ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

BIG DATA ANALYTICS: AN OVERVIEW

S.Jeyabharathy
bharadhina@gmail.com
Department of Information Technology,
Dr.Umayal Ramanathan College for Women, Karaikudi.

ABSTRACT

Big data analytics is the process of collecting, processing and analyzing data sets to create patterns and other useful information. Big data analytics can help us to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with big data basically want the knowledge that comes from analyzing the data. To analyze such a large volume of data, big data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, and forecasting and data optimization. This survey paper presents the existing methodologies commonly used for analyzing text, structured data, web content, mobile analytics, and network.

Keywords: Big data Analytics, Analytics Algorithms, Data Analytics.

I. INTRODUCTION

A. Big data definition

Big data means very large volumes of wide variety of data acquired by capturing, discovery and analysis. It is also defined as datasets whose size is more than the default size that a database software can hold. Big data characteristics are defined by its characteristics such as volume, velocity, veracity.

Table I Difference between Traditional data & Big data

	Traditional Data	Big Data
Volume	GB	Terabyte
Generated Rate	Per hour, day	More Rapid
Structure	Structured	Semi-structured /unstructured
Data source	Centralized	Fully distributed
Data Integration	Easy	Difficult
Data store	RDBMS	HDFS,NoSQL
Access	Interactive	Batch/real time

B. Big data system

Big data system consists of 4 stages for processing

1. Data Generation -Collecting the data from different sources such as sensors, Videos, digital sources.
2. Data Acquisition - Process of obtaining information by data collection, data transmission, Data Preprocessing.
3. Data storage -Storing & managing the large scale datasets.
4. Data analysis -Analytical methods or tools to inspect, transform, model data to extract value.

Table II BIG DATA SOURCES

DataSource	Application	Data Scale	Type
Walmart	Retail	PB	Structured
Amazon	e-commerce	PB	Semi-structured
Google search	Internet	PB	Semi-structured
Facebook	Social network	PB	Structured, unstructured
Health care	Internet of things	TB	Structured, unstructured

II. DATA ANALYSIS

The goal of data analysis is to extract useful information for Recommender system, Decision support system.*Blackettetal [2]* classified data analytics in to 3 types

- Descriptive analytics
- Predictive analytics
- Prescriptive analytics.

All Rights Reserved © 2016 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Descriptive Analytics

It explains or summarizes the data to a form which can be interpreted by the user. It explains the past happenings, with this future can be predicted. The raw data can be counted, aggregated and filtered using mathematical functions. It is applicable in case of analyzing inventory, calculating the amount spent by the customer, analyzing the sales over years.

Eg: Company production, Financials, operations, sales, finance, inventory & customers.

Predictive Analytics (understanding the future)

Predictive analytics "Predict" the future outcome. Having a sample data/uncompleted data ,Predictive analytics uses probabilities to complete the data with a guess. It does not use statistical algorithms. It is mostly applicable in business to predict the customer needs based on sales. It is also applicable in supply chain, inventory operation.

Prescriptive Analytics: (advice on possible outcomes)

It is a new concept which allows users to prescribe different opinions for a selection. This analytics is mainly applicable for decision support system .Prescriptive analysis not only predicts the reason for happening but also provides the recommendation to take actions. This analytics uses business rules, algorithms, machine learning & computations modeling procedures. It is used for optimizing production, scheduling & inventory in supply chain.

III. BIG DATA ANALYTICS

There are six types of research areas in big data analytics such as structured data analytics, Text analytics, web analytics, multimedia analytics, network analytics, and mobile analytics.

A.STRUCTURED DATA ANALYTICS

Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets. Machine learning algorithms are used for processing this type of data. In machine learning algorithms, Statistical machine learning algorithms are used for anomaly detection, Spatial mining can be used for processing sensor data, privacy –preserving data mining is used for extracting information in government documents and health care applications.

B.TEXTDATA ANALYTICS

The Text refers to text in mail communication, industry documents, web content, text posted in social networks and the information stored in the system. Text analytics means extracting meaningful information from unstructured text. Methodologies used for text mining is information extraction, topic modeling, summarization, categorization, clustering, question and answering and opinion mining [1].

Information extraction, extracts the structured information from the existing text. In information extraction ,named-entity recognition (NER) finds entities in text such as person, location, and organization.

Topic modeling, Here the document content will be from various topics in which a topic is generated using probability among the words. Various probability topics are used for content generation as well as meaning of the words. [3].

Summarization generates an abstract from multiple text documents. This is classified as extractive summarization and abstractive summarization [4].

Text categorization points out the main concept in a document by comparing with set of topics. Recent research is on Graph representation and graph mining-based text categorization.[5]

Text clustering is used to cluster the documents. Clustering means grouping similar documents. Documents from multiple subtopics are clustered .Data mining algorithms are used for clustering. [6].

A question answering system find out the best answer for a given question .Here it uses various techniques for analyzing questions, source retrieval, answer extraction and presenting answer. This system is applied in various fields such as education, Websites, Health, Defense. [7]

Opinion mining assesses the opinions from the information published in news, comments in websites and other contents generated by the users. This analysis is useful for collecting the opinions of public and customers about social events and product preferences [8].

