IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

A Survey on Data Mining with Big data - Applications, Techniques, Tools, Challenges and Visualization

S. Usharani[1], K. Kungumaraj[2]

Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodaikanal[1]

Assistant Professor, PG Department of Computer Science, Arulmigu Palaniandavar Arts College for Women, Palani[2]

Abstract: Data mining is analytic process designed to explore data in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information. Data mining tools predict future and behaviors, allowing businesses to make proactive, knowledge-driven decisions. With the amount of data increasing each year, more data is gathered (called Big Data) and data mining is becoming an increasingly important tool to transform this data into information. 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. This survey includes the information about the Data mining with big data - technologies, tools, visualization, Challenging issues and its related work.

Keywords: Big Data, Decision trees, Genetic algorithm, nearest neighbor method, Visualization.

I. INTRODUCTION

Big data includes structured, semi-structured, and unstructured data. This unstructured data contains useful information which can be mined using suitable data mining technology. Big Data are of high level volume, high velocity, high veracity and high variety of information that needs advanced method to process the Big Data. There are different types of data such as relational, structural, textual, semi structured, graph data, streaming data etc can be included in big data. The conventional software tools are not capable of handling such data. Big Data requires extensive architecture also. As the size is too large, decentralized control and different data sources with different types the Big Data becomes much complex and harder to handle. We cannot manage them with the local tools those we use for managing the regular data in real time.

Big data analyze the large amount of information used to uncover the hidden patterns and the other information which is useful. In that visual analysis is used for the purpose of visualization of the data sets in graphics or charts etc., it is very easy to make the decision and attractive form of the data.

II. DATA MINING APPLICATIONS

Data mining is a data analysis approach that has been quickly adapted and used in a large number of domains that were already using statistics. Examples of particular applications areas are:

- Data mining is an interdisciplinary field with wide and diverse applications
- There exist nontrivial gaps between data mining principles and domain-specific applications
- Some application domains
- Retail industry
- Financial data analysis
- Biological data analysis
- Telecommunication industry

Retail industry:
Huge amounts of data on sales, customer shopping history, etc.

- Applications of retail data mining
- Identify customer buying behaviors
- Discover customer shopping patterns and trends
- Improve the quality of customer service
- Achieve better customer retention and satisfaction
- Enhance goods consumption ratios

Design more effective goods transportation and distribution policies.

Financial data Analysis:

- Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality.
- Design and construction of data warehouses for multidimensional data analysis and data mining
- View the debt and revenue changes by month, by region, by sector, and by other factors
- Access statistical information such as max, min, total, average, trend, etc.
- Loan payment prediction/consumer credit policy analysis
- feature selection and attribute relevance ranking
- Loan payment performance
- Consumer credit rating

DOI 10.17148/IJARCCE.2015.41258

Biological data analysis:

- DNA sequences: 4 basic building blocks (nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T).
- Gene: a sequence of hundreds of individual nucleotides arranged in a particular order
- Humans have around 30,000 genes
- Tremendous number of ways that the nucleotides can be ordered and sequenced to form distinct genes
- Semantic integration of heterogeneous, distributed genome databases
- Current: highly distributed, uncontrolled generation and use of a wide variety of DNA data
- Data cleaning and data integration methods developed in data mining will help

Telecommunication industry:

- A rapidly expanding and highly competitive industry and a great demand for data mining
- Understand the business involved
- Identify telecommunication patterns
- Catch fraudulent activities
- Make better use of resources
- Improve the quality of service
- Multidimensional analysis of telecommunication data
- Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc.

III. DATA MINING TECHNOLOGIES

The most commonly used techniques in data mining are:

Decision trees: A decision tree is a structure that includes a root node, branches, and leaf nodes. The decisions generate rules for the classification of a dataset.

Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

Artificial neural networks: ANN is generally presented as systems of interconnected "neurons" which exchange messages between each other. It is non-linear predictive models that learn through training.

Genetic algorithms: It is an adaptive procedure. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Nearest neighbor method: The Nearest Neighbor prediction technique is among the oldest techniques used in data mining. A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset.

Sometimes it is called the k-nearest neighbor technique. Nearest neighbor is a prediction technique that is quite similar to clustering.

IV. DATA MINING TOOLS

The development and application of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes increasingly difficult. Furthermore, we propose criteria for the tool categorization based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and license policies. Every tool has its own advantages and disadvantages. [1] Within data mining, there is a group of tools that have been developed by a research community and data analysis. They are offered free of charge using one of the existing open-source licenses. Data mining tools predict future trends, behaviors, allowing business to make proactive, knowledge driven decisions. [2]There are number of open source tools available for data mining .some of them are explained below.

(i) Weka [Waikato Environment for Knowledge Analysis]

Weka is a machine learning software written in java. It is free software. Several data mining tasks are supported by weka such as data processing, clustering, classification, regression, visualization. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling.

