Data Mining and Knowledge Discovery for Big Data in Cloud Environment

Raid Abd Alreda Shekan
Researcher Lecture, Specialization Computer Science (Information System), University of Babylon, Iraq. E-mail: pure.raed.abd@uobabylon.edu.iq

Ahmed Mahdi Abdulkadium
Researcher Lecture, Specialization Computer Science (Information System), AL-Qasim Green University, Iraq. E-mail: ahmed_mahdi@uoqasim.edu.iq

Hiba Ameer Jabir
Researcher Assistant Lecture, Specialization Computer Science (Multimedia), University of Babylon, Iraq. E-mail: hiba.ameer1976@gmail.com

Received April 06, 2021; Accepted July 14, 2021
ISSN: 1735-188X
DOI: 10.14704 WEB/V18SI04/WEB18186

Abstract

In past few decades, big data has evolved as a modern framework that offers huge amount of data and possibilities for applying and/or promoting analysis and decision-making technologies with unparalleled importance for digital processes in organization, engineering and science. Because of the new methods in these domains, the paper discusses history of big data mining under the cloud computing environment. In addition to the pursuit of exploration of knowledge, Big Data revolution gives companies many exciting possibilities (in relation to new vision, decision making and business growths strategies). The prospect of developing large-data processing, data analytics, and evaluation through a cloud computing model has been explored. The key component of this paper is the technical description of how to use cloud computing and the uses of data mining techniques and analytics methods in predictive and decision support systems.

Keywords

Cloud Computing, Big Data, Machine Learning, Data Mining, Classification, Knowledge Discovery, Data Science, Prediction.
Introduction

1. General

Some have coined the term "information age" for the current era because of the widespread growth of technology and electronic data processing techniques. Thus, we are living in the "information age", wherein a variety of huge volumes of data are generated (Singh AN, Rath RK., 2015), (Kokila PM, Saravanan P, Jagadeesan B, Sharmila R., 2016). Data-intensive technologies are continuously being built and utilized in various scenarios. By far the most significant companies that generate the most data on web databases, like Google+, Twitter, LinkedIn, Facebook, Walmart and Taobao. Generally speaking, computer-intensive software, eBay, Amazon, and cloud-based computing data.

Terabytes or even petabytes of data are generated nearly everyday, leading to a high amount of data; the arriving data is transmitted at a very significant speed and also has real-time specifications requiring fast velocity; a broad variety of structured, semi-structured as well as unstructured data can be met; in order to preserve veracity and value, these variable data must be cleaned prior integrating into the process (Hashem IA, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU., 2015), (Xu J, Huang E, Chen CH, Lee LH., 2015). This 6V attribute is amongst the most widely used descriptions of the Big Data challenge, that's now a major subject for academic science as well as businesses (Figure 1).

![6V Attributes of data processing and manipulation](http://www.webology.org)

**Figure 1 6V Attributes of data processing and manipulation**

a. Big Data

Recently, enormous data volumes in different areas are continuously produced and stored in the large capacity storage devices/vaults. In addition, emerging innovations are growing tremendously in information explosion, whilst developing a need to examine the large amounts of existing data (Zanoon N, Al-Haj A, Khwaldeh SM., 2017). Not only is there a vast amounts of data information, but all the data are continuously being modified and added to them. Furthermore, a broad number of data formats as well
as repositories are available. Text, graphics, animations, videos, recordings, audio and variations are included in the collection of raw data. This type of data in large volumes is known as "Big data" (Agrawal D, El Abbadi A, Ooi BC, Das S, Elmore AJ, 2012). In its simples from Big data are those sets of data that, due to their complexity and dynamic nature, are very hard to manage by merely utilizing data-processing software and applications (Wu C, Buyya R, Ramamohanarao K., 2016), (Neaga I, Hao Y., 2014).

b. Benefits and Utilization of Big Data

Big data may be useful in resolving crucial issues and offer new perspectives for improving clinical researches, particularly in cancer science, world security, transport and logistics, detect and forecast terrorism, as well as resolve environmental and socio-economic problems. In the logistics industry, technical and analytical progress of large datasets is extremely beneficial. Data analytics and big data mining are also decentralized as well as their scale and confidentiality concerns illustrate the scope of approaches. Computer-based machine learning (ML) techniques are thus, important and needed because of enormous amounts of datasets that is an exception for manual commits. The big data is stored on various machine and networked systems. Consequently, the advantages of big data mining and device diversity impose difficulties.

