International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey on Big Data Mining Algorithms

S. Sivasankar[1], T. Prabhakaran[2]

[1]PG Scholar, Department of Computer Science and Engineering, Nandha Engineering College, Erode, Tamilnadu, India

[2]Assistant Professor, Department of Computer Science and Engineering, Nandha Engineering College, Erode, Tamilnadu, India

Abstract: ***In recent years with the explosive development of internet the size of data has grown a large and reached petabytes size. Bigdataisan immense collection of both structured and unstructured data. Due to its large size discovering knowledge or obtaining pattern from big data within an elapsed time is a complicated task. A number of algorithmic techniques have been designed for big data mining in an effective manner. The various mining algorithms like Two-Phase Top -Down Specialization approach (TPTDS), Tree-Based Association rules (TARs), FuzzyC – Means (FCM)algorithm and Associate Rule Mining (ARM) algorithm are surveyed in this paper and the results obtained are compared and evaluated by the parameters such as execution time, information loss and extraction time.***

Keywords: Big data, TPTDS, TAR, FCM, ARM**.**

1. Introduction

Big data is a phrase given to the data set with vast size and composite so that it becomes hard to process with the common data processing applications. The widedata sets are due to the additional information derivable from analysis of a single large set of related data. Big data can be characterized by volume, velocity and variety. Volume is due to transaction based data stored through the years, unstructured data from social media and increasing amounts of sensor and machine-to-machine data being collected. Velocity deals with data streaming in at unique speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations. Variety meant by that processing of data in all formats like numeric data in structured and also in unstructured data contains photos, graphic images, videos, web pages, portable document formatand word processing documents. The challenges in big data are capture, curation, storage, search, sharing, transfer, analysis which includes discovering knowledge or obtaining pattern and visualization.

2. Related Work

During the big data mining instead of miningonly the data, basic features like privacy preservation in electronic health records and financial transaction records should be preserved by this for personal data it ensures the security for the users data and this is done by the data anonymization [1]. Big data mining also deals with the structure by which the mined results are stored and from the results user can get the central idea and answers to the queries easily the data set returned for the answer to queries are also too large [3]. Clustering is also the basic task done in big data mining for the knowledge discovery or obtaining patterns the efficiency of three already available techniques is extended to use for very large data [4]. Mining from Intensive Care Units is the important task because it provides information on patient diseases, conditions, and the medicines to be provided which is very important because from this health condition of the patient can be find out and it provides easy way of finding out the disease [11].

3. Algorithms Used For Big Data Mining

Big data mining is used to discover knowledge or extract pattern from the large data set or big data set which means that the size of the data set is very largeso that it provides the useful information the various algorithms are used for big data mining and in this paper the algorithms surveyed are Two-Phase Top-Down Specialization (TPTDS) approach, Tree-Based Association Rules(TARs), Fuzzy C-Means (FCM) algorithm and Associate Rule Mining (ARM) algorithm, and comparison of these algorithms each algorithm is used to mining from different sources.

3.1 Two-phase top-down specialization (TPTDS)

Two-Phase Top-Down Specialization approach has three components namely data partition, anonymization level merging and data specialization. The two phases of this approach are job level and task level and in both the phases parallelization is achieved. The phases are based on the MapReduce on cloud. Job level parallelization means that multiple MapReduce jobs can be executed simultaneously to make full use of cloud infrastructure resources. Task level parallelization refers to that multiple mapper/reducer tasks in a MapReduce job are executed simultaneously over data splits. To achieve high scalability parallelization of multiple jobs on data partitions in the first phase is done. Finally consistent anonymous data set is obtained by integrating the intermediate results and anonymizes entire data sets. TPTDS approach is mainly used for mining large amount of data with privacy preservation to be achieved.In TPTDS approach calculation is doneby TDS in a useful manner whereas TDS is a continuous process and the metric for information and privacy is information gain per privacy loss (IGPL). Top Down Specialization consists of three steps namely finding the best specialization, performing specialization and updating values of the search metric. MapReduce version of centralized TDS (MRTDS) is a subroutine which performs the calculation required in Two-Phase Top-Down Specialization and it makes the full use of the job level parallelization of MapReduce. MRTDS

anonymize data partitions to generate intermediate anonymization levels information gain after performing specialization and privacy loss specialization can be computed by statistical information derived from data sets.The basic idea of TPTDS is to gain high scalability by making a tradeoff between scalability and data utility and it is done by slight decrease of data utility so that it can lead to high scalability[1].

