Big Data Mining Tools for Unstructured Data: A Review

YOGESH S. KALAMBE
M.Tech-Student
Computer Science and Engineering,
R V College of Engineering,
Bangalore, Karnataka, India.

D. PRATIBA
Assistant Professor,
Department of Computer Science and Engineering,
R V College of Engineering,
Bangalore, Karrnataka, India.

Dr. PRITAM SHAH
Associate Professor,
Department of Computer Science and Engineering,
R V College of Engineering,
Bangalore, Karnataka, India.

Abstract— **Big data is a buzzword that is used for a large size data which includes structured data, semi-structured data and unstructured data. The size of big data is so large, that it is nearly impossible to collect, process and store data using traditional database management system and software techniques. Therefore, big data requires different approaches and tools to analyze data. The process of collecting, storing and analyzing large amount of data to find unknown patterns is called as big data analytics. The information and patterns found by the analysis process is used by large enterprise and companies to get deeper knowledge and to make better decision in faster way to get advantage over competition. So, better techniques and tools must be developed to analyze and process big data. Big data mining is used to extract useful information from large datasets which is mostly unstructured data. Unstructured data is data that has no particular structure, it can be any form. Today, storage of high dimensional data has no standard structure or schema, because of this problem has risen. This paper gives an overview of big data sources, challenges, scope and unstructured data mining techniques that can be used for big data.**

Keywords: **Big Data; Data Analytics; Unstructured Data; Unstructured Data Mining; Analytics as a Service**

I. INTRODUCTION

According to IDC report[1], from 2005 to 2020, the size of data will increase by a factor of 300, from 130 Exabyte to 40,000 Exabyte representing a double growth every two years. The data generated with this rate is named as Big Data [2]

Traditional data management and analysis system is based on structured data, therefore, systems like Relational database management system(RDBMS) are not adequate to process big data.

Relational database management system is not suitable for big data because:

- Traditional database can process only structured data and only 10% of global data is in structured format.
- Traditional database are not scalable as rate of generation of big data is very high. So, to process and store big data, system must be scalable.

IDC defined big data in 2011[3]: "Big data technologies describe a new generation of technologies and architecture designed to economically extract value from very large volumes of a wide variety of data, enabling high velocity capture, discovery and/or analysis ".

In 2011, Mackinsey's report[4] defined big data as " datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze ".

NIST[5] states that, " Big data is where the data volume, acquisition velocity or data representation limits the ability to perform effective analysis using the traditional approaches or requires the use of significant horizontal scaling for efficient processing".

A. *Fundamental characteristics of big data:*

IBM researchers have defined 3V model for big data[6]:

i. Volume

Big data is generated in continuous exponential rate. The rate of data generation is so fast that 90 percent of today’s data is generated in last two years [7]. Twitter generates 12 TB of data daily [8]. Facebook stated that its users registered 2.7 billion "Like" and "Comments" per day[9] in Feb 2012.

ii. Variety

Large size of data is generated by different autonomous sources. Most of the data generated by these sources is heterogeneous. This data has no fixed schema hence it is considered as unstructured data. Previously data has fixed format (often termed as structured data) , KDD algorithm were useful. But big data is either semi structured or unstructured, techniques and tools need to be developed to analyze data in real time.

iii. Velocity

The rate of data being generated is very high. Large set of this data is generated by social media.

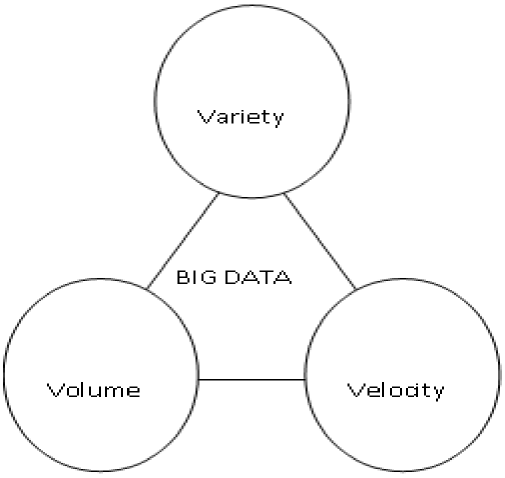


FIGURE I: 3V MODEL FOR BIG DATA[6]

