IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

A Technique for Classifying Unstructured Big Data Files

M. T. Nafis[1*], R. Biswas[2]

[1] Department of Computer Science and Engineering,JAMIA HAMDARD,New Delhi, India
[2] Department of Computer Science and Engineering,JAMIA HAMDARD,New Delhi, India

**Corresponding Authior: tabrez.nafis@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 07/Jun/2018, Published: 30/Jun/2018

Abstract— In the era of technological development, more and more data is being accumulated on daily basis. Therefore it is a challenge to process, store and manage those ever increasing size of data. Before processing, classification of large amount of files needed. In this paper the authors develop a method to classify large amount of unstructured big data files into a small number of groups, each containing structured data files. The objective is that after classification the classified groups will be a better form of resource to the concerned user for further processing.

Keywords— Big data, r-train, identity tag, Classification, Unstructured data.

I. INTRODUCTION

In recent years, the fast development of Internet, IoT, and Cloud Computing have led to the huge growth of data in almost every industry and business area. Big data has rapidly developed into a hot topic that attracts extensive attention from academia, industry, and governments around the world[13].In this work a new method is conceptualized to classify a large amount of unstructured files into few groups so that each group will contain structured files. What kind of groups are to be developed is completely by the prior choice of the concerned user. This method of classification will be useful in the service centers like: Health Care, sports, Agriculture, Education, etc. to list a few only out of many. The method can be applicable to those involved in data analytics, in particular while dealing with large amount of files of common interest, a method which can be regarded as file mining from a file warehouse, where number of files is very large in number.

Our paper is divided into four sections. The first section introduces the need of classification for Big Data. Second section highlights some basics characteristics of Big Data. Third section explains the techniques for classifying unstructured data files.

The last section describes the conclusion of the work.

II. PRELIMINARIES

In today's era the Internet represents a big platform where large sizes of data get accumulated and added every day. The IBM Big Data Flood Infographic indicate that 2.7 ZB of data exist in the digital world today. According to this study there are 100 TB of data updated daily through FB, and a lot of activity on social media will be leading to an approximate of 35 ZB of data generated yearly by 2020. In order to have an idea of the amount of data being generated, one ZB equals 10^{21} bytes, meaning 10^{12} GB. [1]

"Big Data" term was first introduced to the digital world by Roger Magoulas from O'Reilly media in 2005 in order to define a large amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data.[2] Also, in Gartner's IT Glosarry Big Data is defined as high volume, velocity and variety information assets that need cost-effective, innovative forms of information processing for enhanced insight and decision making. [3]

The Big Data has four major characteristics (besides few more):

Volume: signifies the quantity of data accumulated by a company. This data must be used again to get necessary knowledge;

Velocity: signifies the time in which data can be processed. Few activities are very necessary and deserve urgent responses, due to which fast processing enhances efficiency;

Variety: signifies the type of data that constitutes Big Data. This data can be of any one among: structured, semi structured and unstructured;

Veracity: signifies the degree in which a person uses information in order to make decision. So getting the right interrelations in Big Data is very vital for the business strategy. [4]

Among them, Variety, which signifies that origin, may consists of both structured and unstructured data. Structured data are those that are organized in a structured form, which can be searched in a simplest way, leading to specific

knowledge and access it in a fast and compact way. Examples of structured data are data stored in the database.
Unstructured data, on other hand is of opposite features. As it is unstructured in nature, it is a lengthy task to point specific information and use it efficiently. Examples of unstructured data are simple text documents, pictures, videos, Web documents, and.etc. Although different in structure, it speaks about the same elements of the world and relationship among them. Retrieving both unstructured and structured data in an integrated manner is an important research area that is also has commercial value. An efficient classifying technique for accessing these mixed data is of more importance [5].