(A) Key Features:

- It is open source and freely available.
- It provides many different algorithms for data mining and machine learning.
- It provides flexible facilities for scripting experiments.
- It is platform-independent.
- It has kept up-to-date, with new algorithms being added as they appear in the research literature.
- It is easily useable by people who are not data mining specialists.

(B) Advantages:

- It is freely available under GNU General public license.
- There is very large collection of different data mining algorithms.
- It is also suitable for developing new machine learning schemes.[3]
- Weka software contain very graphical user interface, so the system is very easy to access.
- As weka is fully implemented in java programming language, it is platform independent and portable.

(C) Specialization:

- Weka is best suited for mining association rules.
- Stronger in machine learning techniques.
- Suited for machine Learning.

(ii) KEEL [Knowledge Extraction based on Evolutionary Learning]

It is an open source Java software tool that can be used for a large number of different knowledge data discovery

DOI 10.17148/IJARCCE.2015.41258

tasks. It is an application package of machine learning software tools. It contains a wide variety of classical knowledge extraction algorithms, preprocessing techniques (training set selection, feature selection, discretization, and imputation methods for missing values, among others), computational intelligence based learning algorithms, hybrid models, statistical methodologies for contrasting experiments and so forth. It allows performing a complete analysis of new computational intelligence proposals in comparison to existing ones. Moreover, KEEL has been designed with a two-fold goal: research and educational.

(a) Key Features:

- It contains pre-processing algorithms: transformation, discretization, instance selections and feature selections
- It also contains a Knowledge Extraction Algorithms library, supervised and unsupervised, remarking the incorporation of multiple evolutionary learning algorithms
- It has a statistical analysis library to analyze algorithms
- It includes a Java Class library for Evolutionary Computation (JCLEC)
- It contains an user-friendly interface, oriented to the analysis of algorithms
- The software is used through a web interface, sending to the user all the necessary information to perform the designed experiments in the computer that he wishes

(b) Advantages:

- It reduces programming work.
- It includes regression, classification, clustering, and pattern mining and so on.
- It contains a big collection of classical knowledge extraction algorithms, preprocessing techniques, Computational Intelligence based learning algorithms, including evolutionary rule learning algorithms based on different approaches, and hybrid models such as genetic fuzzy systems, evolutionary neural networks etc.[4]

(c) Specialization:

- Machine learning tool.
- Keel is a software tool used for Data Mining problems.

(iii) R [Revolution]

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and Mac OS. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. New technology and ideas often appear first in R.

(A) Key Features:

- R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.
- R is a well-supported, open source, command line driven, statistics package. There are hundreds of extra “packages” freely available, which provide all sorts of data mining, machine learning and statistical techniques.
- R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study.
- It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems

(b) Advantages:

- R is free and open source software.
- R is the most comprehensive statistical analysis package.
- It is a powerful elegant array.
- Ability to make a working machine learning program in just 40 lines of code.
- Numerical programming is better integrated in R.
- R has no license restrictions and so we can run it anywhere and at any time.
- The graphical capabilities of R are outstanding, providing a fully programmable graphics language that surpasses most other statistical and graphical packages.
- The validity of the R software is ensured through openly validated.
- Easier to combine with other statistical calculations.
- Import and export of data from spreadsheet is easier in R.

(c) Specialization:

- Suited for Statistical Computing.
- It has a large number of users, in particular in the fields of social science and bio-informatics.
- It is also a free ware replacement for SPSS.

V. DATA MINING FOR BIG DATA

Data mining is the process of finding correlations or patterns among dozens of fields in large relational database. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Technically, Data mining as a term used for the specific classes of six activities or tasks as follows:

a) Classification
b) Estimation
c) Prediction
d) Association rules
e) Clustering
f) Description

a) Classification

Classification [5] is a set of techniques which are aimed at recognizing categories with new data points. In contrast to cluster analysis, a classification technique uses training data sets to discover predictive relationships.

b) Estimation

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with

DOI 10.17148/IJARCCE.2015.41258

a value for some unknown continuous variables such as income, height or credit card balance.

c) Prediction
It's a statement about the way things will happen in the future, often but not always based on experience or knowledge. Prediction may be a statement in which some outcome is expected.

d) Association Rules
Association rule learning [6, 7] is set of techniques designed to detect valuable relationships or association rules among variables in databases.

e) Clustering
Cluster analysis [8, 9] is based on principles of similarities to classify objects. This technique belongs to unsupervised learning [10, 11] where training data [12] is used.

VI. FEATURES OF BIG DATA

Big data consists of number of features. They are:

- It is huge in size.
- Its data sources are from different phases.
- It is too much complex in nature, thus hard to handle.
- The data keep on changing time to time.
- It is free from the influence, guidance, or control of anyone.

This large storage of data requires large area for actual implementation.

