DOI: 10.21917/ijsc.2021.0342

A FRAMEWORK FOR ANALYSING UNSTRUCTURED DATA IN COMPUTING DEVICES

J. Dhayanithi, M. Marimuthu, G. Mohanraj and P. Neelashkumar
Department of Computer Science and Engineering, Sona College of Technology, India

Abstract

In recent years all the real-world digital equipment are generating enormous amount of data because of technological development. Those data are unstructured and hard to analyse. This framework tool is developed to view the data on different formats depending on the business needs by the customers or end users. Data forms are difficult to interpret manually. Structuring data is needed for the today business world. This system deals with files alone, as there are many problems in unstructured data. Manual listing and maintaining is hard as mistakenly the files are deleted and sometimes lost by the users. A tool was framed for analyzing file properties and showing our analysed report to the user which means they can make decisions based on their own business standards to improve the performance of hard drives. This framework will render different unstructured data analyses and provide some visual representation.

Keywords:

Unstructured Data, Files, Analysis

1. INTRODUCTION

Every part of our daily life has changed to technology-based. Therefore, there is an immense amount of data that has arisen from different digital sources and we need to integrate it with data integration. A machine can store the data in all the drives as a temporary or permanent file. It is tedious to manage the data which is stored in the hard drives and it directly affects the performance of the hard drives. The evolution of data storage in computing device is start with electron tube of storage capacity 512 bytes to blue ray optical disk of storage capacity 25000000 kilo bytes. Now a day people move on to cloud data storage for managing their data in which the storage capacity is limitless.

In general all the available data are classified as Structured and Unstructured. The structured data has been organized into attributes and indexed for stress-free access. The structured data are often referred as relational database and Structured Query Language (SQL) is used to manage and analyse. The data generated through machines are classified as structured data for example, Web logs, sensory data, Point of scale data, etc. Unstructured data does not have any predefined model and it does not have any organization techniques. The data in social media, emails, archives, documents, audio, video and files are classified as unstructured. This data cannot be processed in a particular order, thus it is hard to manage compared to structured data. It was evident that our digital world generates 2.5 Quinton bytes of data every data in which 80% of available data are unstructured and only 20% are in structured manner. The data generation is still going up which states that we all in need of efficient data management techniques. It is very difficult to analyse those data to predict the irrelevant, unwanted and unused data. Even in personal computer (PC), analyzing the available or stored data should be carried out at regular span of time.

The data stored in the computing devices are in unstructured manner. The hard disk in the computing device is partitioned and all the information are stored in undefined manner which means that the computer users have to decide where to store the files in the partitioned area. The metadata about the file/folder is visualized by locating its properties. It shows the information like file format, path, size, date and time of file/folder created, accessed and modified, ownership and attributes. The process of analyzing unstructured data unifying it categories like file format, size and other metadata is called Data classification. In general, the data classification is carried out on two ways namely user and automated. In user, the data classification is done manually by scanning each document. The main advantage of this approach is user can have good judgment on whether the available is sensitive or not. It takes long time and user need patience to classify the data. In the automated data classification approach, a parser tool is used to analyse the documents. This approach takes only less time compare with user centric approach. The data classification is carried out many areas such as email server, webserver and individual enterprise or organization but there is no study or research is carried out on analyzing data in the personal computing devices.

The main purpose of data classification is to detect sensitive files, secure critical data, and track regulated data, optimizing the search, identifying the redundancy and empty records. Classification of file or folder is categorized into four different categories namely qualitative, quantitative, spatial and historical. In Qualitative, the files and folders are classified according to quality such as worm and virus. Quantitative relates to size of the files or folders. In Geographical, the categorization is based on where the files or folder located (Path).The file or folder creation accessed and modified is grouped to Chronological.

The main purpose of classifying the unstructured in computing device is to clean up the disk. The frequent cleaning up of disk improves the efficiency of hard drives, makes the computing device more reliable and maximizes the hard drive memory. The main contribution of this paper is to suggesting a framework for parsing the unstructured data in any computing device and suggesting an advanced cleanup process.

