Data Mining from the Big Data

Satyajee Srivastava*

Abstract

We are already in the era of Exa Bytes where data comes with different V's (Volume, Velocity, Variety etc. [12]. Data sources are decentralized and diverse. Data is dynamic and having complex relationships. Existing data mining tools and technologies are not very effective and have some limitations too. Earlier data stored in Warehouses have some schema standard model which leads to effective and efficient information mining. Now, information industry needs to develop a efficient and reliable NoSQL technology that is able to handle the unstructured and dynamic data. This paper summarizes the methodologies that are already developed, and the key challenges and the privacy issues.

Keywords: Mining,Unstructured, IoT, Data Mining, Complex, Heterogeneity

Introduction

In year 1997, Big Data appeared for the first time in the paper "How much information is there in the world?" published by MichaelLesk. In this paper,author concluded that in future there may be a few thousands petabytes of information that will exist and storage devices such as disks and tapes will reach that level in year 2000.So, after some year we will be able to store all information, nothing will be thrown out [2]. But, we are in the era of Big Data which comes with different V's (velocity, volume, and Variety).After a year, In April 1998, another author John R.Masey,who is the main scientist at SGI(Soka Gakkai International), published a paper entitled "Big Data and the Next Wave of InfraStress". In this paper the term Big Data was used for first time and author talked about the Big Data Technology and Infrastructure Stress [1].

Big Data, Internet of Things (IOT) and web are related to each other. Smart Things means programmed and connected to internet which produces a lot of data every second in various formats such as text files, image files audio files video files etc. Unlimited and unstructured data is the side effect of dynamic web.2006 is the year of web (formally web 2.0) [3]. This innovation comes with very big hope in digital world and it is big innovation as this was the year of web. This is the revolution in the internet world Web and dynamic means where user can comment (e.g. facebook),edit web pages(e.g. Wikipedia), upload and download videos (e.g. youtube) these are the innovation at that time. In 2006, Web 3.0 come in existence , Web 3.0 or semantic web hope to minify human's tasks and given to machines. This includes the semantic technologies [4]. Simultaneously, these innovation create a side effect in the form of Big Data or GIT[13].

Table 1: Internet User Worldwide

Year(First July)	*Internet Users*	*Penetration(%of population with internet)*
2014	2,925,249,355	40.4%
2013	2,712,239,573	37.9%
2012	2,511,615,523	35.5%
2011	2,272,463,038	32.5%
2010	2,034,259,368	29.4%
2009	1,752,333,178	25.6%
2008	1,562,067,594	23.1%
2007	1,373,040,542	20.6%
2006	1,157,500,065	17.6%
2005	1,029,717,906	15.8%
2004	910,060,180	14.1%
2003	778,555,680	12.2%
2002	662,663,600	10.6%
2001	500,609,240	8.1%
2000	413,425,190	6.7%
1999	280,866,670	4.6%

* Assistant Professor, CSE, Galgotia University, Greater Noida, Uttar Pradesh, India. Email: drsatyajee@gmail.com

Year(First July)	Internet Users	Penetration(%of population with internet)
1998	188,023,930	3.1%
1997	120,758,310	2.0%
1996	77,433,860	1.3%
1995	44,838,900	0.8%
1994	25,454,590	0.4%
1993	14,161,570	0.3%

Internet of Things (IoT) and web create a ocean of unstructured data because people and things all are loosely connected. From this ocean of data, information mining (retrieval) is not easy due to the speed and variety of data. Also the speed of Internet users is increasing day by day as shown in the Table 1. In the further sections, some Big Data Mining Techniques, challenges and possible solutions will be discussed.

Rises in the Big Data

In a day, approximately 2.5 quintillion bytes of digital data is produced. 90% of the data present in today's world has been created in the previous two years alone. This data comes from audios, speeches, text files, videos, social networking sites and sensors that are used to gather climate information, images, transaction logs, blogs and cell phone GPS signals etc. This data is big data which is born online [6].

Types of Data

Digital data can be categorized in three types:

Structured Data comes from data bases that have well defined models. This type of data can be represented in the form of tables consisting rows or tuples and columns or fields. Traditional data bases technology such as RDBMS etc are the instances. Sources of structure data are Databases such as Oracle, DB2, Teradata, MySQL, Postgre SQL etc., Spreadsheets, OLTP Systems [8].

Semi-Structured Data is basically a combination of both Structured and Semi-structured data. This kind of data exists in the form of XML (eXtensible Markup Language) Files, Other Markup Language such as HTML, ISON (Java Script Object Notation)[8].