3.1.1 Data partition

Data set D is partitioned into smaller ones $D_{i,}$it is required that the distribution of data records in D_i is similar to D. A data recordhere can be treated as a point in an m-dimension space, where m is the number of attributes, the partitioned data can be grouped together to achieve better anonymization level merging. Intermediate anonymization levels can be derived from the partitioned data and partition of data takes place through Random sampling technique. Data partition takes place in the first phase.

3.1.2 Anonymization level merging

Anonymization meant by the encryption of data which is useful for the privacy preservation. In the second phase Anonymization Level Merging occurs in this all the intermediate levels which are partitioned in data partition are merged. Merging of anonymization levels is done by the merging cuts. Anonymization can be performed by specialization operation which is to replace domain value with its entire child value. Anonymization level can represent the anonymization degree of a data set, the more specific anonymization level a data set has it corresponds with less degree of anonymization. The more general one is selected as the merged one to ensure the merged intermediate level anonymization. In the multiple levels anonymization merge can occurs by the same way in iterative manner.

3.1.3 Data specialization

Data specialization takes place at last and which is done for a data set D for anonymization in a one pass MapReduce job. Final anonymization level is getting by MapReduce version of centralized TDS (MRTDS), by running it on the entire data set Dwhich performs computation in TPTDS. Specialization process is track and manages by the anonymization level.Then, the data set D is anonymized by replacing original attribute values in D with the corresponding domain values. The Map function provides anonymous records with its count and the Reduce function adds the anonymous records and counts their number. An anonymous record and its count represent a Quasi-identifier group. The QI-groups constitute the final anonymous data sets. Quasi-identifiers, representing groups of anonymous records, can lead to privacy breach. Specialization operation performs the replacement of values and here the domain values are replaces with all its child values.

3.1.4 Advantage of TPTDS

Two-Phase Top-Down Specialization approach has an advantage that it anonymizes large scale data sets using the MapReduce framework on cloud.

3.1.5 Limitation of TPTDS

Despite its well usage for the privacy preservation on privacy sensitive large scale data sets, Two-Phase Top-Down Specialization approach hasa limitation that it cannot provide privacy preservation for the data set with large scalability so that scalable privacy preservation aware analysis to be done.

3.2 Tree-based association rules (TAR)

Tree-Based Association Rules is used for mining the Extensible Markup Language (XML) documents and the obtained results can be stored in XML formats this mined data provides the central idea which includes the content as well as the structure of the XML document and answers to queries raised. Association rules describe the co-occurrence of data items in a large amount of collected data. The quality of an association rule is measured by means of support and confidence.Support corresponds to the frequency of the set X and set Y in the data set with the union operation where confidence corresponds to the conditional probability of finding Y, having found XThe association rule is extended in the context of relational databases to make it adapt for hierarchical nature of XML documents. Relationships among subtrees of XML documents are to be finding with the textual contents of leaf elements and the value of attributes [3].

3.2.1 Basic concepts

Induced subtree and rooted subtree are the two types of subtrees can be obtained for an XML document. Tree Based Association Rules also provide information about the structure of frequent portions of XML documents in addition to associations between data values; thus, they are more expressive than classical association rules which only provide frequent correlations of flat values.Structure TAR (sTAR) provides information only on the structure of the document whereas instance TAR (iTAR) provide information both on the structure and on the data values contained in a document.

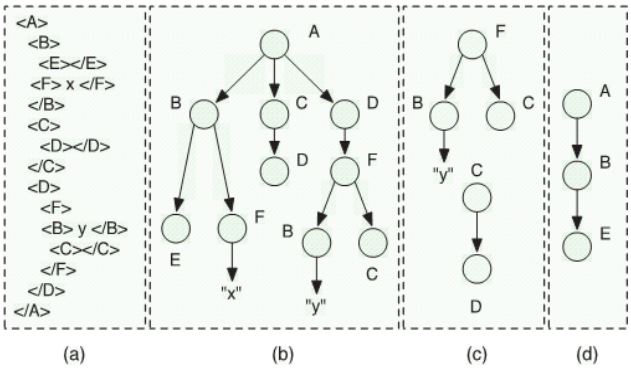

Figure 1:(a) – XML document(b) - Tree- Based Representation(c) - Induced subtrees (d) - Rooted subtree