c. Data Mining

The key aspect of big data is data mining. Data mining is immensely useful in gathering specific data from diverse data sources. It is also immensely useful to complete unique tasks. Cloud-based data mining comprises of the discovery of legitimate and logically valuable factors and new patterns. It has been developed by a sequence of dynamic and incremental steps, in which data can be preprocessed, patterns with an almost clear characterization are searched and several patterns that are interpretable are visualized (Gan W, Lin JC, Chao HC, Zhan J., 2017). Data science and data mining application approaches and practices are generally attributed in relevant business disciplines and analytical studies.

Literature Review

A variety of parallel and distributed algorithms for data mining were discussed in literature in order to accomplish systematic big data mining. Hamrouni, Slimani and Charrada (Hamrouni T, Slimani S, Charrada FB., 2015) have published a very comprehensive bibliography of concurrent data mining algorithms that includes four
broad types of decentralized data mining frameworks, including apriori algorithms, sorting, clusters and streams of data for data mining.

The writers in (Diamantopoulos T, Symeonidis A., 2018) explore how power grid firms are enabled by sophisticated algorithms to make decisions to predict and eliminate faults in the transmission line. The authors thus proposed a hybrid classification, consisting of a cache-based classification as well as a main decision tree classification. The resulting classification is built into a streaming data unit above the decision support tool to translate high speed sampling data into valuable information. In contrast with existing systems, empirical findings suggest that the classification presented by the authors is more credible.

In the context of an Salient Semantic Analysis (SSA)-driven retrieval approach, the scholars in (Gheware SD, Kejkar AS, Tondare SM. 2014) discuss the problem of productivity of Salient Semantic Analysis (SSA) with a view to improving data collection efficiencies. The SSA paradigm is introduced in the first stage for modeling text and query descriptions. These definitions and bag-of-word (BOW) representations are then utilized in conjunction to evaluate requests and documentation in communicative language teaching. Numerous standardized text retrieval conference (TREC) libraries test the suggested solution. Experiments on generic TREC compositions indicate that the models introduced continuously surpass the current methods of recovery that are based on Wikipedia models.

The data mining cloud computing framework context was developed for the effective delivery of data mining feature implementations in a cloud computing environment (Gan W, Lin JC, Chao HC, Zhan J, 2016). A collection of parameter-spreading classification and clustering frameworks have been used to apply the context on the Windows Azure platform. Classifying and clustering utilizing Hadoop including sector/sphere approaches are data mining methods.

In another study (Gheware SD, Kejkar AS, Tondare SM., 2014) the association rules mining procedures have been explored based on big data in cloud computing. The primary aspect was cloud computing, Hadoop, Map Reduce, Apriori and parallel data and mining association rule algorithms. A parallel rule mining model is adopted which adapts to the actual environment of cloud computing. This involves the methods for division of data, the process for allocating datasets, an updated Apriori algorithm as well as the protocol for applying the enhanced earlier algorithms upon Map Reduce.
Methods

The current increase in data is attributed to the following factors: (1) the constantly gathering of data by thousands of apps, like smartphone sensors, social networking sites including mobile applications. (2) the scope for data processing has increased, thus, rendering it much easier than ever before to gather data, making it convenient to buy additional storage instead of choosing whether to delete; (3) Computer based ML techniques and data extraction methods have advanced significantly in recent years, allowing greater efficiency and performance in retrieving information from the dataset to be obtained. Moreover, Hadoop, Hadoop++, Google MapReduce, Spark, Haloop, Twister, etc., are being used as big data and cloud computing platforms for manipulating massive data and performing complex computations. The databases in cloud computing are used for storing large structured and semi-structured data from various applications. Hadoop DB, Hbase and Big Table are the most popular cloud databases. The data warehouse management is also essential in order to introduce an effective Big Data analytics and data mining system. The Pig and Hive are the two important tools for data warehouse management and analysis. Here we will discuss a few of the popular techniques under cloud environment.

