ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 07, Issue, 07, pp.13714-13717, July, 2017

ORIGINAL RESEARCH ARTICLE Open Access

AN AGRICULTURE SURVEY OF BIG DATA MINING APPLICATIONS

[1]Swarupa Rani, A. and [2]Jyothi, S.

[1]Research Scholar, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, AP, India
[2]Dept. of Computer Science, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, AP, India

ARTICLE INFO

Article History:
Received 05th April, 2017
Received in revised form
14th May, 2017
Accepted 26th June, 2017
Published online 22nd July, 2017

Key Words:

Aggregation,
Big data,
Data Mining,
Data streams,
Statistics,
GIS,
Agriculture.

Corresponding author:

ABSTRACT

The data is collected from various sources like laboratory reports, agriculture information web pages, and expert recommendation for the developed framework. After the collection of raw data, the irrelevant or the redundant data that is also known as the noise, should be removed. The next step is to extract the features from cleaned data, normalization of data is done in order to remove the technical variations. Once normalization is complete the data is uploaded on HDFS and save in a file that is supported by Hive. Thus classified data is finally located on the specific place. In the next step HiveQL is used to analyze agriculture data based on features and then prioritize the outcome based on crop disease symptoms and in the last a high priority solution is recommended. In the paper prioritize outcomes are useful for agriculture officers, researchers to easily understand, and helpful for recommending a solution based on evidence from historical data. The major aspire of this paper is to make a study on the concept Big data and its application in data mining. The paper mainly concentrating different types of big data and its application in knowledge discovery.

Copyright ©2017, Swarupa Rani and Jyothi. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Citation: Swarupa Rani and Jyothi. 2017. "An Agriculture Survey of Big Data Mining Applications", *International Journal of Development Research*, 7, (07), 13714-13717.

INTRODUCTION

Big data is massive volume of both structured and unstructured data from various sources such as social data, machine generated data, traditional enterprise which is so large that it is difficult to process with traditional database and software techniques. Big Data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Characteristics of Big Data include 5 Vs. They are Volume, Velocity, Variety and Veracity. Big data is mainly used to spot trends, to determine the quality of research, to prevent disease, to link legal citation etc. It is used in different applications such as Medicine, Physics, Simulation, RFID, Astrometry, Biology etc. There are different types of data such as relational, structural, textual, semi structured, graph data, streaming data etc can be included in big data. These data can be used for Aggregation and Statistics in Data warehouse and OLAP, Indexing, Searching, and Querying, Keyword based searching, Pattern matching (XML/RDF), Knowledge discovery in Data Mining and Statistical Modeling.

MATERIALS AND METHODS

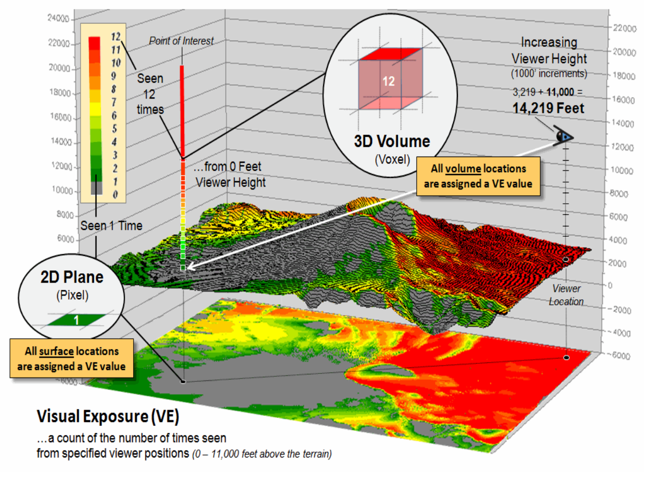
Big data includes structured, semi-structured, and unstructured data. This unstructured data contains useful information which can be mined using suitable data mining technology. We can see that the digital streams that individuals create are growing rapidly. Most of the people are using camera on their own mobile. Big Data are of high level volume, high velocity, and high variety of information that needs advanced method to process the Big Data. Moreover, the conventional software tools are not capable of handling such data.