From the observation of sTARs users can able to know the structure of an XML document, and thus use this approximate schema to formulate a query when no DTD or schema is available. sTARs do not show all possible paths in the XML document but only the frequent paths. sTARs provide a simple path matching and can be used for the optimization of the query process. An index for an XML

data set is a predefined structure whose performance is maximized when the query matches exactly the designed structure, the goal when designing an index is to make it as similar as possible to the most frequent queries.Instance TAR (iTARs) give an idea about the type of content of the different nodes. Instance TAR gives the good knowledge for nodes.

3.2.2 TAR extraction

TAR extraction is the process of extracting knowledge from the XML document by Tree-Based Association Rules. It consists of two steps the first step is mining frequent subtrees meaning that subtrees with a support above a user defined threshold from the XML document and the second step is computing interesting rules which means rules with a confidence above a user defined threshold from the frequent subtrees. After the mining process has finished and frequent TARs have been extracted, they are stored in XML format.Thereason for using TARs instead of the original document is that processing iTARs for query-answering is faster than processing the document. To take full advantage of this, indexes on TARs is introducedtofurther speed up the access to mined trees, path indexes are proposed to quickly answer queries that follow some frequent path template, and are built by indexing only those paths having highly frequent queries. Once the algorithm is applied to all TARs the resultis a set of trees whose nodes contain references to one or more rules and which are stored on XML file.Tree Based Association Rules extraction is an important and useful process.

3.2.3 Advantages of TAR

Tree – Based Association Rules have several advantages which are mining all frequent association rules without imposing any prior restriction on the structure and the content of the rules, storing the mined information in XML format and using the extracted knowledge to gain information about the original data sets.

3.2.4 Limitation of TAR

Updatability of the document storing TAR during the change occurs in original XML data sets and as well as their index is a limitation for Tree-Based Association Rule.

3.3 Fuzzy c-means (FCM)

Fuzzy C-Means algorithmis used for clustering of very large data for the process of mining from labeled data, large image data and Unloadable data. FCMalgorithm is extended to be applied for very large data. The two main approaches for clustering in VL data are distributed clustering which is based on various incremental styles and clustering a sample found by either progressive or random sampling. Approaches provide useful ways to accomplish two objectives such as acceleration for loadable data and approximation for unloadable data. Fuzzy partitions are more flexible than crisp partitions in that each object can have membership in more than one cluster. Sampling methods compute cluster centers from a smaller sample of selected objects. Single pass algorithms sequentially load small groups of data, clustering these manageable chunks in a single pass and then combining the results from each chunk. Data transformation algorithms alter the structure of the data so that data can be more efficiently accessed usually data take the structure of graph structure. The other approaches for optimizing the FCM model are genetic algorithms and particle swarm optimization. FCM is initialized by choosing c objects randomly from the dataset to serve as the initial cluster centers, the algorithm terminates when there are only negligible changes in cluster center locations [4].

3.3.1 Sampling and non-iterative extension

The most basic way to address VL data is to sample the data set and then use Fuzzy C-Means algorithm to compute cluster centers of the sampled data. If the data were sufficiently sampled, the error between the cluster center locations produced by clustering the entire data set and the locations produced by clustering the sampled data should be small. Extension can be used to compute the full fuzzy data.

3.3.2 Algorithms based on weighted fuzzy c means

In Large FCM (LFCM), each object is considered equally important in the clustering solution. The weighted FCM (wFCM) model introduces weights that define the relative importance of each object in the clustering solution. The Single-Pass Fuzzy c-Means (spFCM) iterates over the remaining subsets of data in X, and at each iteration the value of the weighted fuzzy at each iteration wFCM is used to cluster an augmented set of data that is composed of the union of the cluster centers and the sample subset X_1, spFCM computes the new cluster centers by feeding forward the clustercenters from the previous iteration into the data being clustered. Online Fuzzy c-Means (oFCM) clusters all is subsets of objects separately and then aggregates the sets of cluster centers at the end. Bit-Reduced Fuzzy c-Means(brFCM) was designed to address the problem of clustering in large images. Image data can be bit reduced by removing the least significant bits and binning the identical objects. Approximate Kernel Fuzzy c-Means(akFCM)algorithm approximates a solution to kFCM by using a numerical approximation of the cluster center to object kernel distance in that is based on the assumption that the cluster centers can be accurately represented as a linear sum of a subset of the feature vectors.

