International Journal of Research and Applications

ISSN (online): 2349-0020 ISSN (print): 2394-4544 http://www.ijraonline.com/

Review Report

A comprehensive review on classification of data mining systems and a three tier data warehouse architecture

Sravanthi Kurapati

University College of Engineering- Kothagudem, Kakatiya University, Telangana State, India -507115.

ARTICLE INFO

Corresponding Author:
sravanthi.k1982@gmail.com

DOI:
http://dx.doi.org/
10.17812/IJRA.7.25 (5)2020

Manuscript:
Received: 20th Jan, 2020
Accepted: 22nd Feb, 2020
Published: 25th Mar, 2020

Publisher:
Global Science Publishing Group, USA
http://www.globalsciencepg.org/

ABSTRACT

Many data-based modeling studies are carried out in a specific application domain name. Hence, domain-specific understanding and also experience are usually essential in order to develop a purposeful issue statement. In this step, a modeler usually defines a set of variables for the unknown reliance as well as, ideally, a general kind of this dependence as a first hypothesis. This paper provides a review on classification of data mining systems and three tier data ware house architecture.

Keywords: Data mining, data ware house architecture.

1. INTRODUCTION TO ARCHITECTURE OF DATA MINING

Data mining is the method of extracting potentially very valuable information from large datasets which is previously unknown and useful for business purposes. A Typical data mining system may have the following major components: knowledge base, Data Mining Engine, Pattern Evaluation Module, User Interface and Data Warehouse Server (Fig. 1).

Knowledge Base:

This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

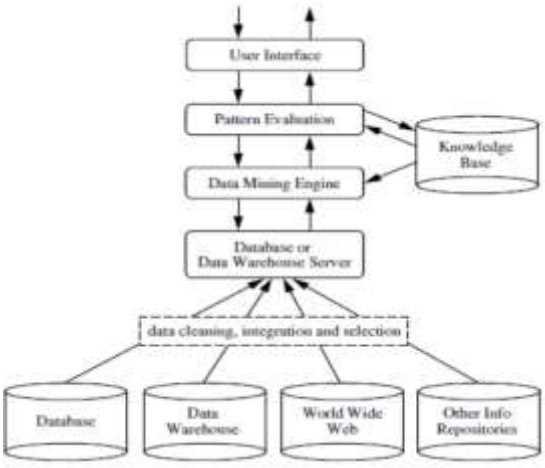

Figure 1: Architecture of Data Mining

Data Mining Engine:

This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern Evaluation Module:

This component typically employs interestingness measures interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process as to confine the search to only the interesting patterns.

User Interface:

This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Data Warehouse Server

The data warehouse server is the machine that stores all the historic data which is ready to be processed. Based on the user request, the data warehouse server fetches the required data, and, thus, the actual datasets can be very personal.

2. DATA MINING PROCESS

Data Mining is a process of finding various models, recaps, and acquired values from a given collection of data.

The basic experimental procedure adapted to data-mining issues entails the complying with actions:

i. State the trouble as well as formulate the theory

Many data-based modeling studies are carried out in a specific application domain name. Hence, domain-specific understanding and also experience are usually essential in order to develop a purposeful issue statement. Unfortunately, numerous application researches tend to focus on the data-mining strategy at the expenditure of a clear issue statement. In this step, a modeler usually defines a set of variables for the unknown reliance as well as, ideally, a general kind of this dependence as a first hypothesis. There may be several hypotheses created for a solitary issue at this phase. The primary step calls for the combined expertise of an application domain name as well as a data-mining design. In practice, it usually indicates a close interaction between the data-mining specialist and also the application specialist. In effective data-mining applications, this teamwork does not drop in the initial stage; it proceeds during the whole data-mining procedure.

ii. Collect the data

This step is worried about exactly how the data are created as well as accumulated. Generally, there are 2 distinctive possibilities. The very first is when the data-generation process is under the control of a specialist (modeler): this method is called a created experiment. The 2nd opportunity is when the professional cannot influence the data- generation procedure: this is referred to as the observational technique. An observational setting, particularly, random data generation, is thought in most data-mining applications. Normally, the sampling circulation is totally unknown after data are accumulated, or it is partly and also unconditionally given in the data-collection treatment. It is extremely important, nonetheless, to comprehend how data collection influences its theoretical circulation, considering that such a priori knowledge can be very helpful for modeling as well as, later, for the final interpretation of outcomes. Additionally, it is important to see to it that the data made use of for approximating a design as well as the data made use of later for testing as well as using a model originated from the same, unknown, sampling distribution. If this is not the instance, the approximated design cannot be successfully made use of in a final application of the outcomes.

iii. Preprocessing the data

In the observational setting, data are generally "collected" from the existing databases, data storehouses, as well as data marts. Data preprocessing typically includes at least 2 usual tasks:

a. Outlier discovery (and removal):

Outliers are unusual data values that are not consistent with a lot of observations. Generally, outliers result from measurement mistakes, coding and also recording mistakes, and also, often, are all-natural, irregular values. Such non representative examples can seriously influence the version generated later on. There are 2 methods for managing outliers:

- Detect and also at some point get rid of outliers as a part of the preprocessing stage, or
- Develop durable modeling approaches that are aloof to outliers.

b. Scaling, encoding, and choosing attributes:

Data preprocessing consists of several actions such as variable scaling as well as various kinds of inscribing. For instance, one function with the array [0, 1] and also the other with the array [−100, 1000] will not have the same weights in the applied method; they will certainly also influence the final data-mining results in different ways. For that reason, it is advised to scale them as well as bring both functions to the very same weight for further analysis. Additionally, application-specific inscribing approaches generally attain dimensionality reduction by providing a smaller number of interesting functions for subsequent data modeling.

