ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Data Extraction Techniques for Data on different Environment: A Review

Robonomics AI Private Limited[1], Bhavna Kamble[2], Milind Nemade[3],Vaishali Wadhe[4]

L-1840, Purva Highlands, Survey#19, Mallasandra Village, Holiday Village Road, Kanakapura Road, Bangalore, India[1]

Student, Department of Electronic Engineering, K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, India [2]

Professor, Department of Electronic Engineering, K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, India [3]

Associate Professor, Department of Electronic Engineering, K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, India [4]

ABSTRACT: In the field of computer science, a data structure is a data management, storage and organization format which enables an efficient access and modification. Data Extraction can be defined as retrieving data from source and then further utilizing it as per required necessities. Data Extraction is a process of extracting unstructured, structure semi-structure data from the user requirement based upon data warehouse or web or any type of automation level etc. Structured data is data that has been organized into a formatted repository, typically like a database, so that its elements can be made addressable for more effective processing and analysis. Unstructured data (or say unstructured information) is the information that either does not have a predefined data model or is not organized in a pre-defined manner. This paper will review and give a basic ideal understanding concept overview to existing techniques on data extraction for data on different environment. This review is aimed to discuss different data extraction approaches together with the basic tools algorithms for extracting favour data from different sources mainly such as OCR, Data Storage and analysing Methods, WEB Based techniques etc..

KEYWORDS: Structure, Unstructured, semi structure, OCR, Web based Technique.

I. INTRODUCTION

In today's era, these 21st century is mostly known as information age, has an affected on every sphere of human life or human phrase in the form of moderation of education (E-Learning),in medical, in banking,, in sports, in business etc. Thus, resulted into a very large volumes of data stored I the form consisting of numerical figures and text documents, to more complex composite information such as raw data, spatial data, multimedia data, and hypertext documents, to take complete advantage of data, and hypertext documents, To take complete advantage of data, the data retrieval is simply not enough, it requires a tool for automatic summarization of data, its extraction of the information which has stored, and the discovery of pattern in the raw data. Data extraction is where data is analysed and crawled through to retrieve relevant information from data sources(like a database) in a specific pattern. Further data process in data workflow.Most of the data extraction comes from unstructureddata sources and different data formats. This unstructured data can be in any form, such as tables, indexes, and analytics etc.

The work in this paper is divided in stages namelyOCR, Data Ware House and Data Analysis. This paper further comprised of paper having research in the fields of extraction data namely study of data analysis model based on Big Data [1].Hybrid OCR techniques review application based on script cursive language [2]. Performance of document

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

image OCR systems for recognizing idea texts on embedded platform [3].Web approach for web data extraction [4].Web Extraction Technique : A Review [5].With the help of these, previously published paper we have tried to put main focus on the technique's used for extracting data Paper is organized as follows. Section II describes introduction to each techniques which are mentioned above in terms of definition, usage, working, advantages and disadvantages, scopes, availability and etc. In Section III applications have discussed.

II. INTRODUCTION FOR VARIOUS TECHNIQUES

A. Optical Character Recognition (OCR)

OCR can be defined as Optical character recognition or optical character reader, which is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine -encoded text, from a scanned document or a photo of that document or a scene-shot.It is widely used in form of information entry from printed paper data records such as passport documents, invoices, bank statements, computerised receipts, business cards, mail, printouts. They can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, text-to-speech or v/s.Early versions were trained with images of each character and worked on one font at a time. Advanced systems are now capable of producing a high degree of recognition accuracy for most fonts, and reproducing formatted output that closely approximates the original page including images, columns, rows, and other non-textual components, etc.

Early optical character recognition was traced to technologies involving telegraphy and creating reading devices for the blind.[6] In 1914, Emanuel Goldberg developed a machine that read characters and converted them into standard telegraph code. Meanwhile, Edmund Fournier dAlbe introduced the Octophone, that's a hand-held scanner when moved across a printed page, produced tones that corresponded to specific letters or characters.[7]
In late 1920s and into the 1930s Emanuel Goldberg developed a searching microfilm archives using an OCR system named as Statistical Machine. Further, in 1931, his invention was granted USA Patent number 1,838,389 which was acquired by IBM. With discovery of smart-phones and smart glasses, OCR used in internet connected mobile device applications to extract text captured using the camera build in device. Due to absence of OCR functionality (in operating system), typically used an OCR API [8][9] and it returns the extracted text, along with information to location of the detected text in the original image back to the device app for further processing (such as text-to-speech) or display.