III. METHOD OF CLASSIFICATION

Consider a researcher having interest in information on a certain area, say, 'Cancer detection', for rigorous study and analysis for his research work. Free (or paid) information is available in the cyber space in very large number on the area of his interest "Cancer detection". We assume that all these large amounts of information are files, but of various sizes, various types, etc. Many files are word files, many are pdf files, many are jpeg, tif etc files, many are zip/rar files, many are MP-3/4 files, many are video files, etc.
A. Ahad ,et al. provides a comprehensive study of the prominent distributed file systems like HDFS, Cassandra and Quantcast file system for handling big data[6]. [7] and [8] explore in detail the relevant use of HDFS in processing files of large sizes. After classifying and clustering, a secure technique have been provided[9].
Thus, before seeking the information it is completely uncertain which structured files are in what numbers. Suppose that by download operation the researcher has got a database D of N number of unstructured files. Suppose that storage of these files are done using ADS[10] distributed system and hence storage is not a problem to him/her but the main problem to him/her is to classify these large amounts of files into various precise classes:

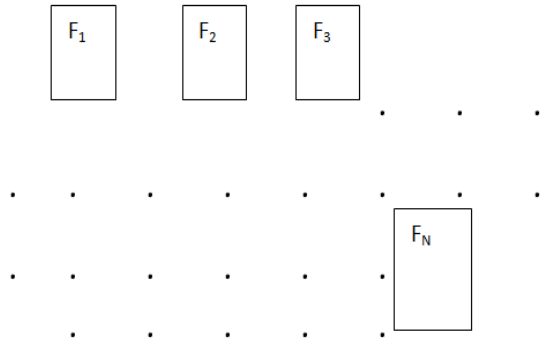

Fig. 1 Big dataset D of large amount (N) of files

Suppose that the following are the categories of files as shown in Fig. 2, in which we have to logically restore all these N files $F_1, F_2, \ldots F_N$ after classifying precisely:

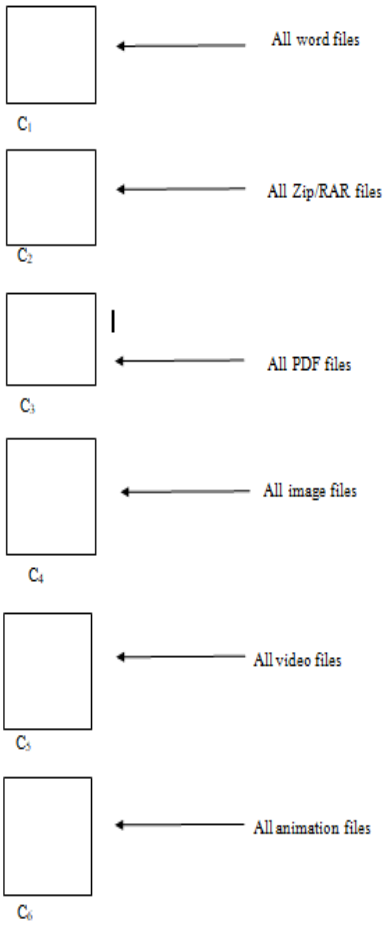

Fig. 2 Classification of Big Data Files

For the sake of presentation of our methodology here, we mention only six classes (groups) above. But it can be more in number, can be scalable in our method with no loss of generality.
The first step for this is to issue "Identity Tag" to all the unstructured files $F_1, F_2, \ldots F_N$.
The Identity tag of a file F_i will be as shown in the Fig 3, if the data structure r-train [11] or atrain [12] be used for storage.
An Identity Tag of a file has the following five fields, which can be designed with more number of fields, with flexible length of the fields to be decided by the developer:

(i) File Name: first 32 bits
(ii) Address : address of the file
(iii) Size : size of the file
(iv) Category : c_1 or c_2, … or c_6
(v) Address of Next Coach : 20 bits (i.e. next tag)

Some of the information to be filled in the field space are available in the 'Property" details of the concerned file. These information are to be fetched from the corresponding

property to fill-up the Identity Tag of the concerned file (say F_i).

File Name	Address	Size	Category	Address of the Next Coach
32 bits	8 bit	6 bytes	3 bytes	20 bits

Classification, as per requirement of the user, may have complex type of constraints, say: to group all TIF files which are at most of 10 MB size or greater than 10 MB. Thus, the above Identity Tag may be made more informatics by increasing its fields and its length(size), which can serve the users better.