TABLE I. COMPARISON BETWEEN TRADITIONAL DATA AND BIG DATA

Characteristics	Traditional data	Big data
Volume	Gigabytes	Terabytes/Petabytes
VELOCITY	Per hour,day..	Fast rate
Variety	Only structured	Structured, semi structured and un structured
Value	Less	High
Data source	Central database	Distributed database

II. DATA ANALYSIS

The purpose of data analysis is to extract information and to find unknown and hidden patterns.
Aim/Purpose of data analysis:

- To check whether data is legal
- To get information for future use.
- To make system robust.

As data is generated in many form. Therefore, method of analysis that need to be performed is different. According to Blackett et al. [10] data analytics can be divided into

i. Descriptive analytics

Descriptive analysis is based on historical data. It mainly concerns with Business Intelligence.

ii. Predictive analytics

It deals with predicting future scope. For this purpose it uses statistical data.

iii. Prescriptive analytics

Prescriptive analytics helps in decision making. It tries to find out optimize solution for a given problem with given constraints.

III. UNSTRUCTURED DATA ANALYTICS

A. *AaaS*

IBM stated that AaaS[11] is the service which can be used to perform analysis of unstructured data. IBM introduced AaaS platform that allows companies to submit data which can be structured or semi structured or unstructured format to perform analysis.

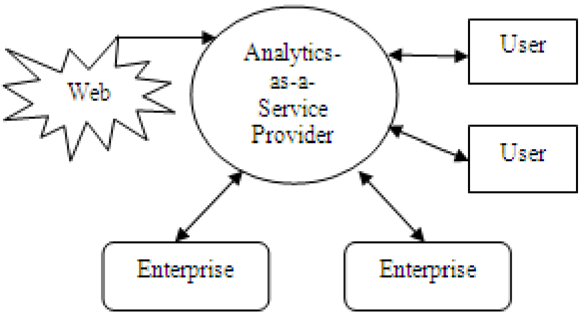


FIGURE II: ANALYTICS-AS-A-SERVICE MODEL PROPOSED BY IBM[11]

IBM soon realized the problem that are faced by AaaS:

- Definition of Service Level Agreement
- Quality of Service examining strategy
- Pricing
- Data management
- Processing models

Sun et al[12] propose AaaS is a next level of SaaS that allows companies to use analytics as a service.

IV. UNSTRUCTURED DATA MINING

Without data mining, meaningful information cannot be extracted from big data. As already stated, existing methodologies are designed to process schema oriented data and cannot be used for unstructured data, Researchers are trying to develop techniques that can perform data mining on unstructured data in real time.

Some of the techniques that can be for data mining are:

- Information Retrieval algorithms based on templates[13],
- Association Rules[14-15],
- Document clustering [16], and so on.

Table II summarizes the methods for data mining.

TABLE II. TECHNIQUES OF UNSTRUCTURED DATA MINING[17]

	Information Extraction	Association Rules	Topics	Terms	Document Clustering	Document Summarization	Re-Usable Dictionaries
Goal	Data Retrieval, KDD process	Discovery in text, file	Topic Recommendation	Establish associative relationships between terms	Organize documents into sub-groups with closer identity	Noise Filtering in large Documents	For Knowledge Collaboration and sharing
Data Representation	Semantics and keywords	Keywords, sentences	Keywords,Lexical chaining	Keywords	Data documents	Terms, topics	Words, sentences
Natural Language Processing	Feature extraction	Keyword extraction	Lexical chaining	Lemmas	Feature extraction	Lexical chaining	Feature extraction
Output	Structured or semi structured Documents	Summary report	Topic linkages	Crawler depth	Visualization	Summary report	Summary report
Techniques	Data mining, Tagging	Semantics, Linguistics	Subscribe Topics association Occurrences	Term generation	Clustering	Document analysis	Annoations, Tagging
Search space	Document, Files Database	Document, Files Database	Document, Files Database	Topics vector space	Databases	Databases	Databases
Evaluation metrics	Similarity of documents and keywords	Similarity, Mapping of features	Association, Similarity	Frequency of terms	Sensitivity	Transition	Word extensibility
Challenge and Future Directories	Lack of research on Data Pedigree	Applies varying approaches to different data sources	Identification of topics of interest may be challenging	Community based so cross-domain adoption is challenging	Can be resource intensive therefore needs research on parallelization	Needs research on data Pedigree	Lack of research on dictionary adaption to new words