Cloud Computing

Cloud computing is a kind of computation focused on the Internet. In order to provide applications from vendors to the customer with vast quantities of data based on industry models, cloud computing has been developed as a service focused computing paradigm to distribute technology and frameworks and application. It supplies PCs and other devices with common tools and data on demand.

1. Private cloud: This approach works with an organizations in particular. Providing more security and command.
2. Public cloud: The general public is equipped with this framework. Pay-by-use or freemium infrastructure, therefore, this model is convenient and affordable.
3. Community cloud: many organizations use the infrastructure and only certain entities are allowed to use it. The technology can be managed and operated by the private organization or by a supplier of cloud services.
4. Hybrid cloud: this is a combination of 2 or 3 cloud setups, like private, public and hybrid/community. It may include associations, accomplishments and/or dedicated cloud-based services.
Cloud platforms include a variety of virtual machines (VMs) and are designed to handle big data, with an extent to where computation can be divided into various parallel operations.

**Knowledge Discovery in Cloud Computing**

KDD[18] is an approach to retrieve new information from complex databases. KDD (Gao K, Wang Q, Xi L., 2014) refers back to a variety of operations. Knowledge discovery in its simplest form represents a non-trivial retrieval of valuable information from huge databases of indirect, previously unrecognized and theoretically usable data. Whilst KDD and data extraction are quite often viewed as simile, data mining is simply a component of the KDD-based method of information visualization. Thus, data mining as well as KDD are also used synonymously as the core ingredient in the KDD method is data mining (Hassan MK, El Desouky AI, Elghamrawy SM, Sarhan AM., 2018). The exploration of knowledge in a database process consists of many phases that progress from raw collected data to new retrieval of information. Process of knowledge discovery may include intermediate components: integration of data, selection of data, cleaning of data, data transformation, pattern searches, i.e. data mining, pattern evaluation, interpretation, analyzation, evaluation findings as well as representation of knowledge. Figure 2 shows a standard process for knowledge discovery (Xu J, Huang E, Chen CH, Lee LH., 2015).

![Figure 2 Process of Knowledge Discovery in Database](http://www.webology.org)
Data Mining in Cloud Computing

The mining of data is part of the method of information finding (Kokila PM, Saravanan P, Jagadeesan B, Sharmila R., 2016). Data Mining and Big Data Analytics are common for cloud computing. Cloud Infrastructure data aggregation helps companies to centralize information and data storage management and ensure that customers have efficient, secure and protected resources. Customers can retrieve meaningful data and valuable information from virtually automated data warehouses by applying data mining techniques under cloud computing environment which minimizes storage and infrastructure costs.

The existing databases hold so much data that important knowledge for decision-making cannot be evaluated manually. The need to automate the extraction of valuable information from vast volumes of data is commonly understood and contributes to an advanced research and finding tool market which is growing rapidly. The exploration of intelligence and data retrieval are methods for discovering strategic data contained in very huge databases. In an attempt to derive useful information and expertise, researchers, informaticians, political scientists, analysts, economists, sociologists, mathematicians, and other researchers seek access to vast volumes of data. Analyzing the raw data and providing derived high-quality information to researchers or policy makers is achieved through automatic discovery software, instead of manual digging by the analysts.

Data Mining Tools

Various data mining tools and techniques include (Siddiqui T, Ahmad A., 2018), (Gheware SD, Kejkar AS, Tondare SM., 2014):

2. *WEKA*: It is an instrument based on java. Use free. Use free. It involves techniques of association, sorting, clustering, simulation and modelling.
3. *ORANGE*: The tool is based on Python. It promotes computer education, interpretation of data content and text classification.
4. *SAS*: This offers model dimensional forecasting and visualizations.
5. *R-Programming*: The FORTRAN- and C-based tools for r-programming. It helps data mining to detect, interpret, clusters and analyze data in time series approach.
6. *KNIME*: Tool help Interface. This data Analysis tool is prominent under linux applications.
The techniques comprise mathematical simulations, math algorithms and approaches of machine learning (algorithms that enhance their performance efficiencies automatically via learning, like the artificial neural network or decision tree models). Data mining thus requires more than data acquisition and management; it also involves the study and forecasting of favorable results.