OCR engines have been developed into many kinds of domain-specified OCR applications, such as receipt OCR, invoice OCR, check OCR, legal billing document OCR.OCR is a field of research in pattern recognition, artificial intelligence and computer vision.There are mainly two types in OCR which always targets typewritten text, one glyph or one character at a time.Intelligent character recognition (ICR) targets one glyph or character at a time from script, usually involving machine learning.Intelligent word recognition targets one word at a time from script, which is useful for languages where glyphs are not separated in cursive script.

A data warehouse is a technique for collection and managing data from various sources to provide meaningful business insights which consists of technologies and components by using the strategic data. It is an electronic storage of a large amount of information, designed for queries and analysis instead of transaction processing and making it available for users in a timely manner to make a difference. Key events in evolution of Data Warehouse In 1960, Dartmouth and General Mills were partners in research project, develop the terms dimensions and facts. In 1970- dimensional data marts for retail sales were introduced by A Nielsen and IRIIn 1983 a database management system which is specifically designed for decision support was introduced by Tera DataCorporationIn the late 1980s there was evolution started in Data warehousing by IBM worker Paul Murphy and Barry Devlin developed the Business Data Warehouse. However, the real concept was discovered by Inmon Bill. Inmon Bill is considered as a father of data warehouse. He had also written about a variety of topics for building, usage, and maintenance of the warehouse the Corporate Information Factory[11]. A Data Warehouse works as a central repository for all data, whose information arrives from one or more

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

data sources. Data may be Structured, Semi-structured or Unstructured data. These data are processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets. By merging all this information from different sources in one place, it helps to be organization to analyse its customers more holistically, which helps to ensure all the consideration, the information available. Data warehousing makes data mining possible. It allows business users to quickly access critical data from sources all in one place. It provides consistent information on various cross-functional activities. It helps to integrate many sources of data to reduce stress on the production system Restructuring and Integration make it easier for the user to use for ad-hoc reporting and analysis by supporting queries for different time periods and trends to make future predictions. Creation and Implementation makes time confusing affair. It can be outdated relatively quickly, thus it is difficult to make changes in data types and ranges, data source schema, indexes, and queries. It is too complex for the average users. Data warehousing project scope will always increase. Sometime warehouse users will develop different business rules. Organisations need to spend lots of their resources for training and Implementation purpose. Data Analysis is done with help of these three following concepts to recognize the data which we are trying to be retrieving, storing or transforming.

Data mining refers to the technique of searching useful and relevant information from the data ware Brought together by the goal of meeting challenges of the previous researches, from various disciplines began focusing on developing more efficient and scalable tools which could handle diverse types of data. This work which culminated in the field of data mining, built upon the methodology and algorithms that researches had previously used.in house. It involves the extraction of hidden, predict from large data base. It is also known as a powerful technology with great potential to analyse important Information in the data warehouse. Its techniques have outcome of the rigorous research and product development.[12].The evolution of data mining process began where business data was first stored on computer, continued with improvements in data access, and generated technologies that allow the user to navigate through the data in real time world[12].Searching useful and relevant information from the data ware brought together by the goal of meeting challenges of the previous researches, from various disciples began focusing on developing more efficient and scalable tools which could the field of data mining ,built upon the methodology and algorithm that researches had previously used. Data mining evolution is compared in below table 1.

Table 1 Evolution of Data Mining

Evolutionary Step	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	Computer, tapes, disks	IBM, CDC	Static data delivery
Data Access (1980s)	RDBMS, SQL, ODBC	Oracle, Sybase, Informix, IBM	Dynamic data deliver at record Microsoft level
Data warehouse and Decision Support (1990s)	OLAP, multi-Dimensional databases, data warehouse.	Pilot, Com share, Arbor, Cognos, Micro strategy	Dynamic data deliver at multiple level
Data Mining (still maturing)	Advanced algorithm, Multiprocessor, computers, massive databases	Pilot, Lock head, IBM, SGI, Numerous Start up (nascent industry)	Proactive information delivery

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Figure 2Dataminingasconfluenceofmanydisciplines