V. AAAS ARCHITECTURE

Lomotey et al proposed tool that uses AaaS for unstructured data.[18] The architecture is shown in figure.

It has four components

- End user
- Front end
- Database
- AaaS framework

The end user communicates with the system through input layer. It is an interface designed in HTML5 and customized.

User specifies two parameters:

- Search criteria selection:

This tells system the type of mining to be performed

- Data source selection:

User can specify data source where data mining to be performed.

The information specified by user is given to Request Parser. Request Parser acts as system interface. AaaS is designed using Erlang[19] programming environment. Output of Input layer is passed to Request parser as input.

The request is given to a Artifact Extraction Definition which validates inputs. User input is invalidated when data source specified by the user is not present.

Data format specified by the user is not supported by the system.

The request is passed to the Semantic Engine layer which can be in either topic format or term format. The semantic engine layer performs iterative checks on request to get better result. Artifact Extraction component specifies request either term extraction or top extraction. Term extraction is specified then it has to check all related keywords as keyword specified by the user may have other similar meaning. The artifact then forwarded to Topic Parser where definition of artifact is checked. Topic parser consists of two components Dictionary and Thesaurus. These structure build table format for each artifact which consists of artifact and their corresponding synonyms and antonyms . These table are updated when new entry is found.

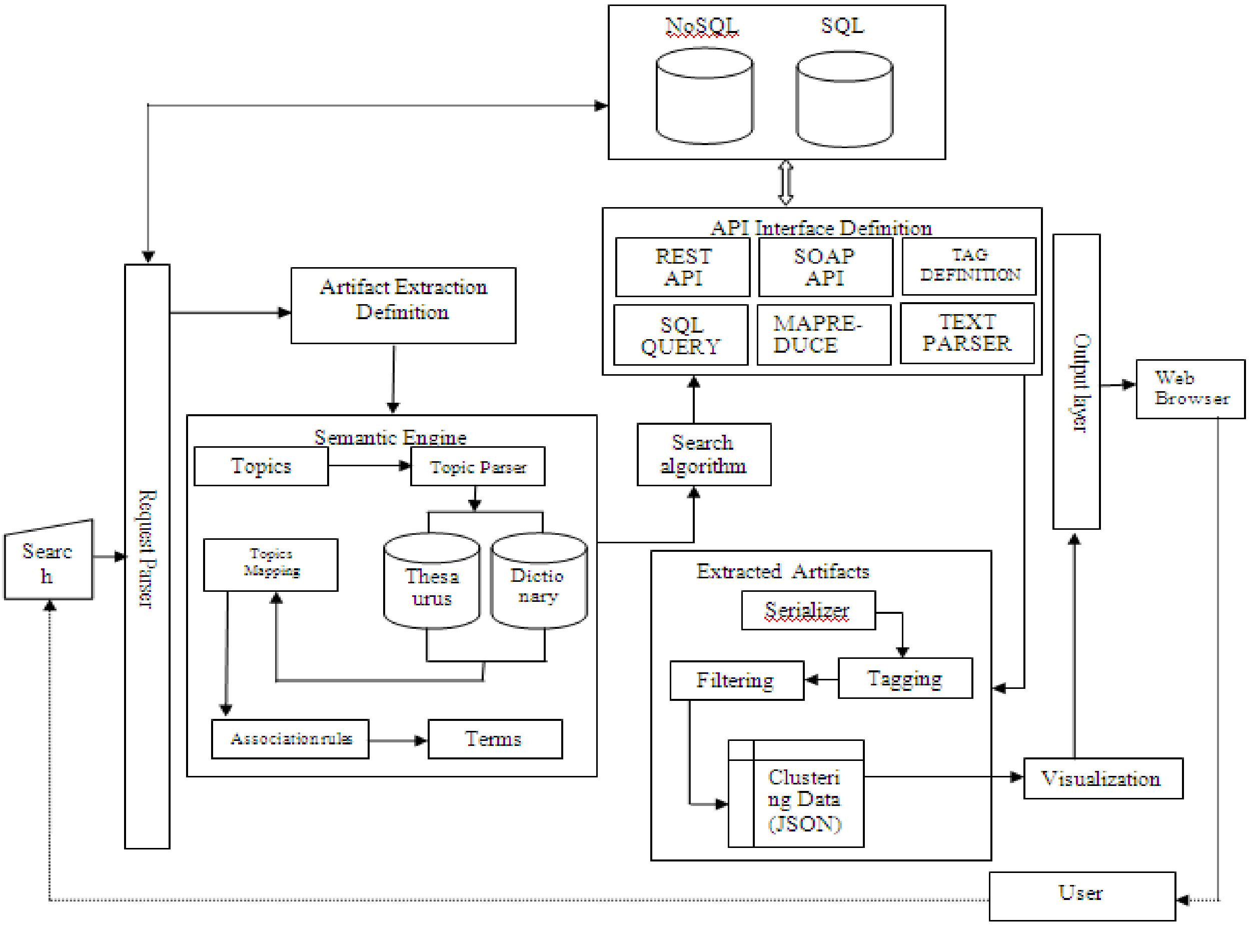
Topics Mapping component creates relationship between artifacts. To link words to each other it uses Bloom filtering methodology. Once all dependencies are found association rules are formed.