C. WEBANALYTICS

The information from web documents are retrieved for knowledge discovery .Web analytics covers 3 areas Web content mining, web structure mining, web usage mining.

1) WEBCONTENT MINING: Web Content mining refers to the extracting useful information from web content consisting of text, images videos etc. [9, 10]. It follows two approaches; Agent based approach and database approach [10, 11]. There are three types of agents, intelligent search agents, Information filtering/Categorizing agent, and personalized web agents [10].

All Rights Reserved © 2016 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Intelligent Search agents - Search information for a query using domain characteristics and user profiles.
Information agents - Filters the data according to the predefined instructions
Personalized web agents - learns user preferences and Creates documents related to those user profiles [10, 11].

In Database approach it consists of database schemas and attributes with defined domains.

2) WEB STRUCTURE MINING: Web structure mining focus on the hyperlink structure of the Web. The links are properly activated to predict the accuracy of learned models. For an independent link it leads to a wrong conclusion. So, Here correct linking should be done for potential correlations and improve the predictions accuracy. [13]. two algorithms that have been proposed to lead with those potential correlations: HITS and Page Rank.

3)WEB USAGE MINING: It automatically discovers the patterns in Click streams, user transactions and other associated data collected or generated as a result from one or more websites. The main goal is to capture, model and analyze the behavioral patterns and profiles of user visiting website. The resultant patterns are represented as pages, objects that are frequently accessed by user groups [1].

D. NETWORK ANALYTICS

Social networks contain a tremendous amount of linkage and content data. Linkage data will be in graph structure, communication between entities. Content data will be as text and images. These data are analyzed as linkage based structural analysis and content based analysis.

Linkage based structural analysis

It focuses on link prediction, community detection, social network evolution, social influence analysis. Social network structure is like graph where each vertex refers to a person edge represents Association between nodes is done by link prediction. The techniques used for link prediction are Feature based classification, Probabilistic approaches and linear algebraic approaches [1].

E. MOBILE ANALYTICS

The huge volume of mobile applications leads to mobile analytics in handling mobile data (noisiness, redundancy, sensitivity)

Data from mobile apps and sensors (wireless sensor data) are analyzed for real time monitoring the individual health.

RFID allows a sensor to read a unique product identification code (EPC) associated with tag. This tag is used to track the location, monitor physical objects. Here RFID data analytics to find out the noisy and redundant.
In recent days body sensor networks have been emerged for monitoring a person health [1].

F. MULTIMEDIA ANALYTICS

In [1] Multimedia data consists of audio, video, image. Here data analytics is done to understand the semantic content in multimedia data. The sources of data are Health & patient media, Surveillance department, user groups. Methodologies used here are Summarization, annotations, Indexing and retrieval, Recommendation, Event detection.

Summarization - Extracting salient words from Original data
Annotation - Assigning images, videos a set of Labels describing their content in Semantic levels.
Multimedia Indexing & Retrieval - Indexing is done for quick access.
Multimedia Recommendation -Suggesting the multimedia contents with high quality for user preferences.

1) CONTENT BASED FILTERING: Information filtering deals with delivery of items selected from a large collection of user's interest. Filtering system classifies the unseen items as positive items relevant to the user and negative for irrelevant user. For this Learning method is used , It select a function based on training set of m input vectors that can classify any item in the collection. The function h(x) classifies an unseen items as positive or negative at once. The threshold determines the item relevant to user or irrelevant to user. While displaying the content the similarity of the document is also considered [15].

2) COLLABORATIVE FILTERING: Collaborative filtering is a successful recommender technique. It provides item recommendations based on user opinion. The main goal of CF is to recommend new items to a user based on the user's previous linkings. Opinions can be explicitly given by user as rating score .CF either predict the likeness of an item or recommend an item liked by the user.CF

All Rights Reserved © 2016 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

identifies the common interest and recommend for others users [16].

IV. COMMON METHODOLOGIES

Data analysis methods are commonly classified as 3 types [1]

A.DATA VISUALIZATION

This method represents data in clarity using graph, chart and maps. The user can understand easily from these representations.

B.STATISTICAL ANALYSIS

This method is used for analyzing description and inference data sets. Descriptive analysis is for creating summary for the acquired data. Inferential statistical is to point inferences about the process. It uses the following techniques Aggression, factor analysis, and clustering, discriminate analysis.

C. Data mining

It discovers patterns in large data sets. Some common data mining algorithms are [14]

1) C4.5: Constructs a classifier in the form of decision tree.C4.5 is given a set of data that are already classified. Classifier is a tool in data mining, takes data to classify and attempts to predict the new data belong to which class. This is supervised learning generates a decision tree and uses it to classify.

2) K-means: K-means creates k groups from a set of objects. So that the members in the groups are similar. It uses clustering technique for exploring a data set. Cluster analysis is a algorithm designed to form groups, such that the group members are similar than non-group members. It uses unsupervised learning. It is simple, faster and more efficient than other algorithms suitable for large data sets. It works only on continuous data.

