate file finder tool. It can find duplicate files instantly and users can compare and manage duplicate files easily. It searches duplicate files based upon Cyclic Redundancy Check (CRC) method. The user can search for duplicate files in their desktop computer as well as on network computers. They can rename, delete, or move the duplicate files [6]. This tool is used to find the duplicates based on the file size and file names. But the main disadvantage of this tool is not analyzing the content of the files. The following parameters are used in this tool detect the duplicate files.

- CRC
- CRC and File Name
- File Name
- File Size
- File Name and File Size

2.2 FAST DUPLICATE FILE FINDER

Fast Duplicate File Finder is the easiest way of finding and eliminating useless duplicate files from computer. It is the fastest and most accurate program of its kind on the market. It will scan the files, show a detailed report, provide easy to use tools to resolve duplicates safely and quickly. User can free up to 50% of their disk space [7].

This Tool helps to minimize disk space usage Helps to organize the user's media files or documents makes their file searches to run faster decrease size of backups. The main goal of this tool is to find the duplicate files based on file name and file extension [7]. Hence the same extension and the file name also same, then it will find that file is duplicate file. The parameters to check the duplicates are as follows,

- File Name
- File extension
- Similarity

2.3 DUPLICATE FILE ERASER

Duplicate File Eraser is a very small utility for PC to search and delete duplicate files. In this we can add a drive or folder to the list and click Start button to start a search. It can search using MD5, SHA1 or CRC32 method to find true duplicate files. The user can delete duplicate files after search easily by clicking 'Select all except one' to select only duplicate files. The program doesn't require installation and suitable for different types of OS. Important factors considered for duplicate file detection are as follows [8],

- File Size
- File Extension
- Algorithms
- File Type

2.4 EXACT DUPLICATE FINDER

Exact Duplicate Finder is a free duplicate file finder tool for Windows. It is used to find duplicate files in any location user specify using byte by byte comparison method. It can find files that are saved with different names. It shows the result in a grouped manner based on their location. It has various predefined criteria based on file types [9]. We can find all files or use any predefined criteria to find duplicate files. The factors are,

- File Extension
- File size with same Extension

ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences &Management Research

2.5 DUPLICATE COMMANDER

Duplicate Commander is a freeware application that allows the users to find and manage duplicate files on their PC. Duplicate Commander comes with many features and tools that allow end user to recover their disk space from those duplicates [10]. Some of the common features are,

- Search for duplicates based on filename or actual data
- When searching for duplicates based on file name, we may also factor in file extension, size, or timestamp.
- Duplicates found can be limited to certain file types
- Duplicates found can be limited to a certain file size bracket
- Search locations can be pinned and accessed easily later

2.6 CLONESPY

CloneSpy is a duplicate file finder and remover. It has various options to search for duplicate files. We have to first add a drive or folder to its pool list. We can search for duplicate files using one of the four criteria given [11]. It also has an option to automatically delete duplicate files (use it with caution) based on user search. User can also export the list of duplicate files in a TXT format. This tool process the files which

- are duplicates
- are duplicates having the same file name, title, or extension
- have the same file name or title
- have the same file name or title, and similar size
- are zero bytes long

This tool also handle the duplicate files. They are,

- deleting redundant files
- moving redundant files to a specific folder
- exporting a list of all equal files without removing any files
- replacing redundant files with shortcuts or hard links (NTFS) to retained files

2.7 DUPLICATE FILE FINDER

Duplicate File Finder is a free software for Windows to find and remove duplicate files. It can find various types of duplicate files including pictures, documents, spreadsheets, MP3 files, etc. User can scan for duplicate files in their disk and the program can delete the duplicate files easily. We can export the list of duplicates to TXT, CSV, or HTML format [12]. Some of the key features are,

- Powerful search engines (byte by byte and SHA-1)
- Find files with same contents, same name and zero size with the same extension.
- Find duplicate pictures, videos, songs (mp3, wma, ogg).
- Fastest among duplicate file finders and duplicate file cleaners.
- Works with removable media devices like pen drives, external hard disks, etc.
- Search local PC and over network.