Big Data requires extensive architecture also. Different types of data such as Social data – Customer feedback forms for Customer Relationship Management (CRM) in Social media sites such as Twitter, Face book, LinkedIn etc. Machine-generated data in sensor readings, satellite communication. Traditional enterprise data such as and ledger information. Employee information, customer information etc are referred as big data.

Characteristics of Big Data

There are mainly four characteristics for big data. They are Volume, Velocity, Variety and Veracity. Volume means vast amount of data generated in every second. It is a scale characteristic. The data is in rest state. Machine generated data are examples for these characteristics. Nowadays data volume is increasing exponentially. The second generated characteristics of big data are velocity or speed. Velocity is the speed at which data generated. The streaming data may not be massive and its state is in motion. It should have high speed data. Example is data created through social media. The data is begin generated fast and need to be processed fast. Online Data Analytics includes these types of big data. E-Promotions and health care monitoring are examples. In e-promotion, **b**ased on our current location and our purchase history, what we like will send promotions right now for store next to us. In Healthcare monitoring, sensors monitoring our activities and body. Any abnormal measurements require immediate reaction can be immediately identified through this.

Variety is another important characteristic of big data. Various data formats, types, and structures can be referred here. The type of data may include different varities such as Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…It also includes s static data and streaming data . A single application can be generating by collecting many types of data. To extract the knowledge all these types of data need to be linked together. Veracity means data in doubt. The uncertainty of data can be found due to the inconsistency and incompleteness. The messiness of data (Abbreviation, colloquial speech etc) may result the veracity.


Point of Interest
Seen 12 times
Seen 1 Time
12 11 10 9 8 7 6 5 4 3 2 1 0
...from 0 Feet Viewer Height
3D Volume (Voxel)
12
Increasing Viewer Height (1000' increments)
3,219 + 11,000 = 14,219 Feet
All volume locations are assigned a VE value
2D Plane (Pixel)
1
All surface locations are assigned a VE value
Viewer Location
Visual Exposure (VE)
...a count of the number of times seen from specified viewer positions (0 – 11,000 feet above the terrain)
24000 22000 20000 18000 16000 14000 12000 10000 8000 6000
22000 20000 18000 16000 14000 12000 10000 8000 6000 4000 2000 0 -2000 -4000 -6000

CHALLENGES IN HANDLING BIG DATA

The challenges in handling big data includes in technology. The technology needs new architecture, algorithms, techniques for its implementation. It also requires technical skills .So experts are needed for this new technology to deal with big data.

The correction and correlation of data makes more complexity. The main challenges need to be faced by the enterprises or media when handling Big Data are capturing of big data, its duration, storage, sharing of big data and its analysis, visualization of the massive data etc. Connection and correlation of data which describes more about relationship among the data.

Application of Big Data in Data Mining

In data mining a number of different data repositories can be involved. Data mining should be applicable to any kind of data repository as well as to transient data such as data streams. The challenges and techniques of mining may differ for each of the repository systems. Advanced databases or information repositories require sophisticated facilities to efficiently store retrieve and update large amounts of complex data. They also provide fertile grounds to raise many challenging research and implementation issue for data mining. For data mining in object relational system, techniques need to be developed for handling complex object structures, complex data types, class and sub class hierarchies, property inheritance and methods and procedures. Data mining techniques can be used to find the characteristics of object evaluation or the trend of changes for objects in the database. Such information can be useful in decision making and strategy planning. For example stock exchange data can be mined to uncover trends that could help to plan investment strategies. (Marek Kretowski, 1984) Geographic databases have also numerous applications ranging from forestry and ecology planning to provide public service information regarding the location of cables, pipes or sewage system. They are also useful for vehicle navigation. Spatiotemporal database that change with time is also a big data in which information can be mined. (Hunt, 1962) Streams of data flow in and out of an observation pattern dynamically. They may be huge infinite volume, dynamically changing in nature. Usually multi level, multidimensional on-line analysis and mining should be performed on stream data. Even if the web pages are fancy and informative to readers, they can be highly unstructured and lack pattern. Data mining can often provide additional help to the web search services which include big data. Data mining are used to specify the kind of patterns to be found in data mining task. The tasks can be classified as predictive and descriptive.