3.3.3 Advantage of FCM

The advantage of Fuzzy C-Means algorithm is the clustering of very large data or big data.

3.3.4 Limitations of FCM

Fuzzy C - Means algorithm has several limitations as developing and investigating scalable solutions for VL fuzzy clustering. To examine where kernel solutions can be used, is it possible to use cluster validity indices to choose the appropriate kernel. Quality of clustering is measured by requiring full access to the objects vector data.

3.4 Associate rule mining (ARM)

Associate Rule Mining (ARM) is the mining algorithm used to mine from the clinical data. Associate rule mining is a method to reveal meaningful relations between variables in data bases. ARM has been widely adopted in applications such as heart disease prediction, healthcare auditing and neurological diagnosis in hospitals. An item can be

numerical or categorical depending on the data type of the variable. Two important metrics aresupportand confidence whichevaluate the frequency and level of association of a rule.In order to discover frequent and confident association rules, the mining process requires users to specify two minimum values as thresholds to drop infrequent and unconfident rules, which are minimum support ($Supp_{min}$) and minimum confidence ($Conf_{min}$). Rules are considered to be frequent if their supports are at least $Supp_{min}$and confident if their confidences are at least$Conf_{min}$. The goal of ARM is to findall frequent and confident rules based on these two users specified values [11]

3.4.1 Advantage of ARM

Associate Rule Mining algorithm has an advantage that it generate quantitative and real time decision support rules for Intensive Care Unit it is done by predicting the characteristics of ICU stay.

3.4.2 Disadvantage of ARM

Manual specification of variables for interest and cut points by clinicians is the limitation of ARM, development of automatic feature selection can be done.

4. Comparison of Big Data Mining Algorithms

Big data mining algorithms are compared based upon the important criteria for mining such as the speed or the timely manner in which the knowledge discovery or pattern extraction occurs that is run time or extraction time and depending upon the size of data and the content of data the run time will surely differs.

Table 1:Comparison of algorithms

Algorithm/ Approach	Performance Criteria	Usage
Two-Phase Top-Down Specialization (TPTDS) approach	Execution time and Information Loss	Privacy Preservation of data
Tree-Based Association Rules (TAR) approach	Extraction time and Answer time	Mining from semistructured (XML) document
Fuzzy c-Means (FCM) algorithm	Run time	Clustering of data
Associate Rule Mining (ARM)	comorbidity	Mining from ICU(clinical) data

The table shows the algorithm comparison.Size of data includes Megabyte, Gigabyte, Terabyte etc…., and the content of data includes both structured and as well as unstructured data. Each of the algorithms satisfies the different purpose, the algorithm done the mining process from the different data sources their performance criteria or factors and usage are listed in the table.

5. Conclusion

There are several big data mining algorithms such as Two-Phase Top-Down Specialization approach (TPTDS), Tree-Based Association rules (TARs),Fuzzy c–means (FCM) algorithm and Associate rule mining (ARM) algorithm are surveyed in this paper and each algorithm has used for mining from different data sets and they satisfy the different purpose or usage. These algorithms have their advantage as well as disadvantage too. The mining algorithms for big data helps to discover knowledge or extract patterns but thealgorithm suffers at the time of the scalability of data takes place.