The rest of this paper is organized as follows. Section 2 reviews the background and related work of analyzing and classifying unstructured data. Section 3 explains the tool proposed for analyzing unstructured data in computing devices and also experimental results of the tool is explained in section 4. Section 5 enlightens the conclusion and future work to be carried out.

2. RELATED WORK

Most of the statistical research in business until recently dealt with structured data alone. With some of the organizations opening up to the concept of using unstructured data created by

customers such as complaints, service logs, social media etc. Many researchers are investigating the impact of such types of text data on business performance. One common field is under research to study the impact of online reviews on the sales of goods and services. The impact of online reviews has been researched in depth on only a few fields such as movies. While most of these works are analyzing the influence of feedback, few works include methods for integrating the variables into a predictive model. One such research stream in this field focuses on incorporating and combining expert analysis with statistical forecasts, as [1] and [2] have shown that, based on previous expert experience; improvements are made to statistical methods that have led to more reliable predictions.

Yu et al. [3] analysed vast volumes of online reviews of the films. He found that both the emotions conveyed in the reviews and the consistency of the reviews has a major effect on the success of potential sales. The senses are observed using Probabilistic Latent Semantic Analysis (PLSA). They suggest an autoregressive sentiment-aware model for sales prediction based on the emotions, and that model is further strengthened by considering a review quality factor.

Krasser et al. [4] proposed a system which used the SVM for the purpose of classification. This framework collects data on the actions of email senders based on distribution of global sending. It will evaluate them and assign a confidence value to any IP address that sends an email message. Classification of SVM is successful but that method always fails. This is because the spammer does not send spam mails from a set of IP address, instead they use separate IP addresses to send spam. In a variety of studies, researchers analysed texts from different social network services (SNS), blogs, and news to analyse associations between stock prices and public emotion as a reaction to social events and news [5]-[9].

Many research works have been carried out on classifying unstructured data with domain specific but lack of research in analyzing data in hard drives. In many aspects the data are stored in hard drives [10] [13] and it very important to manage those hard drives in an efficient way. Ardeshir et al. [11] had done an exhaustive case study on real time data driven quality monitoring on hard disk drives. In their study, the time to failure of hard disk drives are monitored in real time using Self-Monitoring, Analysis and Reporting Technology. Another promising work of analyzing the hard disk drives was carried out by Jing et al. [12]. In this approach the failure of hard disk drives are predicted using random forest technique.

In order to improve the computer functionally there are many disk cleanup software are available in the market. The most popularly used disk clean up software are CCleaner, Avast Cleanup, CleanMyPC, Treesize, Glary Utilities Pro and WinZip Utilities Suite. The main drawback of all available disks clean up software is platform dependent and its performance was limited. Disk clean up software should be regularly updated and it gave additional overhead to the user.

3. TOOL FOR ANALYSING UNSTRUCTURED DATA

The proposed tool for analyzing unstructured data comprises of two modules namely scanning module and analytical module. The tool architecture was shown in Fig.1. It was developed using python 3 and it is platform independent. In order to initiate the analysis process the python application should run in the computing device. It initially identifies the available number of partition in the computing device. For every identified partition determine whether OS is installed in it. In such cases, iterate to the next partition and the scanning process will begin. In the scanning module, and there is no human interaction carried out. All the files in the computing device or in the personal computer are taken into consideration during the scanning phase.