Data mining draws on ideas shown abovein figure 2. Data mining task is classified into attributes this attribute is commonly known as “Target” or “Dependent variable” while they said for prediction are known the “Explanatory” or “Independent variables”. There are two task mainly classified as namely Predicted Task – The objective is to predict attribute based on values of the other figure.3. Core data mining task is to derive patterns (Trends, Cluster, Trajectories, Anamolies) summarized underlying relationship in data. They are often exploratory in nature and frequently required post processing techniques to validate and explain result.[12]

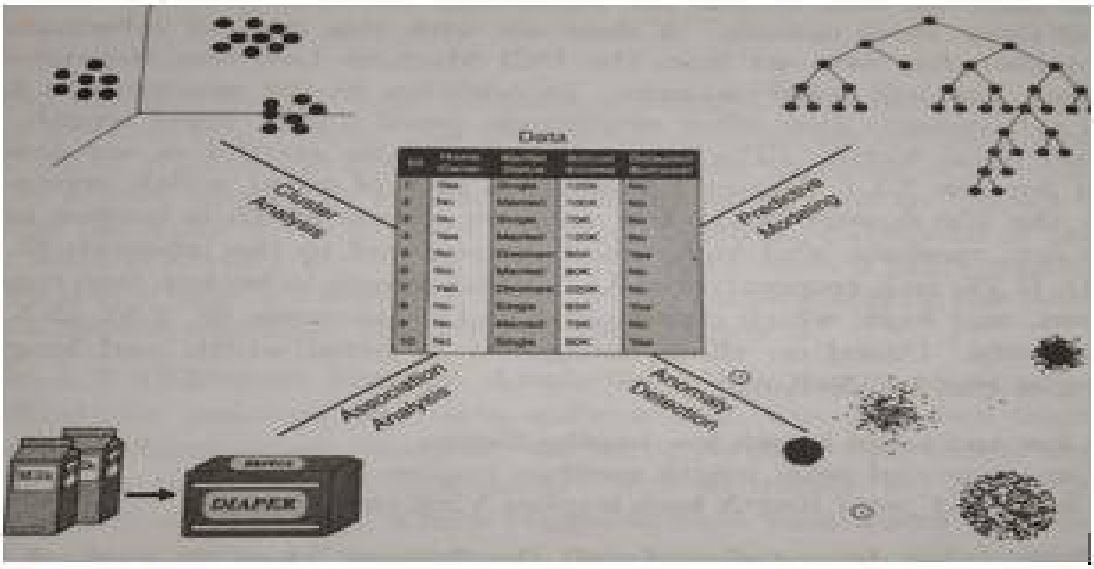

Fig.3. Core data mining task.

Data mining techniques includes association rule mining, finds association relationships among a large set of data item. With massive amount of data continuously being collected and stored, many industries are becoming interested in mining association rules from its database.The classification of large datasets is a classic problem in data mining. It is generally applied in data base with large no of records belong to one of given classes. Clustering is the organization of data in classes when the class label is unknown. In such a way, examples belong to same group which are like each other.Prediction is used for forecasting in a business context or forecasting of missing numerical values or predicting trends in time - related data. Once a classification model has been developed based on a training set, the class label of

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

an object can b for seen based on attribute values on the object. In case-based reasoning, given a new problem, the techniques consider past situation where similar tasks were accomplished successfully. N network is use as a mathematical model for adjusting the weights which is called a learning rule. Outlier Analysis discovers outliers or noise from database. According to Hawking, (1980); Outliers is an observation that derivate so much from other observation as to arouse suspicion that it was generated by different mechanism. Text mining derives high quality information from the text. It applies same types of analysis, such as knowledge discovery and trend analysis to unstructured textual data, this data mining applies to structured data. Spatial mining is growing with the increasing significance of large geo-spatial datasets, such as MAs, repositories of remote sensing images, and decennial census. In temporal mining a sequence is concerned with funding statistically equivalent patterns between data examples where the values are delivered in a sequence. Mainly used in structure data mining. There are two type of sequence mining-[12]. Time-series analysis comprises methods to understand time-series data. This is used to predict future event based on past event. Issues and challenges in Data Mining are Limited information, Noisy and missing data, User Interaction, Level of uncertainty, and Data updates.