**Data Mining Techniques**

Data mining in Cloud Computing enables businesses to centralize software and data storage management to ensure their customers enjoy effective, consistent and stable service (Petre RS., 2012), (Holzinger A, Dehmer M, Jurisica I, 2014). The data mining effort is usually used to construct a comprehensive descriptive framework of data mining or a statistical model of data mining. In particular, the descriptive data mining paradigm is used to describe the overall data features of a dataset stored in the repository. The predictive mining framework speculates existing data to make forecasts (Figure 3) (Wu C, Buyya R, Ramamohanarao K., 2016).

![Figure 3 Methods for Data mining in Cloud Computing environment](http://www.webology.org)

A number of data mining methods as seen in Table 1 (Neaga I, Hao Y., 2014) would accomplish the objective of a descriptive and predictive models. These methods may be utilized to find the attributes of object transformation. The information derived from the big data could also be used to make decisions as well as to forecast future developments and activities. Table 2 enlists few of the algorithms used for data mining in the mentioned methods (Table 1).
### Table 1 Models for data mining

<table>
<thead>
<tr>
<th>Data Mining</th>
<th>Predictive Mining</th>
<th>Descriptive Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Regression Analysis</td>
<td>Clustering</td>
</tr>
<tr>
<td>Prediction</td>
<td>Time-series Analysis</td>
<td>Association Rules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sequence Discovery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Summerization</td>
</tr>
</tbody>
</table>

### Table 2 Few techniques and algorithms used in data mining applications

<table>
<thead>
<tr>
<th>Technique</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>KNN, Naïve Bayes, C4.5, Decision tree, ID3, Lazy decision tree</td>
</tr>
<tr>
<td>Association Rule</td>
<td>Apriori and advanced apriori algorithms</td>
</tr>
<tr>
<td>Regression</td>
<td>Linear models, SVM, Multivariate Linear Regressions</td>
</tr>
<tr>
<td>Clustering</td>
<td>K-medoids, Self-organizing F maps, HAC, CLARANS</td>
</tr>
<tr>
<td>Time-Series Analysis</td>
<td>Las Vegas algo, Randomized algorithms</td>
</tr>
</tbody>
</table>

### 1. Classification

Classification involves guidelines on machine learning that are specific to additional knowledge that usually involve pre-processing of data, modeling architecture, allocation of the learning/feature and validation/evaluation. The generalized continuous and categorical functionality of classification is foreseen. For example, we may identify applications for bank loans as safe or risky through classification models. The classification is the model derivation evaluating an object's class depending on its characteristics.

The Random Forest, decision tree, AdaBoost and Naive Bayes techniques of data mining are effective and frequently used. Random Forest is the effective training framework that seeks to reduce uncertainty and maximize output quality. Through integrated ensemble learning across various training datasets, source parameter values for each tree within the forests are classified. An excellent example is the Naive Bayes classification which determines uncertain class identification conditions. This results in the identification, banking, categorization of text, advertising and knowledge representation of medical items. AdaBoost is by far the most appropriate binary grading approach that links a multitude of soft learning modules to increase their class segregation (Hashem IA, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU., 2015).
2. Characterization

Characterization of data is a descriptive summary, which generates objects that are known as characteristic laws of the specific attributes of subjects within the target class. The data for a user class are typically accessed by means of a database query and run using a description framework to retrieve the significant data at various abstract concentrations.

3. Clustering

For data exploration, clustering methods are beneficial. Assuming that there are several instances but no apparent natural groups, natural groupings could be identified with clustering-based data mining frameworks. A cluster comprises a set of data items that somehow mimic each other. Clustering is a form of grouping, though, is a non-supervised approach for learning that groups data fragments together into clusters of significance focused on their similarity indexes (Zanoon N, Al-Haj A, Khwaldeh SM., 2017), (Petre RS., 2012). A successful clustering process creates high-quality clusters that ensure a low similarity between the clusters and high similarities between the clusters; in other terms, the one cluster members are often similar to one another than from the cluster members of other group (Saigal S, Mehrotra D., 2012). Subsequently, clustering for data mining in cloud computing environment typically includes the following techniques or analysis tools: potential algorithm for data indexing, k-means, k-medoids, k-means, etc.

There are various kinds of clustering strategies for various subjects, that include a multitude of algorithms. The most important are: Model-based, grid-based, density-based, partitioning, hierarchical method, etc. Clustering Algorithms are classified by various kinds of grouping techniques (Hassan MK, El Desouky AI, Elghamrawy SM, Sarhan AM., 2018).