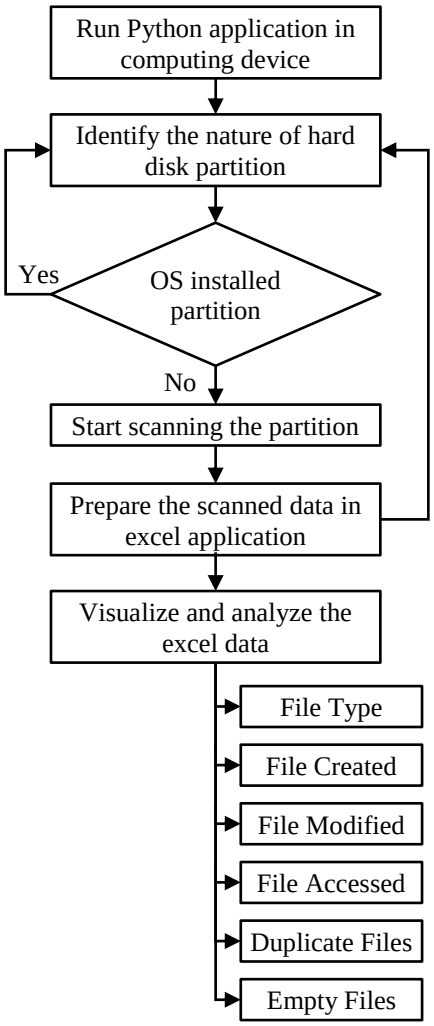

Fig.1. Architecture for analyzing unstructured data

The output of the scanning module is taken as input to the analytical module. The files in the system will be scanned initially, and all the metadata are collected and stored in Excel file which is given as input to the analysis process. The entire analysis process is carried out in three sequences of steps.

Step 1: Scans the system.

Step 2: Extract the metadata of files.

Step 3: Maintaining file metadata as source, conducting analysis such as counting files, file type, when the file is created, when the file modified, when the file is last accessed, duplicate files and empty files.

This analysis processes will help the user who using the computing device to make a decision on cleaning up the files or

folders. In the entire hard disk, the partition where the OS was installed is never touched because all the system supported files contains many sensitive files that are located. The proposed tool was run in a personal computing device having the configuration of 500 GB hard disk with four partitions namely C, D, E and G. In C partition, the operating systems were installed and it will never consider for the analysis process. Once the Scanning process is completed it outputs an excel sheet.

4. EXPERIMENT RESULTS

The analysis process outputs in the form of visual representation and provide guidance to the user to make valid decision on the files and folders available in the hard drives. The Fig.2 depicts the graph of top 15 files based on its size. It takes the x-axis as file extension and the y-axis as number of files. In all the drives except partition C, 195 files are PDF, 194 files are word document, and 176 files are mp4 format. The bar chart clearly shows that the user mostly works on PDF and word document. The Fig.3 reflects the graph of top 15 percentages of files available in the hard drives. The visual representation made it easier for the user to access the details.

In Fig.4, the graph was represented with respect to the file last accessed in year wise. In our system, 0.8% files were accessed during 2015, 0.2% files were accessed in 2016 year, 19.8% files were accessed in 2017 year, 41.8% files are accessed in 2018 year and 37.4% files are accessed in 2019 year. From the figure it was clearly stated that 0.8% files in the hard drives are never accessed since 2015 and user will take a decision on taking backup of those files or delete the files form the hard disk. The Fig.5 Depict month wise file last accessed. In our system, 48.5% of files were last accessed in January month on year 2019, 22.8% of files were last accessed in February month and 28.6% of files were last accessed in March month.