Web Mining is the application of data mining techniques to discover patterns from the Web. Mining techniques in web can be categorized into three-man areas. Web Content Mining describes the automatic search of information resources available online and involves mining web data content. Web structure mining is to generate a structural summary about the website and the web page. Based on the topology of the hyperlinks, it will categorize the web pages and generate the information, such as similarity and relationship between different website. Web Usage Mining is application of data mining techniques to discover user navigation patterns from the web.

Big data is a field that treats of ways to analysed, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.[13] Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. Big data was originally associated with three key concepts: volume, variety, and velocity. Other concepts later attributed with big data are veracity [14]. Relational database management systems, desktop statistics [clarification needed] and software packages used to visualize data often have difficulty handling big data. The work may require ”massively parallel software running on tens, hundreds, or even thousands of servers”[15]. What qualifies as being” big data” varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target.” For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration” [16]

2)Big Data-History Trends

The term has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term.[17][18] Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.[19] Big data philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data.[20] Big data ”size” is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many exabytes of data.[21] Big data requires a set of techniques and technologies with new forms of integration to reveal insights from data sets that are diverse, complex, and of a massive scale.[22]

A 2016 definition states that ”Big data represents the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value”.[23] Similarly, Kaplan and Haenlein define big data as ”data sets characterized by huge amounts (volume) of frequently updated data (velocity) in various formats, such as numeric, textual, or images/videos (variety).”[24] Additionally, a new V, veracity, is added by some organizations to describe it,[25] revisionism challenged by some industry authorities.[26] The three Vs (volume, variety and velocity) have been further expanded to other complementary characteristics of big data:[27][28]

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Machine learning: big data often doesn't ask why and simply detects patterns[29] Digital footprint: big data is often a cost-free by-product of digital interaction[28][30][better source needed] A 2018 definition states "Big data is where parallel computing tools are needed to handle data", and notes, "This represents a distinct and clearly defined change in the computer science used, via parallel programming theories, and losses of some of the guarantees and capabilities made by Codd's relational model." [31]

The growing maturity of the concept more starkly delineates the difference between" big data" and" Business Intelligence":[32]

Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends, etc. Big data uses inductive statistics and concepts from nonlinear system identification [27] to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density [33] to reveal relationships and dependencies, or to perform predictions of outcomes and behaviours.[32][34]

3) Big Data-Characteristic

Big data can be described by the following characteristics:[27][28].Volume The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. Variety in the type and nature of the data. This helps people who analysed it to effectively use the resulting insight. Big data draws from text, images, audio, video; plus, it completes missing pieces through data fusion. Thus, Velocity in this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

Big data is often available in real-time which is compared to small data, big data are produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.[35]

It is the extended definition for big data, which refers to the data quality and the data value.[36] The data quality of captured data can vary greatly, affecting the accurate analysis.[25] Data must be processed with advanced tools having analytics and algorithms to reveal meaningful information.

For example, to manage a factory one must consider both visible and invisible issues with various components. Information generation algorithms must detect and address invisible issues such as machine degradation, component wear, etc. on the factoryfloor.[37][38][39]

4) Big Data-6'C Architecture

Big data analytics for manufacturing applications is marketed as a "5C architecture" (connection, conversion, cyber, cognition, and configuration).[40] Factory work and Cyberphysical systems may have an extended "6C system":Hence, the 6C system namely includes-

1) Connection (sensor and networks)
2) Cloud (computing and data on demand) [41][42]
3) Cyber (model and memory)
4) Content/context (meaning and correlation)
5) Community (sharing and collaboration) [41][42]
6) Customization (personalization and value)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

III. APPLICATIONS

In this Proposed paper consist of various data Extraction approaches in this research paper is comprised of main categories: The First category is focus on OCR technique The Second category is focus on Data ware House technique The Finale category is focus on Data Analysis technique which consist of data mining ,Web mining and Big Data Analysis as discussed. Here are application for each techniques discussed below:

A. Application for OCR Technique

1..Used for extracting the data and entering into specified destination for business documents, e.g. check, passport, invoice, bank statement and receipt.
2.Used in Automatic number plate recognition.
3.Used in airports, for passport recognition and information extraction for passengers
4.Due to automatic insurance documents key information can be extracted
5.Can be used in extracting business card information and saving itinto a contact list
6. More quickly make textual versions of printed documents. book scanning for Project Gutenberg
7.Making electronic images of printed documentssearchable such as Google Books.
8.Converting handwriting in real time to control a computer such as pen computing.
9.Defeating CAPTCHA anti-bot systems, though these are specifically designed to prevent OCR. Thus, this purpose can also be usedto test the robustness of CAPTCHA anti-bot systems.
10.Assistive technology for blind and visually impaired users. e is in Medical field in 1974.