Different types of data mining system

There are different types of data mining system which can be used with big data. The main techniques used with data mining are as follows.

Classification

Classification is the process of finding a model or function that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of asset of training data. The model can be represented in various forms such as classification rules, decision tree, mathematical formulae or neural networks. (Arijay Chaudhry) Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded.

Evolution Analysis

Evolution analysis is used with time series data of previous years. Regularities in such time series data is used to predict future trends in stock market prices, contributing to decision making regarding stock investments.

Outlier Analysis

Outlier analysis may be detected using statistical tests that assume a distribution or probability model for the data or using distance measures where objects that are a substantial distance from any other cluster are considered outliers.

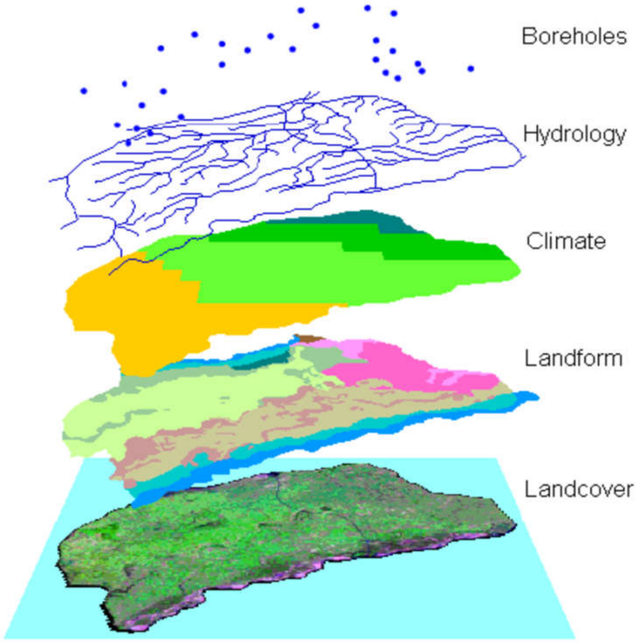
Cluster Analysis

In cluster Analysis, there are no class labels in the training data sets.

The labels are generating using this technique. The objects in a cluster are grouped based on their similarity. Then rules are formed from the clusters .The major clustering methods includes portioning methods, hierarchical methods, density based methods, model based methods and constraint based clustering method. If the cluster contains large number of data or big data, then it has to used methods like frequent pattern based clustering or high dimensional data clustering.

TOOLS FOR HANDLING BIG DATA

There are many tools are currently available for handling big data some of them are follows. Map Reduce (Arijay Chaudhry) is a programming model for handling complex combination of several tasks and it was published by Google. It is a batch query processor and can run an ad hoc query for whole dataset and get the results in a sensible manner which has to be transformative.



It has two steps. 1. Map: Queries are divided into sub queries and allocated to several nodes in the distributed system and processed in parallel. 2. Reduce: Results are assembled and delivered. Oracle has introduced the total solution for the scope of enterprise which requires Big Data. Oracle Big Data Appliance (Breiman et al., 1984) is a tool to integrate optimized hardware and extensive software into Oracle Database 11 to endure the Big Data challenges.