2.8 ALLDUP

AllDup is a free software for searching and removing duplicate files on your computer. It has a fast search algorithm to find duplicates of any file type. The user can find text, documents, pictures, music, or movies with high speed. We can search duplicates with various criteria like file name, file extension, size, create

ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences &Management Research

date, modified date, file contents (byte by byte compare) etc. Removing duplicate file will recover valuable disk space [13]. Key features of this tool are,

- Search for duplicates of music and video files
- Search for duplicates of executable and any other files
- Export the search result to TXT or CSV file
- Find duplicates with a combination of the following criteria: file content, file name, file extension, file dates and file attributes!
- Detailed log file about all actions
- List non-duplicate files
- The unnecessary duplicates can be deleted permanently or copied/moved to a folder of your choice
- Search for hard links
- Many flexible options helps you to select unnecessary duplicates automatically

2.9 DUPLICATE FILE HUNTER

Duplicate File Hunter is a free and easy to use duplicate file finder software. It can search for duplicate files on the drive or folder which is specified by user. It takes few minutes to show the results, but the results are more accurate. It displays the name, path, size, and CRC32 for each file in the result[14]. We can be sure before deleting any duplicate file by matching CRC32 value.

- Search these folders for duplicates;
- Select any file from folder 01 in the search results;
- Select the "Safely Select Duplicates From This Folder in All Groups" feature using the popup menu of the search result list;
- Click the Delete button.

2.10 SLIM CLEANER

SlimCleaner is a disk administration programming. It additionally has a Duplicate File Finder tool. It has three settings: IntelliMatch Accurate Scan, Moderate Scan, and Quick Scan. Likewise has an alternative to look with particular document types or all record types. It has a rundown of the organizers to be disregarded. We can discover copy records and from the outcomes we can choose copy documents to erase. By erasing copy records we can recoup profitable disk space [14]. Some of the features are

- Most robust engine for analyzing and cleaning unneeded files that slow down a PC.
- SlimCleaner Free's new cleaning engine is fast and powerful, analyzing entire computers in as little as one second.

III. COMPARATIVE ANALYSIS

For this analysis, the synthetic folder was created and it has 15 files. These files are with same extension and also with different extension. This folder also has the duplicate files based on the content and duplicate file names. The size of the files is varied from 20 kb to 26758 kb. Table 1 shows the input files for this analysis.

Table 1 : Input Document Description

S.No	Name	Type	Size(Kb)
1	Big Data (1) - Copy	Microsoft Word Document	20
2	Big Data (1) - Copy	Adobe Acrobat Document	177
3	Big Data (1)	Microsoft Word Document	20
4	Big Data Analytics in Social Media	Microsoft Word Document	569

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

5	Ch8	Adobe Acrobat Document	1987
6	Chap Proposal (1) - Copy	Microsoft Word Document	540
7	Chap Proposal (1)	Microsoft Word Document	540
8	Cluster Tech	Microsoft Word Document	348
9	Clustering	Microsoft Word Document	348
10	Clustering	Adobe Acrobat Document	541
11	CProposal	Microsoft Word Document	23
12	Document Cluster	Microsoft Word Document	374
13	Figures	Microsoft Word Document	458
14	I_K_2015_KER	Adobe Acrobat Document	26758
15	New word doc	Microsoft Word Document	980

Table 2 describes the duplicate files based on their file names. This table shows only files which are having the same file name. Here extension and the content of the file doesn't matter.

Table 2: Duplicate Files Based on File Names

S.No	File Name	Type
1	Big Data (1) - Copy	Microsoft Word Document
2	Big Data (1) - Copy	Adobe Acrobat Document
3	Clustering	Microsoft Word Document
4	Clustering	Adobe Acrobat Document

Table 3 illustrates the duplicate files based on their file size. This table shows only files which are having the same file size.

Table 3: Duplicate Files Based on File size

S.No	File Name	Size (Kb)
1	Big Data (1) - Copy	20
2	Big Data (1)	20
3	Chap Proposal (1) - Copy	540
4	Chap Proposal (1)	540

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

5	Cluster Tech	348
6	Clustering	348

Table 4 illustrates the duplicate files based on their content of the file. This table shows only files which are having the exact file content.

Table 4: Duplicate Files Based on Content

S.No	File Name	Type
1	Big Data (1) - Copy	Microsoft Word Document
2	Big Data (1) - Copy	Adobe Acrobat Document
3	Big Data (1)	Microsoft Word Document
4	Clustering	Microsoft Word Document
5	Clustering	Adobe Acrobat Document

These same input only given to all the duplicate file detector tool and the tools will give the different results shown in table 5.