B. Application for Data Warehouse Technique

1.Used In the Airline sector, it is used for operation purpose like crew assignment, analyses of route profitability, frequent flyer program promotions, etc.
2.Used in Banking sector, It is widely used in the banking sector to manage the resources available on desk effectively. Few banks also used for the market research, performance analysis of the product and operations.
3.Used in Healthcare sector, also used forData warehouse to strategize and predict outcomes, generate patient's treatment reports, share data with tie-in insurance companies, medical aid services, etc.
4.Can be used in Public sector where data warehouse is used for intelligence gathering. It helps government agencies to maintain and analysed tax records, health policy records, for every individual.
5.Used in Investment and Insurance sector,in this sector, the warehouses are primarily used to analysed data patterns, customer trends, and to track market movements.
6.Used inRetail chains where the Data warehouse is widely used for distribution and marketing. It also helps to track items, customer buying pattern, promotions and used for determining pricing policy.
7.Used for Telecommunication Sector where the data warehouse is used in this sector for product promotions, sales decisions and to make distribution decisions.
8.Used in Hospitality Industry utilizes warehouse services to design as well as estimate their advertising and promotion campaigns where they want to target clients based on their feedback and travel patterns.

C. Application for Data Analysis Technique

1. Application for Data Mining Technique-

Application of data mining in the business area include customer segmentation, market basket analysis, risk management, fraud detection, delinquency tracking, demand prediction, direct marketing and promotion of products, and sales forecasting.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

2. Application for BIG Data Technique-

1. Big Data in Education industry-

A. Customized and dynamic learning programs: Customized programs and schemes for everyone can be created using the data collected on the bases of a student's learning history to benefit all students. This improves the overall student results

B. Reframing course material: Reframing the course material according to the data that is collected based on what student learns and to what extent by real time monitoring of what components of a course are easier to understand.

C. Grading Systems: New advancements in grading systems have been introduced as a result of proper analysis of student data.

D. Career prediction: Proper analysis and study of every students' records will help in understanding the students' progress, strengths, weaknesses, interests and more. It will help in determining which career would be most appropriate for the student in the future. The applications of big data have provided a solution to one of the biggest pitfalls in the education system, that is, the one-size-fits-all fashion of academic set up, by contributing in e-learning solutions.

E. Example: The University of Alabama has more than 38000 students and an ocean of data. In the past when there were no real solutions to analysed that much data, some of that data seemed useless. Now administrators can use analytics and data visualizations for this data to draw out patters with students revolutionizing the university's operations, recruitment and retention efforts.

2.Big data in Healthcare industry-

Now healthcare is yet another industry which is bound to generate a huge amount of data. Following are some of the ways in which big data has contributed to healthcare

A. Big data reduces costs of treatment since there is less chances of having to perform unnecessary diagnosis.

B. It helps in predicting outbreaks of epidemics and helps in deciding what preventive measures could be taken to minimize the effects of the same.

C. It helps avoid preventable diseases by detecting diseases in early stages and prevents it from getting any worse which in turn makes the treatment easy and effective.

D. Patients can be provided with the evidence-based medicine which is identified and prescribed after doing the research of past medical results.

E. Wearable devices and sensors have been introduced in healthcare industry which can provide real time feed to the electronic health record of a Patient.

F. Example-One such technology is from Apple. Apple has come up with what they call Apple Health Kit, Care Kit and Research Kit. The main goal is to empower the iPhone users to store and access their real time health records on their phones.

3.Big data in Government industry –

A. Governments, be it of any country, come face to face with a very huge amount of data on almost daily basis. Reason being, they must keep track of various records and databases regarding the citizens, their growth, energy resources, geographical surveys and many more. All this data contributes to big data.