References

[1] Xuyun Zhang, Laurence T.Yang, Chang Liu and Jinjun Chen, “A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud”, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, vol .25, no.2, FEBRUARY 2014.
[2] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, “Data Mining with Big Data”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol.26, no.1, JANUARY 2014.
[3] Mirijana Mazuran, Elisa Quintarelli and Letizia Tanca, “Data Mining for XML Query-Answering Support”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol.24, no.8, AUGUST 2012.
[4] Timothy C.Havens, James C.Bezdek, Christopher Leckie,Laewrence o.Hall and Marimuthu Palaniswami, “Fuzzy c-Means Algorithms for Very Large Data”, IEEE TRANSACTIONS ON FUZZY SYSTEMS, vol .20, no.6, DECEMBER 2012.
[5] Jialei Wang, Peilin Zhao, Steven C.H.Hoi and Rong Jin, “Online Feature Selection and Its Applications”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol.26, no.3, MARCH 2014
[6] Tias Guns, Siegfried Nijssen and Luc De Raedt, “k-Pattern Set Mining under Constraints”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol.25, no.2, FEBRUARY 2013.
[7] Marc Sole and Josep Caroma, “Region Based Foldings in Process Discovery”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol.25, no.1, JANUARY 2013.
[8] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic and Mirjina Ivanovic, “The Role of Hubness in Clustering High-Dimensional Data”,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol.26, no.3, MARCH 2014.
[9] Liang Wang, Xin Geng, James Bezdek, Christopher Leckie and Kotagiri Ramamohanarao, “Enhanced Visual Analysis for Cluster Tendency Assessment and Data Partitioning”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol.22, no.10, OCTOBER 2010.
[10]Evan Wei Xiang, Bin Cao, Derek Hao Hu and Qiang Yang, “Bridging Domains Using World Wide Knowledge for Transfer Learning”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol.22, no.6, JUNE 2010.
[11]Chih-Wen Cheng, Nikhil Chanani, Janani Venugopalan, Kevin Maher and May Dongmei Wang, “icuARM-An ICU Clinical Decision Support System Using Association Rule Mining”, IEEE

Journal of Translational Engineering in Health and Medicine, vol.1, DECEMBER 2013.
[12] Dragos Bratasanu, Ion Nedelcu and Mihai Datcu, " Bridging the Semantic Gap for Satellite Image Annotation and Automatic Mapping Applications", IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, vol.4, no.1, MARCH 2011.
[13] Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo and Chengqi Zhang, "Combined Mining: Discovering Informative Knowledge in Complex Data", IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS", vol.41, no.3, JUNE 2011.
[14] Yunpeng Chai, Zhihui Du, David A. Bader and Xiao Qin, "Efficient Data Migration to conserve Energy in Streaming Media Storage Systems", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, vol.23, no.11, NOVEMBER 2012.
[15] Aliaksei Makarau, Gintautas Palubinskas and Peter Reinartz, "Alphabet-Based Multisensory Data Fusion and Classification Using Factor Graphs", IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, vol.6, no.2, APRIL 2013.
[16] Frieder Ganz, Payam Barnaghi, and Francois Carrez, "Information Abstraction for Heterogeneous Real World Internet Data", IEEE SENSORS JOURNAL, vol.13, no.10, OCTOBER 2013.
[17] Alexander Rind, Tim Lammarsch, Wolfgang Aigner, Bilal Alsallakh and Silvia Miksch, "TimeBench: A Data Model and Software Library for Visual Analytics of Time-Oriented Data", IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL.19, NO.12, DECEMBER 2013.
[18] Zhipeng Tan, Wei Zhou, Dan Feng and Wenhua Zhang, "ALDM: Adaptive Loading Data Migration in Distributed File Systems", IEEE TRANSACTIONS ON MAGNETICS, VOL.49, NO.6, JUNE 2013.
[19] Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu- Thu Le, Le-Minh Nguyen, Susumu Horiguchi and Quang-Thuy Ha, "A Hidden Topic-Based Framework toward Building Applications with Short Web Documents", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.23, NO.7, JULY 2011.
[20] Thanh Tran, Gunter Ladwig and Sebastian Rudolph,"Managing Structured and Semistructured RDF Data Using Structure Indexes", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.25, NO.9, SEPTEMBER 2013.

Author Profile

S. Sivasankar received the B.Tech degree in Information Technology from Nandha College of Technology in 2012. He is currently doing his M.E degree in Computer Science and Engineering in Nandha Engineering College.

T. Prabhakaran received the B.Tech degree in Information Technology from Kongu Engineering College in 2006 and the M.E degree in Computer Science and Engineering from Nandha Engineering College in 2010. He is currently working as an Assistant Professor in Nandha Engineering College. His research interests include Cloud Computing, Privacy and Big Data.