The Real-time application of Big Data can also be in agriculture Information System on Cloud (Charissis, 1993b). Crop growth and pest management (CGPM) is an emerging technique to store the climate Information Record and exchange the data over the network, which is stored at the cloud for accessing the data log anytime and anywhere. CGPM includes variety of data such as structured, unstructured, and semi-structured. In CGPM, we propose machine generated data by acquiring the GIS information pattern or crop of the pests for saving the entire Agriculture data of the crop growth. It uses GIS spatial information capturing the crop and pest generation Identification.

RESULTS

Big Data are used to be included for finding the crop conditions, for identifying the market trends, for increasing the innovations, for retaining the customers, for performing the operations efficiently. Flood of data coming from many sources must be handled using some non-traditional database tools. It provides more market value and systematic for the upcoming generation. Big data has a variety of application and influence in the Agriculture field of data mining.

Conclusion

To implement data mining techniques, we can use big data concept. Big data presents more opportunities for research and reference in the public sector as well in technical progress. The challenges in data analyzing can be overcome by capturing the techniques in big data.

REFERENCES

Choudhary, A. K., Harding, J. A. and Tiwari, M. K. 2008. "Data Mining in Manufacturing: A Review Based on the Kind of Knowledge", Journal of Intelligent Manufacturing, Volume 20, Number 5, 501-521.

Arijay Chaudhry and Dr. P.S. Deshpande. Multidimensional Data Analysis and Data Mining, Black Book.

Breiman, L., Friedman, J. H., Olshen, R. and Stone, C. 1984. Wadsworth, Belmont. Classification and Regression Trees.

Charissis, G., V. Moustakis, and G. Potamias. 1993b. Diagnosis of acute abdominal pain children: alication case study. Heraklion, Greece: FORTH.

Clark, P., and R. Boswell. 1991. Rule induction with CN2: Some recent improvements. In Proceedings of the Fifth Working Session on Learning, 151163.

Senthil Kumar, G. 2011. “online message categorization using Apriori algorithm” International Journal of Computer Trends and Technology- May to June Issue.

Green, M. 1980. Pediatric diagnoses. Philadelphia, Pa.: W. B. Saunders.

Hand, D, John Wiley & Sons, Chichester, 1997. “Construction and Assessment of Classification Rules.

Hipp, J., Güntzer, U., Nakhaeizadeh, G. 2000. "Algorithms for association rule mining --- a general survey and comparison". ACM SIGKDD Explorations Newsletter 2: 58.

Hunt, E. B. 1962. Concept learning: An information processing problem. New York: Wiley.

Karmaker et al. 1992. “Incorporating an EM Approach for Handling Missing Attribute Values in Decision Tree Induction”.

Kononenko, I. 1997. Semi Naive Bayesian classifier. In Proceedings of EWSL91, 206219. Tom M. Mitchell, . Machine Learning, Singapore, McGraw Hill. 1991.

Marek Kretowski, Marek Gizes, 1984. sBialystok Technical University, Poland "Classification and Regression Trees".

Mucherino A. Petraq papajorgji P.M. Paradalos. A survey of data mining techniques alied to agriculture CRPIT.3(3): 555560. 1998.

Deepika, N. et al. "Association rule for classification of heart attack patients", (IJAEST) International Journal of Advanced Engineering Sciences And Technologies Vol No. 11, Issue No. 2, 253 – 257

Piatetsky-Shapiro, Gregory; and Frawley, William J., eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.

Pei, Jian; Han, Jiawei; and Lakshmanan, Laks V. S., Mining frequent itemsets with convertible constraints, in Proceedings of the 17th International Conference on Data Engineering, April 2–6, Heidelberg, Germany, 2001, pages 433-442.

Quinlan, J.R. 1986. Induction of Decision trees. Machine Learning.

Shelly Gupta et al. "data mining classification techniques applied for breast cancer for diagnosis and prognosis "Indian Journal of Computer Science and Engineering (IJCSE).

UCI Machine Learning Repository http://mlearn.ics.uci.edu/databases. Usama et al. "On the Handling of Continuous Values Attributes in Decision Tree Generation". University of Michigan, Ann Arbor.

****