B. The proper study and analysis of this data helps the Governments in endless ways.

Few of them are:

1.Welfare schemes: In making faster and informed decisions regarding various political programs.

To identify the areas that are in immediate need of attention.

To stay up to date in the field of agriculture by keeping track of all the land and livestock that exists.

To overcome national challenges such as unemployment, terrorism, energy resource exploration and more.

2.Cyber security: Big Data is hugely used for deceit recognition Governments are also finding the use of big data in catching tax evaders.

C. Example: The Food and Drug Administration (FDA) which runs under the jurisdiction of the Federal Government of US leverages from the analysis of big data to discover patters and associations in order to identify and examine the expected or unexpected occurrences of food-based infections.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

4.Big Data in Media and Entertainment industry

A.With people having access to various digital gadgets the generation of large amount of data is inevitable and this is main cause of rise in big data in media and entertainment industry. Other than this, social media platforms are also another way in which huge amount of data is being generated.

B.Although business in media and entertainment industry have realized the importance of this data and they have been able to leverage from it to help their businesses grow.

C.Some of the benefits extracted from the big data in media and entertainment industry:

1. Predicting the interests of audiences. Optimized or on-demand scheduling of media streams in digital media distribution platforms.

2. Getting Insights into customers reviews and pinpointing their animosities. Effective targeting of the advertisements for media

D. Example: Spotify, which is an OnDemand music providing platform, uses big data analytics and collects data from all the users around the globe and then uses the analysed data to give informed music recommendations and suggestions to every individual user.

IV. CONCLUSION

Thus, In this paper we have studied various types of Data extraction techniques which is comparatively analysed on application based. Here, in this paper it has described techniques that, they have their own recalls and precise rates. Thus, due to this proposed paper we get to known about their original, characteristics and Working models of their version. A comprehensive review of this taxonomy would turn to be a great opportunity to predefine or cultivate a fixed formatted structure in upcoming application exploration of state new relevant and precise data extraction approaches. And, therefore, we assumed that this paper gives us basic conceptual idea of Data and their types. Their roles do it plays in today's worlds. And, also, we have Understood various Extraction techniques which are mention above in this paper.

REFERENCES

[1] Study of data analysis model based on big data technology Jinhua Chen ; Qin Jiang; Yuxin Wang; Jing Tang 2016 IEEE International Conference on Big Data Analysis (ICBDA)

[2] Azam Beg; Faheem Ahmed; Piers Campbell 2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks

[3] Kai Wang; JianmingJin; Qingren Wang 2009 10th International Conference on Document Analysis and Recognition

[4] Mohd Amir Bin MohdAzir, Kamsuriah Binti Ahmad978-1-5386-04755/17copyright 2017 IEEE

[5] " Miss N.V. Kamanwar"," Prof.S.G. Kale",2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (WCFTR'15),978-1-4673-9214-3/16 copyright 2016 IEEE

[6] Schantz, Herbert F. (1982). The history of OCR, optical character recognition. [Manchester Centre, Vt.]: Recognition Technologies Users Association. ISBN 9780943072012.

[7] d'Albe, E. E. F. (1 July 1914)." On a Type-Reading Optophone". Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 90(619): 373375. Bibcode:1914RSPSA. 90..373D.doi:10.1098/rspa.1914.0061.

[8] " Extracting text from images using OCR on Android". 27 June 2015.

[9] " [Tutorial] OCR on Google Glass". 23 October 2014.

[10] Tappert, C. C.; Suen, C. Y.; Wakahara, T. (1990)." The state of the art in online handwriting recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence. 12 (8): 787. doi:10.1109/34. 57669..

[11] " Data ware Housing", author -" REEMA THAREJA", ISBN-13:978-019-569961-6" ISBN-10:0-19-569961-0".

[12] " Introduction to Data Mining", author-" Pang-Ning tan", "Michael steinbach," Vipin Kumar", ISBN-978-81-317-1472-0". (PEARSON)

[13] Breur, Tom (July 2016)." Statistical Power Analysis and the contemporary" crisis" in social sciences". Journal of Marketing Analytics. 4 (23): 6165. doi:10.1057/s41270-016-0001-3. ISSN 2050-3318.

[14] Jacobs, A. (6 July 2009)." The Pathologies of Big Data". ACMQueue.