These rules are programmed in JSON script. After JSON script is submitted to Search Algorithm layer which perform search through database.

API Interface Definition layer generates queries to perform searching. As database has different structure hence different APIs are defined for different data sources.

The output of query is passed to the Serializer component which performs redundancy check, organiztion of same artifacts and combining of search result. After these activities Tagging is done on data. Once tagging is finished , data is given to Filtering component. Filtered data is grouped into clusters by Clustering component and it is passed to a Visualization layer. The result is shown to user based on the platform user is running.

FIGURE III: THE ARCHITECTURE OF AAAS TOOL[18]



VI. KEY CHALLENGES AND ISSUES OF BIG DATA

A. *Privacy and security*

This is one of the most important issue related with big data as it consists of sensitive and private data. The database consists of personal information if analyzed can give facts about a person and that information may be confidential so that person may not like to disclose that information to other people.

B. *Data access and sharing information*

The rate at which big data is generated is high so it must be maintained timely manner as data require to make decision has to be accurate.

C. *Space issues*

There are many diverse sources of big data which are continuously generating data. IBM states that every year 2.5 quintillion bytes of data is generated at the tremendous speed. As data generated in sheer volume, storage devices must have high storage capacity.

D. Processing issues

Big data is changing rapidly, it is a challenge to process data in real time to keep its integrity. Sometimes data need to be moved from one device to another for processing but as it is very difficult to move terabyte of data in real time. Every time data to be processed need to be scanned first which takes a lot of time. So building an indexes may reduce scanning and processing time.

E. Robust

It is not possible to build a system which is 100% reliable and fault tolerant. But system must be developed to handle any kind of failure such that it will able to recover from failure and damage done will be minimum. Big data technology has to deal with large volume of diverse data. As data is divided into small sets and these small sets are processed individually on each respective node. If any error occurs then that process can be restarted again.

F. Scalability

Heterogeneous autonomous sources continuously generate data with exponential rate. Methodologies designed to handle big data must be capable of processing this ever increasing data in real time.

VII. FUTURE SCOPE

From the current scenario, it is clear that Big Data will stay for long. Data is generated at exponential rate, though big data has its own advantages but the challenge is how to analyze and perform operation on data mining. So, we need a system tools which can manage big data and perform mining not only on structured data but also on unstructured data efficiently.