3) SVM (Support Vector Machine): It is a classification method. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. For example, if we only had two features like Height and Hair length of an individual, we'd first plot these two variables in two dimensional space where each point has two co-ordinates (these co-ordinates are known as Support Vectors)

4) Apriori: It is Learning Association rules applied on a database having more number of transactions. Association rules are data mining techniques for correlations and relations among variables in a database. It identifies the frequent item sets from datasets. It performs Join, prune, repeat operations.

5) AdaBoost: It is a boosting algorithm which constructs a classifier. Classifier takes a collection of data to predict or classify. Boosting is an ensemble learning algorithm, which takes multiple learning algorithms and combines them. This is supervised learning .It is simple and can work with a large variety of data.

6) KNN: k-Nearest neighbors is a classification algorithm. It doesn't create any classification model instead it just stores the labeled data. It looks at the k-closest labeled point's i.e, k-nearest neighbors. Using neighbor's classes it classifies the new data. For continuous data, KNN uses Euclidean distance. For Discrete data it first transform into continuous data and uses hamming distance to find the closest neighbor. It is expensive when applying on large dataset. It cannot be applicable for noisy data. Finding distance between 2 neighbors is difficult for larger range of values. It requires greater storage.

7) NAÏVE BAYES: It is a collection of classification algorithms. It is used to predict a class with a set of features and find probability.

$$P(CA|F1, F2) = \frac{P(F1|CA).P(F2|CA).P(A)}{P(F1).P(F2)}$$

Where,
CA is Class A, F1 is Feature 1 and F2 is Feature 2,
P is Probability

Navie Bayes involves simple algorithm such as counting, multiplying, and dividing. Once the tables are calculated probabilities for all the classes are calculated and choose highest probability. It gives a accurate result. It is effective for span filtering.

8) CART:CART stands for Classification and regression. It is a decision tree learning technique so that the output will be either classification or regression. Classification predicts a class while regression predicts a numeric or continuous value. The output of classification is classes & the output of regression is numbers.

All Rights Reserved © 2016 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

V.CONCLUSION

In this paper, I have presented the concept of big data which covers the sources of big data, after retrieving the big data it is processed using four steps Data generation, Data acquisition, Data Storage, Data analysis. After processing the information extracted is analyzed, the analysis can be predictive analysis, descriptive analysis, and Prescriptive analysis. For analyzing variety of Big data different Data mining algorithms were used. The mostly used algorithms for analyzing are also explained in this paper about their purposes.

REFERENCES

[1] Han hu, Yonggang Wen,Tat-Seng Chua, Xuelong L "Toward Scalable for Big Data Analytics: A Technology Tutorial", IEEE translations and content mining ,Vol2 , 2014
[2] G. Blackett. (2013). Analytics Network-O.R. Analytics [Online].Available: http://www.theorsociety.com/Pages/ SpecialInterest/AnalyticsNetwork_anal%ytics.aspx
[3] D. M. Blei, ``Probabilistic topic models," Commun. ACM, vol. 55, no. 4,pp. 77_84, 2012.
[4] H. Balinsky, A. Balinsky, and S. J. Simske, ``Automatic text summarization and small-world networks," in Proc. 11th ACM Symp. Document Eng.,2011, pp. 175_184.
[5] M. Mishra, J. Huan, S. Bleik, and M. Song, ``Biomedical text categorization with concept graph representations using a controlled vocabulary," in Proc. 11th Int. Workshop Data Mining Bioinform., 2012, pp. 26_32.
[6] J. Hu et al., ``Enhancing text clustering by leveraging wikipedia semantics," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retr., 2008, pp. 179_186.
[7] M. T. Maybury, New Directions in Question Answering. Menlo Park, CA,USA: AAAI press, 2004.
[8] B. Pang and L. Lee, ``Opinion mining and sentiment analysis," Found.Trends Inform. Retr., vol. 2, nos. 1_2, pp. 1_135, 2008.
[9] Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations. Vol. 2, 1-15.
[10]Inamdar, S. A. and shinde, G. N. 2010. An Agent Based Intelligent Search Engine System for Web Mining. International Journal on Computer Science and Engineering, Vol. 02, No. 03.
[11]Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.
[12] Miguel Gomes da Costa Júnior Zhiguo Gong, Av. Padre Tomás, S.J., Taipa, Macao S.A.R., China, Web Structure Mining: An Introduction, Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China
[13] L. Getoor, Link Mining: A New Data Mining Challenge. SIGKDD Explorations, vol. 4, issue 2, 2003.
[14] (2015).Top 10 datamining algorithms [Online].Available http://www.kdnuggets.com/2015/05/top-10-data- mining-algorithms-explained.html/3
[15] Robin van Meteren, Maarten van Someren, "Using Content-Based Filtering for Recommendation".
[16] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms", ACM

All Rights Reserved © 2016 IJARTET