[15] Magoulas, Roger; Lorica, Ben (February 2009)." Introduction to Big Data". Release 2.0. Sebastopol CA: O'Reilly Media (11).

[16] John R. Mashey (25 April 1998)." Big Data ... and the Next Wave of InfraStress" (PDF). Slides from invited talk. Usenix. Retrieved 28 September 2016.

[17] Steve Lohr (1 February 2013)." The Origins of 'Big Data': An Etymological Detective Story". The New York Times. Retrieved 28 September 2016.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

[18] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. 7: 15.

[19] Dedi, N.; Stanier, C. (2017)." Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery". 285. Berlin; Heidelberg: Springer International Publishing. ISSN 1865-1356. OCLC 909580101.

[20] Everts, Sarah (2016)." Information Overload". Distillations. Vol. 2 no. 2. pp. 2633. Retrieved 22 March 2018.

[21] brahim; Targio Hashem, Abaker; Yaqoob, Ibrar; BadrulAnuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015)." big data" on cloud computing: Review and open research issues". Information Systems. 47: 98115. doi:10.1016/j.is.2014.07.006.

[22] De Mauro, Andrea; Greco, Marco; Grimaldi, Michele (2016)." A Formal definition of Big Data based on its essential Features". Library Review. 65: 122135. doi:10.1108/LR-06-2015-0061

[23] Kaplan Andreas; Michael Haenlein (2018) Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence, Business Horizons, 62(1) [24]" What is Big Data?". Villanova University.

[25] Grimes, Seth." Big Data: Avoid 'Wanna V' Confusion". InformationWeek. Retrieved 5 January 2016.

[26] Hilbert, Martin." Big Data for Development: A Review of Promises and Challenges. Development Policy Review". martinhilbert.net. Retrieved 7 October 2015.

[27] DTSC 7-3: What is Big Data? YouTube. 12 August 2015.

[28] Mayer-Schnberger, V., Cukier, K. (2013). Big data: a revolution that will transform how we live, work and think. London: John Murray.

[29] " Digital Technology Social Change". Canvas.instructure.com. Retrieved 8 October 2017.

[30] Fox, Charles (2018-03-25). Data Science for Transport. Springer.

[31] " avec focalisation sur Big Data Analytique" (PDF). Bigdataparis.com. Retrieved 8 October 2017.

[32] Billings S.A." Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains". Wiley, 2013

[33] " le Blog ANDSI DSI Big Data". Andsi.fr. Retrieved 8 October 2017.

[34] Les Echos (3 April 2013)." Les Echos Big Data car Low-Density Data? La faible densiten information commefacteur discriminant Archives". Lesechos.fr. Retrieved 8 October 2017.

[35] Kitchin, Rob; McArdle, Gavin (17 February 2016)." What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets". Big Data Society. 3 (1): 205395171663113. doi:10.1177/2053951716631130.

[36] Onay, Ceylan; ztrk, Elif (2018)." A review of credit scoring research in the age of Big Data". Journal of Financial Regulation and Compliance. 26 (3): 382405. doi:10.1108/JFRC-06-2017-0054.

[37] Big Data's Fourth V

[38] Lee, Jay; Bagheri, Behrad; Kao, Hung-An (2014)." Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial Informatics". IEEE Int. Conference on Industrial Informatics (INDIN) 2014.

[39] Lee, Jay; Lapira, Edzel; Bagheri, Behrad; Kao, Hung-an." Recent advances and trends in predictive manufacturing systems in big data environment". Manufacturing Letters. 1 (1): 3841. doi: 10.1016/j.mfglet.2013.09.005.

[40] . Imscenter.net. Retrieved 16 June 2016

[41] Wu, D., Liu. X., Hebert, S., Gentzsch, W., Terpenny, J. (2015). Performance Evaluation of Cloud-Based High-Performance Computing for Finite Element Analysis. Proceedings of the ASME 2015 International Design Engineering Technical Conference Computers and Information in Engineering Conference (IDETC/CIE2015), Boston, Massachusetts, U.S.

[42] Wu, D.; Rosen, D.W.; Wang, L.; Schaefer, D. (2015)." Cloud-Based Design and Manufacturing: A New Paradigm in Digital Manufacturing and Design Innovation". Computer-Aided Design. 59 (1): 114. doi: 10.1016/j.cad.2014.07.006.