(a) Hierarchical Based: ROCK, Echidna, Chameleon, CURE, and BIRCh.
(c) Model Based: COBWEB, SOMs, EM, and CLASSIT.
(d) Grid-Based: OptiGrid, CLIQUE, Wave-cluster, and STING.
(e) Density Based: DBLASSD, DENCLUE, OPTICS, and DBSCAN.

4. Association Rule

For market-based research, an interaction model is sometimes used to try to establish ties or associations in a variety of products. In marketing and advertising, catalogue creation
as well as other corporate decision-making methods, for this the market basket analysis is commonly employed. For instance: '70% of people who purchase pasta, wine and sauce also buy a kernel bread' (Hicham GT, Lotfi E., 2017), a common law of association of this type says.

Under cloud computing environment association rule typically includes the following methodologies: apriorialgorithm, parallel association rule data mining, strengthened apriorialgorithm, the field / sphere and MapReduce algorithms, etc. Advanced apriori's algorithm is often utilized to address common problems associated with traditional Apriori data mining in the cloud computing environment.

5. Regression

Regression is a method with limited data model error. Regression analysis is still the most frequently utilized mathematical methodology for estimation and prediction of numerical values, where its usage overlaps greatly with the machine-learning domain. Regression analysis is often used for understanding whether any of the independent variables are correlated with and analyzing the types of certain associations between the dependent variable. Regression analysis may be used under restricted situations to deduce statistical correlation among dependent and independent variables under study. Consequently, they are often unsatisfactory and deliver false positive outputs (Han J, Kamber M, Pei J., 2011).

6. Time Series Analysis

Sequences of timestamps, measured normally at intervals of the same duration. The time series analysis is the method by which mathematical methods are used to model and expose a series of data points that depend on time. Time Series Analysis is a way to use an algorithm to determine events in the future, based upon previously observed events (Saigal S, Mehrotra D., 2012).

7. Neural Networks

The MLP-based neural networks are fitted with 3 connected layers (viz. input layer, output layer and hidden layer), of neurons or perceptrons and often are referred to as backpropagation networks. The input is generated in the neural network as the basic non-linear function which is the characterization of the inputs at the output layer. The nodes of the 3 layers are closely linked, but they are not interlinked. The nodes of that
layer. The given input is linked to the hidden layer nodes and thereafter transferred to the output layer (Hicham GT, Lotfi E., 2017).

For prediction purposes, neural networks may be used for data mining mostly with multiple networks. The input layer used for the given inputs displays the input characteristics and, on the basis of its values, is wrong or valid, depending on their presence. The secret layer embraces the input layer type and transfers it to the output layer, the output layer generates the results of the data mining prediction models, provided with possible input event predictions (Sinkov A, Asyaev G, Mursalimov A, Nikolskaya K., 2016).

![Architecture of a neural network](image)

**Figure 4 Architecture of a neural network**

8. Bayesian Algorithm

Bayesian algorithm can be implemented for predictive data mining, the estimation percentage of this algorithm is smaller than other algorithms, since it evaluates the conditions for inputs and output and does an evaluation that depicts the correlations among inputs and the output (Siddiqui T, Ahmad A., 2018). It can be significant in the initial detection processes that can be used for timely identification of identification of illnesses or serious diseases.

9. Benefits of the Data Mining

- Enhances the organization's benefit levels.
- Boost decision-making.
- Can be carried out quickly with a modern method.
• Provides the basis of knowledge for the particular field of expertise.
• Prediction and research automation.
• It guarantees a potent method to evaluate a large volume.

Conclusion

Cloud infrastructure data processing applications are an important feature of strategic, knowledge-based actions made by today's companies, helping them foresee emerging developments and behavioral biases. This paper presented a survey of various approaches and tools for data mining under cloud computing such as Classification, Clustering, Association Rules, Prediction, Time Series Analysis, Regression and Sequence Discovery. These innovative methods offer a number of uses in the retail, telecommunications, pharmaceutical, etc. sectors. These techniques foresee potential developments and conducts that empower businesses to build constructive and existing variable knowledge in various forms understandable by the man.

References