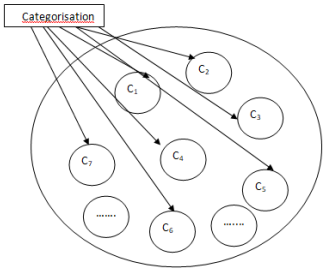

Fig. 3 Categorisation of files

C_1 = All images each of size (<= 1 GB)
C_2 = All images each of size (1 < s <= 10 GB)
C_3= All images each of size (10 < s < 100 GB)
C_4= All videos each of size(<= 1 GB)
C_5= All videos each of size(1 < s <= 10 GB)
C_6= All videos each of size (10 < s < 100 GB)
C_7= All audios each of size(<= 1 GB)
C_8= All audios each of size(1 < s <= 10 GB)
C_9= All animations each of size(<= 1 GB)
C_{10}= All animations each of size(1 < s <= 10 GB)
C_{11}= All pdf each of size(<= 1 GB)
C_{12}= All pdf each of size(1 < s <= 10 GB)
C_{11}= All word files each of size(<= 1 GB)
C_{12}= All word files each of size(1 < s <= 10 GB)

Storing of Identity tags in memory can be easily done by using r-train[10] data structure(for structured data) or r-atrain[11] data structure(for unstructured data), and in case of big data ADS system[9] may be adopted.
This work can be used in predicting malicious program using classification techniques and for secure data in MapReduce[14,15].

IV. CONCLUSION

In many cases, a large number of data files on a pre-chosen area of interest or on a phrase of interest are downloaded or collected by the user and in this work it is presumed that storing these large numbers of files is not an issue under study for this work. In that case, initially every user (here, a researcher) seeks to become more disciplined by classifying them into several groups each of structured nature for next course of action and/or application. In this work the authors have proposed a method on how to classify a large number of unstructured data files into several pre-defined groups precisely.

REFERENCES

[1] G. Noseworthy, Infographic: Managing the Big Flood of Big Data in Digital Marketing, 2012 http://analyzingmedia.com/2012/infographic-big-flood-of-big-data-in-digital-marketing.
[2] H. Moed, The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature, 2012, ResearchTrends, http://www.researchtrends.com
[3] Gartner, Big Data Definition, http://www.gartner.com/it-glossary/big-data.
[4] P. Zikipoulos, T. Deutsch, D. Deroos, Harness the Power of Big Data, 2012, http://www.ibmbigdatahub.com/blog/harness-power-big-data-book-excerpt.
[5] C. Zhu, Q. Li, L. Kong, and S. Wei, A combined index for mixed structured and unstructured data, Proc. - 2015 12th Web Inf. Syst. Appl. Conf. WISA 2015, pp. 217–222, 2016.
[6] Ahad, Mohd Abdul, Biswas, Ranjit, Comparing and Analyzing the Characteristics of Hadoop, Cassandra and Quantcast File Systems for Handling Big Data. Indian Journal of Science and Technology, [S.l.], 2017. ISSN 0974 -5645.
[7] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, The Hadoop Distributed File System, IEEE (2010).
[8]Dhruba Borthakur, HDFS Architecture Guide, Apache Foundation https://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf, (2008).
[9] Md Tabrez Nafis, Ranjit Biswas, A Secure Clustering Technique for Unstructured and Uncertain Big Data. In Springer proceedings of 1st International Conference on Advanced Computing & Intelligent Engineering(ICACIE)(Springer),451,2016.
[10] Biswas, Ranjit., Atrain Distributed System (ADS): An Infinitely Scalable Architecture for Processing Big Data of Any 4Vs, in Computational Intelligence for Big Data Analysis Frontier Advances and Applications: edited by D.P. Acharjya, Satchidananda Dehuri and Sugata Sanyal, Springer International Publishing Switzerland 2015, Part-1, 1-53 (2015).
[11]Ranjit Biswas, r-Train (Train) : A New Flexible Dynamic Data Structure, INFORMATION : An International Journal (Japan), Vol.14(4) April'2011, page 1231-1246.
[12] Ranjit Biswas, Heterogeneous Data Structure R-Atrain, INFORMATION : An International Journal (Japan), Vol.15(2) February'2012, pp 879-902(2012) International Information Institute of Japan & USA).
[13]Jin, X., Wah, B.W., Cheng, X.,Wang, Y., Significance and Challenges of Big Data Research, Journal Of Big Data Research(Elsevier),2017.
[14] K. Thyagarajan, N. Vaishnavi, "Performance Study on Malicious Program Prediction Using Classification Techniques", International Journal of Computer Sciences and Engineering, Vol.6, Issue.5, pp.59-64, 2018.
[15]H. Kousar, B.R.P. Babu, "*Efficient Map/Reduce secure data using Multiagent System*", International Journal of Computer Sciences and Engineering, Vol.6, Issue.5, pp.144-149, 2018.