Table 5: Results for Various File Detection Tools

S.No	Tool Name	Limitations	Result
1	Duplicate Filter	This tool is used to check the files only with the same extension. Also, it will check the file names.	No Duplicates Found.
2	Fast Duplicate File Finder	This tool will check the dulicate files based on their contents. But they are in the same extension.	Duplicates Found. 1. Big Data (1) – Copy 2. Big Data (1)
3	Duplicate File Eraser	Based on the size this tool will find the duplicates with the same extension.	Duplicates Found. 1. Big Data (1) – Copy 2. Big Data Analytics in Social Media 3. Cluster Tech 4. Clustering
4	Exact Duplicate Finder	This tool will check the file names with the same extension.	No Duplicates Found.
5	Duplicate Commander	This tool used to search for duplicates based on their file size or actual data.	No Duplicates Found.
6	CloneSpy	This tool will check the files based on their contents. But they are in the same extension.	Duplicates Found. 1. Big Data (1) – Copy 2. Big Data (1)
7	Duplicate File Finder	Based on the size this tool will find the duplicates with the same extension. Also, it will find the	Duplicates Found. 1. Big Data (1) – Copy 2. Big Data Analytics in Social Media

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences &Management Research

		duplicates based on content with the same extension.	3. Cluster Tech 4. Clustering 5. Big Data (1) – Copy 6. Big Data (1)
8	AllDup	This tool will check the duplicate files based on their contents. But they are in the same extension.	Duplicates Found. 1. Big Data (1) – Copy 2. Big Data (1)
9	Duplicate File Hunter	This tool will check the file names with the same extension.	No Duplicates Found.
10	Slim Cleaner	This tool will check the duplicate files based on their contents. But they are in the same extension.	Duplicates Found. 1. Big Data (1) – Copy 2. Big Data (1)

IV. CONCLUSION

The text mining processes the unstructured information and it extracts meaningful numeric tokens. The text mining has numerous methods such as information retrieval, document similarity, information extraction, clustering and classification. This research work analyzes the performance of ten open source duplicate file detection tool. From this analysis, we found that some of the tools will detect the duplicate file within the same extension or same folders. Some of the tools are detecting and deleting the files within the same folders. The main disadvantages of these tools are, they are not able to detect the duplicate files with other extensions and other folders based on content. This is the important research challenge in the tool duplicate file detection. Hence, there is a need to develop a tool which detects and delete the duplicate files with different extensions and different folders by analyzing their content.

REFERENCES

1. *Ankur Singh Bist, Pattern Matching Algorithms for Computer Virus Detection, International Journal of Engineering Sciences & Research Technology, Singh 2(1), P.No.28-29, 2013.*
2. *Bin Wang, Zhiwei Li, Mingjing Li and Wei-Ying Ma, Large-Scale Duplicate Detection for Web Image Search, Multimedia and Expo, IEEE International Conference, 353-356, 2006*
3. *Bo Hong and Demyn Plantenberg, Duplicate Data Elimination in a SAN File System, In Proceedings of the 21st IEEE / 12th NASA Goddard Conference on Mass Storage Systems and Technologies, 2004.*
4. *Vishal Jain, Mayank Singh, Ontology Based Information Retrieval in Semantic Web: A Survey, International Journal of Information Technology and Computer Science(IJITCS), vol.5, no.10, pp.62-69, 2013. DOI: 10.5815/ijitcs.2013.10.06*
5. *StephaneDucasse, Matthias Rieger& Serge Demeyer, A Language Independent Approach for Detecting Duplicated Code, Proceeding IEEE International Conference on Software Maintenance, 109 – 118, 1999.*
6. *http://www.mindgems.com/products/Fast-Duplicate-File-Finder/Fast-Duplicate-File-Finder-About.htm*
7. *http://www.f2ko.de/en/index.php*
8. *http://download.cnet.com/Exact-Duplicate-Finder/3000-2248_4-75375597.html*
9. *http://download.cnet.com/Duplicate-Commander/3000-2248_4-75300625.html*
10. *http://www.clonespy.com/*
11. *http://download.cnet.com/Duplicate-File-Finder/3000-2248_4-10300084.html*
12. *http://www.allsync.biz/en_index.php*
13. *http://download.cnet.com/Duplicate-File-Hunter/3000-20432_4-75118901.html*
14. *https://www.slimwareutilities.com/*