ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Unstructured Data Analysis-A Survey

K.V.Kanimozhi[1], Dr.M.Venkatesan[2]

Assistant professor, School of Computing science and Technology, Vellore Institute of Technology University, Tamilnadu, India[1]

Associate professor, School of Computing science and Technology, Vellore Institute of Technology University, Tamilnadu, India[2]

Abstract: Recent years have observed the ability to gather a massive amount of data in a large number of domains. As the data is collected in unprecedented rate, the analysis, rather than the storage of this data becomes a challenge. According to the IDC estimation 90% of data is unstructured data which is a fastest growing data whereas the remaining is the structured data, unstructured data refers to information that either does not have predefined data model or does not fit into relational database for information access. This unstructured data are being continuously comes from various sources like satellite images, sensor readings, email messages, social media, web logs, survey results, audio, videos etc. Due to the large volume of unstructured data there is a big challenge for all the industry currently to analyse and extract a meaningful value from it. Traditional methods are adequate for analysis of structured data but these methods are not appropriate for large volume of unstructured data in order to extract knowledge.
This paper presents the summary about unstructured data analysis for the beginners or the people from academia who is interested in analysis of unstructured data to extract the knowledge to improve the business processes and performance.

Keywords: Unstructured data, structured data, data mining

I. INTRODUCTION

Due to the fast growth of internet there has been big volume of information is produced and shared by various administrations in nearly every business, industry and other fields. Due to this high explosion it's really a big challenge to store, manage and access knowledge from this. Research has shown that over 95% of the digital universe is unstructured data. According to these studies, 80% of all stored organizational data is unstructured (1,2).Unstructured data does not have any data structure(i.e. exists within a database).Examples of unstructured data may include

(i) Textual: documents, presentations, spread sheets, scanned images, etc.
(ii) Imagery: multimedia files, streaming video, etc.
(ii) HUMINT: reports, audio files , and gestures
(iii) Sensors: seismic, acoustic, magnetic, sonar, etc. and
(iv) Environmental: weather (3).

This presents a critical challenge for large data technologies specifically in the area of data exchange because unstructured data must be structured before knowledge can be extracted (3).

II. BACKGROUND STUDY ON UNSTRUCTURED DATA

Today knowledge is the biggest asset of all companies so maximum of the knowledge is recorded in unstructured format. Unstructured data is a significant part of the Big Data explosion, and 91% of the respondents in the survey say they are aware of unstructured datafiles within their enterprise systems.More than one-fourth of the survey respondents now say that the majority oftheir enterprise data is unstructured. Unstructured data in their enterprises will surpass structured data within next three years by Industry group (5).

A. Structured data:

Structured data is a data included in relational database system. Examples are database tables, objects, tags, reports, indexes etc. Being structured and highly organized it can be managed by SQL and its multiple variations developed by IBM, ADO.net, ODBC and many RDBMS support, Due to explicit semantics and structure efficient search is possible for focused content by simple and straightforward search engines.

B. Semi structured data:

Semi structured data is one type of structured data but lacks the data model structure or do not conform a formal or rigid structure. This semi structured data do not require a schema definition it is rather optional and contains tags or other markers to separate semantic elements and enforce hierarchies of records fields within the data .Semi-structured data is increasingly occurring since full-text documents and databases are not the only forms of data on the Internet, and different applications need a medium for exchanging information (4). Language like XML or other mark-up language, Java script object notation (JSON) is used to manage semi structured data. To convert the semi structured data to structured data traditional data mining techniques or natural processing language is used.

C. Unstructured data:

Unstructured data is data comes from machines generated or human generated and it is broadly classified into two types

i) Non-Textual unstructured data is a multimedia data like still images, videos, and MP3 audio files
ii) Textual unstructured data examples are like email messages, collaborative software and instant messages,

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

memos, word processor documents, PowerPoint presentations etc.

And the different standards for unstructured data are open XML, SMTP, SMS, CSV and Information and content exchange.

iii) Unstructured data- Literature Survey:

There are different methods to address this problem space of unstructured analyticsThis paper explores the some of the different techniques and paradigms are being developed and implemented for unstructured data analytics.

Dr.Goutam Chakra borty [7] et.al in 2014 proposed an outlook at how to organize and analyse textual data for extracting insightful customer intelligence from a large collection of document and for using such information to improve business operations and performance. Using SAS text miner and SAS sentiment analysis studio artificial neural network regression model is used for variable selection to predict the target variable.

Yuanming Huang [8] et.al. in 2010 proposed the theory and methods on massive unstructured audio video intelligent information process in emergency system mechanisms like visual computing, cognitive modelling, cognitive science will be the latest achievement of the method of mathematical analysis for unstructured multimedia data in emergency system. This research results can build mass information intelligent service platform technical support system framework, breaking the mass of information intelligence servicesin a number of technical bottleneck.