VII. BIG DATA CHALLENGES

Current technology, architecture, management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data. Data have number of challenges that are related to its complexity which are:

- How to Understand the Unstructured Data.
- Capturing of important data
- How to store, analyze and understand the data.
- Most important challenge is Privacy and security

VIII. BIG DATA VISUALIZATION

Big Data visualization is not as easy as traditional small data sets. Visualization can be thought of as the "front end" of big data. The In existing tool there are challenges regarding Human perception means not proper understanding of data and limited screen space not proper visibility of objects properly. Due to which data is not helpful in decision making and taking required information. Structured and unstructured data is taken fir the visual analysis and converted into visualization form and that help in gaining knowledge and making choices. Visualization used the three important terms: Data, Information and Knowledge. Visualization approaches are used to create tables, diagrams, images, and other intuitive display ways to represent data. In large-scale data visualization, many researchers use feature extraction and geometric modeling to greatly reduce data size before actual data rendering. Choosing proper data representation is also very important when visualizing big data [13].

IX. RELATED WORK

According to [14] relevant Big Data Visualization methods classification and have suggested the modern tendency towards visualization-based tools for business support and other significant fields. The approach of utilizing VR, AR and MR for Big Data Visualization is presented and the advantages, disadvantages and possible optimization strategies of those are discussed. For visualization problems discussed in this work, it is critical to understand the issues related to human perception and limited cognition. Only after that, the field of design can provide more efficient and useful ways to utilize Big Data. It can be concluded that data visualization methodology may be improved by considering fundamental cognitive psychological principles and by implementing most natural interaction with visualized virtual objects.

According to [15] presents the Mondrian data visualization software. In addition to standard plots like histograms, bar charts, scatter plots or maps, Mondrian offers advanced plots for high dimensional categorical (mosaic plots) and continuous data (parallel coordinates). All plots are linked and offer various interaction techniques. A special focus is on the seamless integration of categorical data. Unique is Mondrian's special selection technique, which allows advanced selections in complex data sets.

From [16] described main problems of Big Data visualization and approaches of how we can avoid them. A review on existing methods for data visualization in application to big data and taking into account the described problems. Summarizing the result, have provided a classification of visualization methods in application to Big Data.

X.CONCLUSION

Big Data is going to continue growing. This data is going to be more diverse, larger, and faster. We discussed some insights about the topic, and what we consider are the main concerns and the main challenges for the future. Big Data mining will help us to discover knowledge that no one has discovered before. Future works in this field can be held in the following areas: making decisions and recommendation for visualization method selection for concrete Big Data classes, formalization of requirements and restrictions for visualization methods applied to one or more Big Data classes and research of visualization methods applicability for different scales.

REFERENCES

[1]. S.R.Mulik, S.G.Gulawani. "Performance comparison of Data Mining tools in mining Association Rules", International Journal of Research in IT, Management and Engineering (IJRIME), Volume 1, Issue 3, ISSN: 2249- 1619.

[2]. Ralf Mikut and Markus Reischl Wiley Inter disciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011.

[3]. Witten, I.H., Frank, E. "Data Mining: Practical machine Learning tools and techniques", 2nd addition, Morgan Kaufmann, San Francisco(2005).

DOI 10.17148/IJARCCE.2015.41258

[4]. Alcala-Fdez, J.,L., del Jesus, M.J., Ventura, s., Garrell, J.M, Otero, J., Romero,C., bacardit, j., Rivas, V.M., Fernandez, J.C., Herrera., F. "KEEL: A software tool to Assess Evolutionary Algorithms to Data mining Problems", Soft computing Volume 13, Issue3, pages 307-318(2009).

[5]. Aggarwal CC. Data classification: algorithms and applications. CRC Press; 2014.

[6]. Zhang C, Zhang S. Association rule mining: models and algorithms. Springer; 2002.

[7]. Cleophas TJ, Zwinderman AH. Machine learning in medicine: part two. Machine learning in medicine: Springer; 2013.

[8]. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis, vol. 344. John Wiley and Sons; 2009.

[9]. Anderberg MR. Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks, vol 19, Academic press; 2014.

[10]. Hastie T, Tibshirani R, Friedman J. Unsupervised Learning. Springer; 2009.

[11]. Fisher DH, Pazzani MJ, Langley P. Concept formation: knowledge and experience in unsupervised learning. Morgan Kaufmann; 2014.

[12]. McKenzie M, Wong S. Subset selection of training data for machine learning: a situational awareness system case study. In: SPIE Sensing Technology + Applications. International Society for Optics and Photonics; 2015.

[13] C.L. P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, 275 (10), August 2014, pp. 314-347.

[14]. Olshannikova et al. Journal of Big Data (2015) "Visualizing Big Data with augmented and virtual reality: challenges and research agenda"

[15]. "Interactive Data Visualization using Mondrian" by Martin Theus, University of Augsburg, Department of Computer oriented Statistics and Data Analysis,. 14, 86135 Augsburg, Germany, martin.theus@math.uni-augsburg.de

[16]. Evgeniy Yur'evich Gorodov and Vasiliy Vasil'evich Gubarev "Analytical Review of Data Visualization Methods in Application to Big Data" Journal of Electrical and Computer Engineering Volume 2013 (2013)

DOI 10.17148/IJARCCE.2015.41258