VIII. CONCLUSION

Big data is evolving over time and its size is getting doubled in every two years. Previously all tools were developed for schema oriented data. Though big data offers many opportunities but extracting data is new challenge.

AaaS service is a good option to analyze and perform mining on big data.

IX. ACKNOWLEDGEMENTS

The authors thank Dr.S.Sridhar, Professor and Dean, Cognitive & Central Computing Facility of R.V.College of Engineering, Bangalore, India for reviewing this article and communicating to this Journal for publication. The authors also thank Dr.G.Shobha, Professor and Head, Department of CSE, R.V.College of Engineering, Bangalore, India and Dr.N.K.Srinath, Professor and PG-Dean , for the support given.

X. REFERENCES

[1] J. Gantz and D. Reinsel, ``The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," in Proc. IDC iView,IDC ANAL. FUTURE, 2012.

[2] M. R. WIGAN, AND R. CLARKE, "Big Data's Big Unintended Consequences," Computer , vol.46, no.6, pp.46-53, June 2013, doi:10.1109/MC.2013.195

[3] J. Gantz and D. Reinsel, ``Extracting value from chaos," in Proc. IDC iView, 2011, pp. 1_12.

[4] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. San Francisco, CA, USA: McKinsey Global Institute, 2011, pp. 1_137.

[5] M. Cooper and P. Mell. (2012). Tackling Big Data [Online]. Available: http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/f%csm_june2012_cooper_mell.pdf

[6] P. C. ZIKOPOULOS, C. EATON, D. deROOS, T. DEUTSCH, AND G. LAPIS, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data," Published by McGraw-Hill Companies,2012 https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Big%20Data%20University/page/FREE%20ebook%20%20Understanding%20Big%20Data.

[7] "IBM What Is Big Data: Bring Big Data to the Enterprise,"

http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[8] K. RUPANAGUNTA, D. ZAKKAM, AND H. RAO, "How to Mine Unstructured Data," Article in Information Management, June 29 2012,

http://www.information-management.com/newsletters/dataminingunstructured-big-data-youtube--10022781-1.html

[9] S. Marche, ``Is Facebook making us lonely," Atlantic, vol. 309, no. 4, pp. 60_69, 2012.

[10] G. Blackett. (2013). Analytics Network-O.R. Analytics [Online].

Available: http://www.theorsociety.com/Pages/SpecialInterest/AnalyticsNetwork_anal%ytics.aspx

[11] IBM Research, “Analytics-as-a-Service Platform,” Available:

http://researcher.ibm.com/researcher/view_project.php?id=3992

[12] X. SUN, B. GAO, L. FAN, AND W. AN, "A Cost-Effective Approach to Delivering Analytics as a Service," IEEE 19th International Conference on Web Services (ICWS 2012), vol., no., pp.512,519, 24-29 June 2012, doi: 10.1109/ICWS.2012.79

[13] J. Y. HSU, AND W. YIH, “Template-Based Information Mining from HTML Documents,” American Association for Artificial Intelligence, July 1997.

[14] M. DELGADO, M. MARTÍN-BAUTISTA, D. SÁNCHEZ, AND M. VILA, “Mining Text Data: Special Features and Patterns,” Pattern Detection and Discovery, Lecture Notes in Computer Science, 2002, Volume 2447/2002, 175-186, DOI: 10.1007/3-540-45728-3_11

[15] Q. ZHAO AND S. S. BHOWMICK, “Association Rule Mining: A Survey,” Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003.

[16] L. HAN, T. O. SUZEK, Y. WANG, AND S. H. BRYANT, “The Textmining based PubChem Bioassay neighboring analysis,” BMC Bioinformatics 2010, 11:549 doi:10.1186/1471-2105-11-549

[17] R. K. LOMOTEY AND R. DETERS, “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, 27 June–2 July 2013.

[18] Lomotey, R.K., and R. Deters. “Towards Knowledge Discovery in Big Data.” Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on, April 7, 2014, 181–91. doi: 10.1109/ SOSE. 2014. 25.

[19] Erlang Programing Language,

http://www.erlang.org/