Unstructured Data which occupy a very large portion of the digital data distribution chart creates problem these days. There is no predefined model for this data. Text file, web logs, machine logs, Web Pages, Images, Audios, Videos, Body of Email, Text messages, Charts, Social media data, Word Document are the names of few [8].

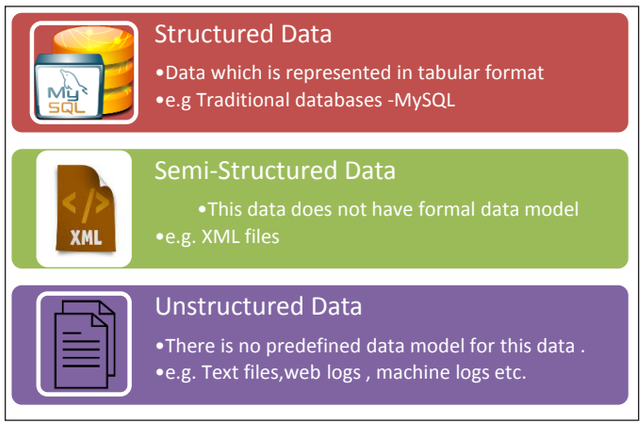

Figure 1: Types of Data

Current Data Distribution

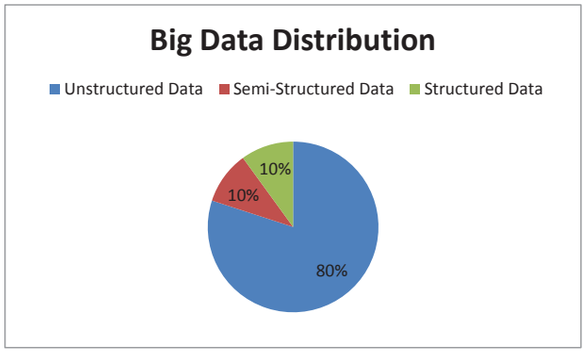

Figure 2: Data Distribution

As the Figure 2 clearly explains distribution of digital data and shows that 80 % of the data is in unstructured form. Only 20% of the digital data is present in rest of two types. So, mining the information from this 80% data is quite difficult due to large amount of data that changes and comes very fast.

Unstructured Knowledge Mining Techniques

Knowledge mining from the large amount of unstructured data requires an attention of researchers, so that, this data can give actionable insights. Traditional information mining technique such as Relational Data Base

Management System (RDBMS) are not much flexible to convert unstructured data to structured format, so new techniques are introduced such as NoSQL, Hadoop, Spark are name of some tools used in Big data Manipulative. Following table represents a brief summary on information extraction, topics, terms, and clustering.

This table contains methodologies that are discussed below:

Information Extraction

Information talks about the KDD process where the unstructured data sources are used. The KDD process contains the series of activities i.e. search through data warehouses, databases, and other repositories and further data cleaning and data integration procedures that are used to remove outliers or unnecessary information. Then this clean data stored in a single repository where data mining tools are used for pattern evaluation to perform knowledge based activities. Existed KDD system have a schema so data cleaning and integration was straight used but current sources of data are unstructured so the information extraction is required to convert unstructured data into semi-structured or structured form before the further processes such as data cleaning and integration process [9].

Topics

Topics are used where keywords extracted on the basis of users' subscription. In online systems, such as hotel booking news articles flight etc. topics methodologies are used. A mechanism based on these related keywords (known as Lexical Chaining) is applied to extract messages that are subsequently published [9].

Terms

Terms are used to build a network of associative relationships among features. For example, artifact are considered as features. The basic advantage of term mining is to optimize the search space vector due to processing time reduced instead of going through the whole document that is required for topics methodology [9].

Document Clustering

Document clustering is a process of further categorizing the documents that are closely related to each other. This clustering process needs a degree of similarity among either the terms or topics or documents. Clustering primarily divided into two types:

(i) K-means clustering and

(ii) Hierarchical clustering. The main problem with K-means clustering is it makes poor clusters. Thus the hierarchical clustering concept is used to over-come the problem of K-means methodology. [9]

So, there are various methodologies such as : Association Rule Mining, Information Retrieval algorithms based on templates, Term Crawling, Document Summarization ,Topic Tracking and Topic Maps, Helmholtz Search Principle, Document Clustering, Re-Usual Dictionaries etc.[10].

Another methodology is also discussed in January 2014 named as HACE theorem. HACE theorem means Very large volume with **h**eterogeneous and **a**utonomous sources and distributed control that needs to explore **c**omplex and **e**volving relationships between the data [11]. In this methodology, a framework is prepared with three tiers as first is Big Data Mining platform that explain the data accessing and computing techniques(central one), second tier is Data Privacy and details domain knowledge(middle one) , algorithms related to Big Data mining (outer most).