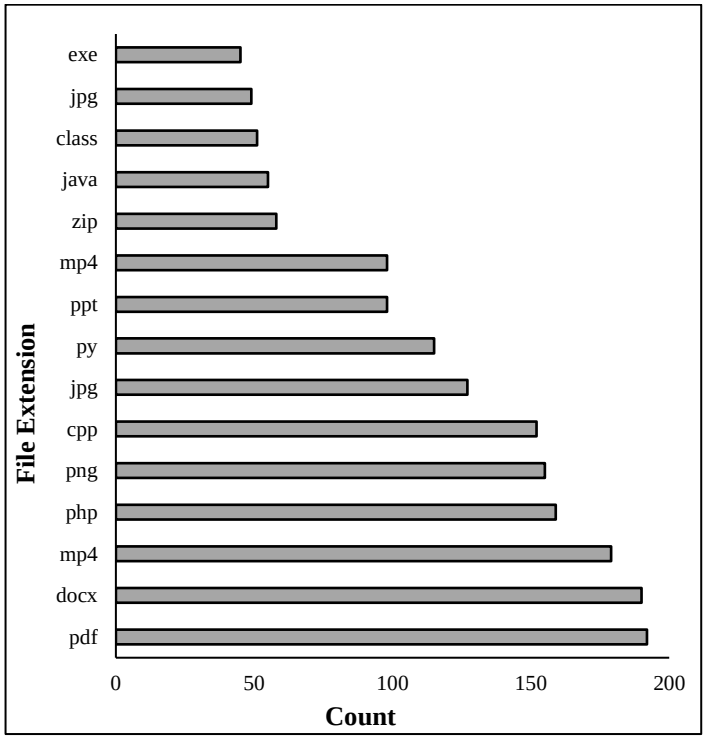

Fig.2. File Count Analysis

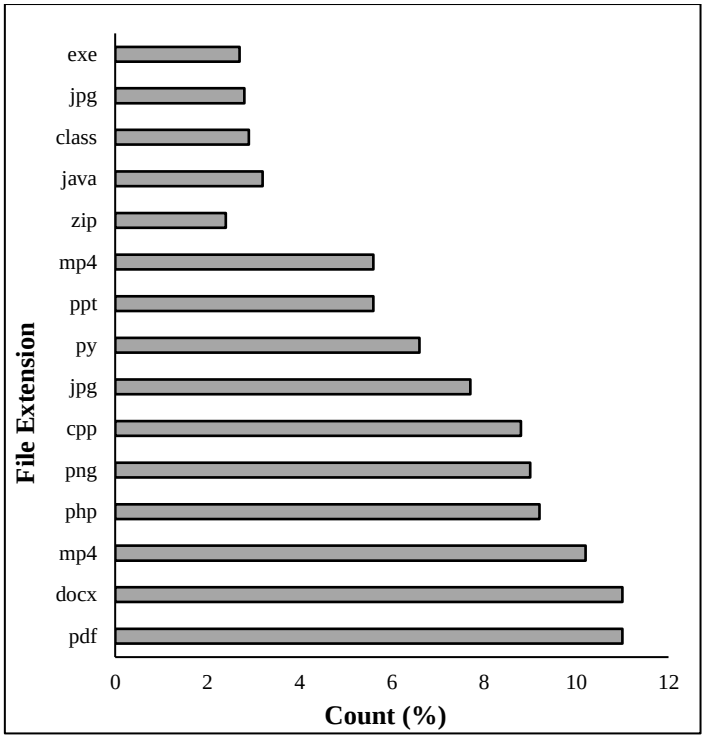

Fig.3. Files Percentage Analysis

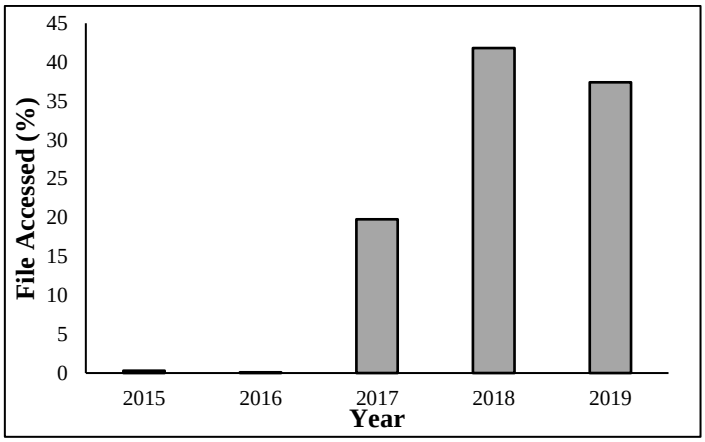

Fig.4. File last accessed Year wise

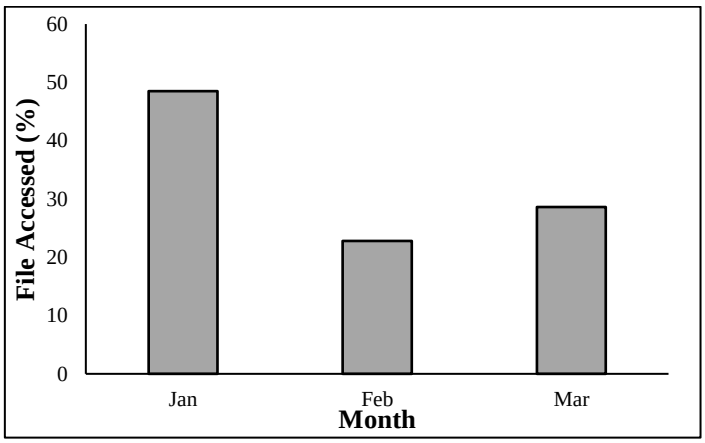

Fig.5.File Last accessed Month wise

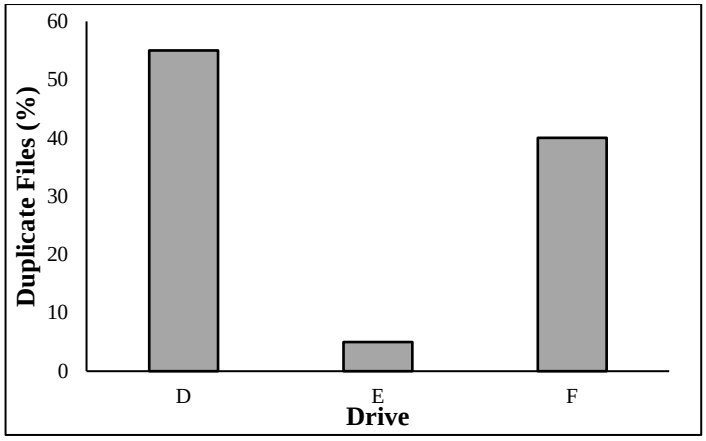

Fig.6. Duplicate/Redundant Files Analysis

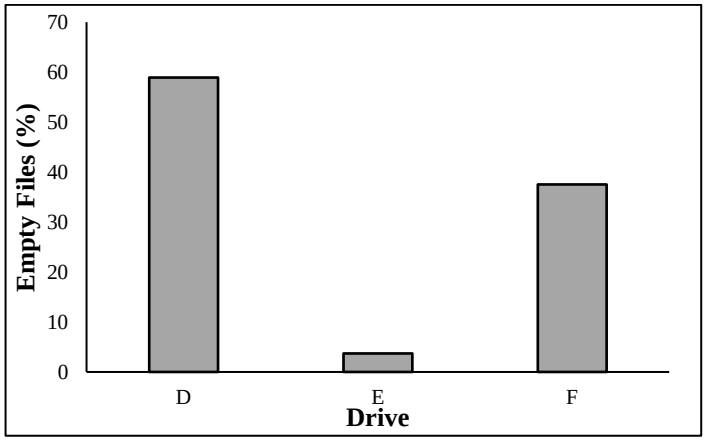

Fig.7. Empty Files

The Fig.6 represents the amount of files duplicated in various drives. D drive having 55% duplicate file, E drive having 5% duplicate file and G drive having 40% duplicate file. D drive which has more duplicate file and E drive which has less duplicate data. The proposed define and report to the user regarding the files which are generated for multiple times. The existence of duplicate files has been tested and described in excel folder, and includes even the empty files. Finally. it shows the amount of bytes lost due to duplicate data.

Table.1. Overall Processing Time

Process	Time (s)
Scanning process	2.12
Grouping of Files	0.56
Last Accessed Year wise process	0.44
Last Accessed Month wise process	3.45
Duplicate file process	7.79
Empty file process	0.64
Folder wise process	12.38

The proposed tool only checks the duplication/redundant in the same extension. There is a possibility of having same file content in different extension. Such files are also need to be identified for the efficient use of the hard drives.