KapilBakshi [13] in 2012 describes about modern approach solutions to solve unstructured data by various building blocks and techniques for Map Reduce and HDFS, HBase and their implementation in an open source Hadoop framework. This paper focuses on the infrastructure planning (compute, network, and storage systems), and reviews Hadoop design criteria and implementation considerations. Hadoop includes many technologies, including Map Reduce, which interacts with the infrastructure elements while analysing data.

Jaein Kim et.al [14] in 2013 proposed novel framework for the real-time analysis of unstructured big data such as video, image, sounds and text. RUBA framework analyses the big data using CEP engine and uses CQL to modify the analysis conditions in real-time without re-executions of system. In addition, RUBA framework provides functions to manage several distributed analysis systems using the method of CQL management easily. Implementation of Object Monitoring System applied RUBA framework confirms the availabilities of proposed framework through real-time data analysis and modifying of analysis conditions.

Jochen De Weerdt et.al [9] in 2012 presents a solution strategy to leverage traditional process discovery techniques in the flexible environment of incident management processes. In such environments, it is typically observed that single model discovery techniques are incapable of dealing with the large number of different types of execution traces. Accordingly, they propose a combination of trace clustering and text mining to enhance process discovery techniques with the purpose of retrieving more useful insights from process data.

Hsin-Ying Wu [11] et.al in 2014 proposed an analytical process to interpret the dialogue between young entrepreneurs and their audience of Facebook Pages and collected consumer feedback from social networks, like FB. The interpretation of the dialogues into meaningful statistics, especially when attempting to model, cluster, and analyse the critical elements of posted Internet content, requires new text analysis techniques and methodologies. CKIP (Chinese Knowledge and Information Processing) is applied to extract the key phrases from the Chinese language dialogues. Then clustering is used to generate the critical points that customers care about and then to explore key factors that attracts customers and resolves their needs.

S.Geetha [18] et.al in 2012 proposed an approach to extract the data from unstructured data by organizing into structured way in the form of Data Relations. Set of rules are used to interpret domain knowledge from the unstructured data. By segmenting the data using POS and its syntactic structure present in the input unstructured text, which will help us to categorize the data into entities, actions and construct the relations among these entities and actions. This approach applied in the "News Retrieval System" which collected news from Various Pages and processed on the basis of Page ranking and displayed on the Single page in an effective way.

Dr.Muhammad Shahbaz [19] et.al in 2014 proposed solution in this work is the development of a System (Sentiment Miner). It will provide features to process and classify text files (reviews and appraisals) for opinion mining at sentence level using Natural language Processing techniques and Opinion Mining algorithms. The prototype of a final product; a Semantic Search Enginewill facilitate in document retrieval for analysis whenever required.

III. TOOLS/TECHNIQUES TO HANDLE UNSTRUCTURED DATA

The different techniques used to search analyse and deliver unstructured data (5) are

a) Content management system
b) Relational Database
c) Data Mining
d) Text Analytics. Federal search or enterprise search data base
e) Non-relational database
f) Real time data visualization tools
g) E-discovery application

The new technologies for unstructured data era (5) are

a) Log monitoring and reporting tools
b) In-memory databases
c) NOSQL databases
d) Hadoop
e) MPP data warehouses.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

These technologies bring high value information in real time instead waiting to store and perform operations like traditional methods.

IV. DISCUSSION

According to our view most of the literature papers methods go fine for small amount of data only by categorizing textual and non-textual unstructured type but not suitable for real time environment. And very few papers analyse streaming data called unstructured both textual and non-textual data in motion, With the relentless growth of unstructured data, as most of the technique consume today, there are so many technologies are emerging continuously still real-world issues for decision making and gaining knowledge is a great challenge. The main future challenges of unstructured data on big data environment are

a) Optimal architecture for large scale data analysis(16) ,
b) Distributed Mining(16),
c) Visualization(16)
d) Useful data gets lost due to large new flow of unstructured data(16)
e) Storage and Management Services
f) Security and other technical challenges.

V. CONCLUSION

Since all the enterprises like manufacturing, IT Tools and Services, Information Services, Government, Education, non-profit organization, Financial Insurance works with majority of data like unstructured data which has been increasing significantly every day and also it is big business asset and not liability, Hence analysis of unstructured big data both textual and non-textual together called streaming or unstructured data in motion is a really recent challenge .This paper suggests the answer that by providing awareness and education and also by establishing good governance policies and alternative technologies we can resolve this difficulty in easier way which is very important about the unstructured large scale data . These innovative approaches to computation make analytics possible and make them affordable.