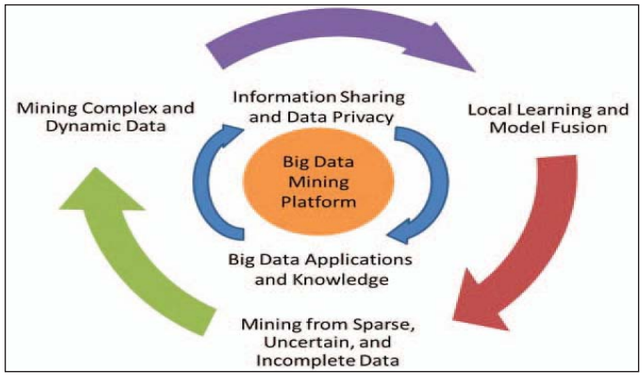

Figure 3: Big Data Processing Framework [11]

Challenge with this framework is associated with every tier. Inner most Tier concentrates on low level accessing and processing data. Second Tier focuses on the challenges such as privacy and information sharing, and application

knowledge of Big Data, the domain which concentrates on high – level semantics, user's privacy problems and domain knowledge of application. The outer most Tier (third) Includes the challenges on actual data mining algorithms. Tier III concentrates on algorithm designs to handle the problems raised by large volume, decentralized distribution of data. Complex and dynamic nature of data. This Tier constitutes the three stages. First contains incomplete, heterogeneous, uncertain, sparse, multisource data are preprocessed by data fusion techniques. Second concentrates on mining of complex and dynamic WEEE [12] data will be performed after preprocessing. From Third local learning global knowledge is obtained and then model fusion is tested and relevant information is given feedback to the preprocessing stage. After that the model and parameters are adjusted based on this feedback [11].

Conclusion

Predefined model and technology slowly go forward to obsolete because data is no more in structured format so requires new enhance models and tools to analyze data to extract information so that knowledge can be obtained and actionable insights can be taken in account. Big Data is large in volume, fast in velocity, various in variety, heterogeneous and diverse in data sources, complex in relationships. Every methodology discussed in this paper has some limitations and problems. For supporting Big Data Mining, we need high performance computing platforms. For this, a system may be designed carefully so that unstructured data can be able to find and solve complex relationships to make a useful pattern, and this further help to make legitimate patterns to know and predict the trends and future.

REFERENCES

Mashey, J. R. (1998). Big Data and the Next Wave of Infra Stress.

Gil Press. (2013). A Very Short History of Big Data. Retrieved from http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

Anderson, P. (2007). *What is Web 2.0? Ideas, technologies and implications for education.* JISC Technology and Standards Watch, Feb.

Aghaei, S., Nematbakhsh, M. A., & Farsani, H. K. (2012). Evolution of the world wide web: From Web 1.0 TO WEB 4.0. *International Journal of Web & Semantic Technology (IJWesT)* 3(1).

Anonymous *Internet Live Stats* (elaboration of databy *International Telecommunication Union (ITU) and United Nations Population Division). Retrieved from* http://www.internetlivestats.com/internet-users/#definitions.

"IBM What Is Big Data: Bring Big Data to the Enterprise. Retrieved from http:// www-01.ibm.com/software/data/bigdata/, IBM, 2012.

Acharya, S., & Chellappan, S. (2015). Willey big data and analytics. ISBN : 978-81-265-5478-2.

Simlilearn. What is Big Data /Hadoop Tutorial/Big Data and Hadoop Training. (2015) Retrieved from https://www.youtube.com/watch?v=CKLzDWMsQGM

Lomotey, R. K., & Deters, R. (2013). RSenter: Tool for Topics and Terms Extraction from Unstructured Data Debris, Proc. of the 2013 IEEE International Congress on Big Data (BigData Congress 2013), (pp:395-402), Santa Clara, California.

Lomotey, R. K., & Deters, R. (2014). Towards Knowledge Discovery in Big Data, 2014 IEEE 8th International Symposium on Service Oriented System Engineering.

Wu, X. (2014). Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE, Data Mining with Big Data, IEEE Transactions on Knowledge and Data Engineering, 26(1).

Laney, D. (2001) *3-d data management: controlling data volume, velocity and variety. META* Group Research Note, 6 February.

Srivastava, S., & Srivastava, R. (2012). *Significance of reverse logistics system to control e-waste,* 2(1).

Srivastava, S., & Srivastava, R. (2012). Adoption of Green Information Technology (GIT) In India-A Current Scenerio. *Journal of Information and Operations Management,* 3.1(61).