The Fig.7 represents the percentage of files that are created with empty content in various drives. Those files are also properly handled for the efficient use of hard drives. D drives have 58.9% empty files, E drive has 3.7% empty files and G drive has 37.5% empty files. The tool will not delete any files or folders automatically but it will report to the user about the status. Only user has to decide what action needs to be taken on improving the performance of hard drives.

The overall processing time of the tool is narrated in Table.1 along with the time taken in seconds. The process of scanning takes 2.12 seconds. Segregation of files in terms of last accessed year and month takes 0.44 and 3.45 seconds respectively. The process of identifying duplicate files and empty files takes 7.79 and 0.64 seconds. Finally, folder analysis process takes 12.38 seconds.

5. CONCLUSION AND FUTURE WORKS

The proposed tool will be used for analyzing unstructured file data in any local environment. The tool outputs percentage of data, year-wise data access, month-wise data access, duplicate, unused and empty files report with visual representation. This approach is further extended to analyse the records in mail server and cloud server. When it comes to cloud server, enormous and more powerful algorithm would be required to analyse the data. In many computing device there is a possibility of same documents are available in different file extension. The proposed framework analyses the redundancy within the same extension and it should be further extended by incorporating machine learning algorithm to find the redundancy in different file extensions.

REFERENCES

[1] A. Davydenko and R. Fildes, “Measuring Forecasting Accuracy: The Case of Judgmental Adjustments to SKU-Level Demand Forecast”, *International Journal of Forecasting*, Vol. 29, No. 3, pp. 510-522, 2013.

[2] P.H. Franses and R. Legerstee, “Do Statistical Forecasting models for SKU-Level Data Benefit from Including Past Expert Knowledge?”, *International Journal of Forecasting*, Vol. 29, No. 1, pp. 80-87, 2013.

[3] Xiaohui Yu, Yang Liu, Xiangji Huang and Aijun An, “Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 4, 2012.

[4] Yuchun Tan and Sven Krasser, “Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis”, *Proceedings of IEEE International Conference on Global Telecommunications*, pp. 1-5, 2008.

[5] J.R. Nofsinger, “Social Mood and Financial Economics”, *The Journal of Behavioral Finance*, Vol. 6, No. 3, pp. 144-160, 2005.

[6] W. S. Chan, “Stock Price Reaction to New and No-News: Drift and Reversal after Headlines”, *Journal of Finance Economics*, Vol. 70, No. 2, pp. 223-260, 200

[7] V. Nadine, “Lagging Behind? Emotions in Newspaper Articles and Stock Market Prices in the Netherlands”,

Journal of Public Relations Review, Vol. 42, No. 4, pp. 548-555, 2016.

[8] G. Ranco, D. Aleksovski, G. Caldarelli, M.Grcar, and I. Mozetic, "The Effects of Twitter Sentiment on Stock price Returns", *PLOS One*, Vol. 10, No. 9, pp. 1-12, 2015.

[9] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market", *Journal of Computer Science*, Vol. 2, No. 1, pp. 1-8, 2011.

[10] P.K. Gopalakrishnan and S. Behdad, "Usage of Product Lifecycle Data to Detect hard Disk Drivers Failure Factors", *Proceedings of International Conference on Design Engineering Technical Conferences and Computers and Information in Engineering*, pp. 1-8, 2017.

[11] Ardeshir Raihanian, Mashhadi, Willie Cade and Sara Behdad, "Moving towards Real-time data Driven quality monitoring: A Case Study of hard Disk Drives", *Procedia Manufacturing*, Vol. 26, No. 1, pp. 1107-1115, 2018.

[12] Jing Shen, Jian Wan, Se-Jung Lim and Lifeng Yu, "Random Forest based Failure Prediction for Hard Disk Drives", *International journal of Distributed Sensor Networks*, Vol. 14, No. 11, pp. 1-9, 2018.

[13] N. Kuznietsova and M. Kuznietsova, "Data Mining Methods Application for Increasing the Data Storage Systems Fault-Tolerance", *Proceedings of International Conference on System Analysis and Intelligent Computing*, pp. 1-4, 2020.