REFERENCES

[1] J. Gantzandetal. The Expanding Digital Universe: A Forecast of World Wide Information Growth through 2010. 2007 mar; IDC Whitepaper
[2] C. White. Consolidating, Accessing, and Analysing Unstructured Data. 2005 Dec. Business Intelligence Network article. Powell Media. LLC
[3] Erik P. Blasch, Stephen Russell, Guna Seetharaman.Joint Data Management for MOVINTData-to-Decision Making, ISIF 11.14th International Conference on Information Fusion; 2011 July 5-8; Chicago, Illinois: USA; 2011. 978-0-9824438-3-5 ©2011 ISIF
[4] M. Sukanya, S. Biruntha.Techniques on Text Mining. ICACCCT 12.IEEE International Conference on Advanced Communication Control and Computing Technologies; 2012.p.269-271
[5] J. McKendrick. Survey on unstructured data, produced by Unisphere Research 2011; Available from: http://www.ciosummits.com/media/pdf/solution_spotlight/ Marklogic_ 20 11- survey.pdf, 2011.
[6] Motoi Iwashita, Ken Nishimatsu, Shinsuke Shimogawa. Semantic Analysis Method for Unstructured Data in Telecom Services. IEEE13. IEEE 13th International Conference on Data Mining Workshops;p.789-795.doi:10.1109/ICDMW.2008.79
[7] Dr. Goutam Chakra borty,Murali Krishna Pagolu. Text Analytics and Sentiment Analysis of Unstructured Data. Applications of Mining.
[8] Yuanming Huang, Yujie Zheng. Research on Theory and Method on massive Audio Video Unstructured Information Intelligent Process in Emergency System. 978-1-4244-6928- 4/10/$26.00©2010 IEEE.
[9] Jochen De Weerdt, Seppe K.L.M. vanden Broucke, Jan Vanthienen, and Bart Baesens. Leveraging Process Discovery with Trace Clustering and Text Mining for Intelligent Analysis of Incident Management Processes. WCCI 12. IEEE World Congress Computational Intelligence; 2012 June 10-15; Brisbane. Australia; 2012
[10] Quan Liu, Yongjun Peng.A Method of Unstructured Information Process in Computer Teaching Evaluation System Based on Data Mining Technology.IEEE 13 .International Conference on Communication Systems and Network Technologies; 978-0-7695-4958-3/13 $26.00 © 2013 IEEEDOI 10.1109/CSNT.2013.147
[11] Hsin-Ying Wu, Kuan-Liang Liu, Charles Trappey. Understanding Customers Using Facebook Pages: Data Mining Users Feedback Using Text Analysis. IEEE [12].Proceedings of the 2014 IEEE 18th International Conference on Computer Support Cooperative Work in Design; 978-1-4799-3776-9/14/$31.00 ©2014 IEEE.
[13] Dr. Muhammad Shahbaz, Dr. Aziz Guergachi, Rana Tanzeel ur Rehman.Sentiment Miner: A Prototype for Sentiment Analysis of Unstructured Data and Text. 978-1-4799-3010-9/14/$31.00 ©2014 IEEE.
[14] Kapil Bakshi.Considerations for Big Data: Architecture and Approach.978-1-4577-0557- 1/12/$26.00 ©2012 IEEE
[15] Jaein Kim, Nacwoo Kim, Byungtak Lee., Joonho Park, Kwangik Seo, Hunyoung Park.RUBA: Real-time Unstructured Big Data Analysis Framework. 978-1-4799-0698-7/13/$31.00 ©2013 IEEE
[16] Erik Cambria ,Björn Schuller , Yunqing Xia, Catherine Havasi. New Avenues in Opinion Mining and Sentiment Analysis. 1541-1672/13/$31.00 © 2013 IEEE
[17] Wei Fan, Albert Bifet. Mining Big Data: Current Status, and Forecast to the Future.
[18] Luís Filipe da Cruz Nassif, Eduardo Raul Hruschka, Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection. IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, 2013 Jan; 8(1):46-54
[19] S. Geetha, Dr. G.S. Anandha Mala, Effectual Extraction of Data Relations from unstructured Data. Seiscon 2012.Third International Conference on Sustainable Energy and Intelligent System; Tiruchengode, Tamilnadu, India on 27-29 December, 2012.
[19] Dr. Muhammad Shahbaz, Dr. Aziz Guergachi, Rana Tanzeel ur Rehman.Sentiment Miner: A Prototype for Sentiment Analysis of Unstructured Data and Text. 978-1-4799-3010-9/14/$31.00 ©2014 IEEE