These 2 classes of preprocessing jobs are just illustratory examples of a large spectrum of preprocessing tasks in a data-mining procedure.

iv. Quote the design

The selection as well as application of the ideal data-mining method is the major task in this stage. This process is not simple; usually, in practice, the application is based on numerous designs, as well as selecting the most effective one is an added task.

v. Interpret the design as well as draw conclusions

In many cases, data-mining designs need to assist in decision making. Thus, such versions need to be interpretable in order to be useful since human beings are not most likely to base their decisions on facility "black-box" models. Note that the objectives of precision of the model and also accuracy of its interpretation are somewhat inconsistent. Generally, basic models are a lot more interpretable, however they are also much less exact. Modern data-mining methods are expected to produce highly precise outcomes using high dimensional designs. The problem of analyzing these versions, likewise really vital, is

considered a separate task, with certain methods to confirm the outcomes. A customer does not want thousands of pages of numeric outcomes. He does not comprehend them; he cannot sum up, analyze, and also utilize them for effective decision making.

3. CLASSIFICATION OF DATA MINING SYSTEMS

The data mining system can be identified according to the following criteria:

- Data Source Modern Technology Stats
- Machine Learning Information Science Visualization
- Other Self-controls

Some Other Classification Criteria:

- Classification according to kind of Databases mined
- Classification according to kind of knowledge mined
- Classification according to kinds of techniques utilized
- Classification according to applications adapted

Classification according to kind of databases mined

We can classify the data mining system according to kind of databases mined. Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example if we classify the database according to data model then we may have a relational, transactional, object- relational, or data warehouse mining system.

Classification according to kind of knowledge mined

We can classify the data mining system according to kind of knowledge mined. It is means data mining system are classified on the basis of functionalities such as:

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

Classification according to kinds of techniques utilized

We can classify the data mining system according to kind of techniques used. We can describes these techniques according to degree of user interaction involved or the methods of analysis employed.

Classification according to applications adapted

We can classify the data mining system according to application adapted. These applications are as follows:

- Finance
- Telecommunications
- DNA
- Stock Markets
- E-mail

4. THREE TIER DATA WAREHOUSE ARCHITECTURE

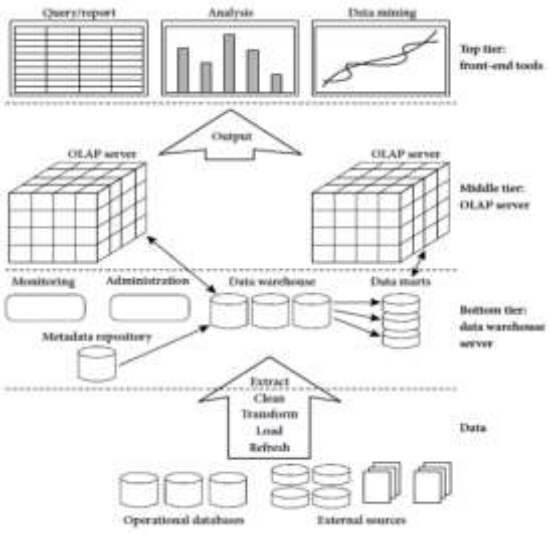

Figure 2: Three tier architecture

Tier-1:

The bottom tier is a warehouse database server that is almost always a relational database system (Fig. 2). Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning, and transformation (e.g., to merge similar data from different sources into a unified format), as well as load and refresh functions to update the data warehouse . The data are extracted using application program interfaces known as gateways. A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.

Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection).

This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

Tier-2:

The middle tier is an OLAP server that is typically implemented using either a relational OLAP (ROLAP) model or a multidimensional OLAP.

- OLAP model is an extended relational DBMS that maps operations on multidimensional data to standard relational operations.
- A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

Tier-3:

The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

5. META DATA REPOSITORY

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

A metadata repository should contain the following:

- A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.
- Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails).
- The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports.
- The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security.
- Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.

6. CONCLUSION

Data-preprocessing steps need to not be taken into consideration completely independent from various other data-mining stages. In every version of the data-mining process, all activities, with each other, can define brand-new and enhanced data sets for succeeding versions. Typically, an excellent preprocessing approach gives an optimal representation for a data-mining method by incorporating a priori expertise in the form of application-specific scaling and encoding. This paper has provided a rreview on classification of data mining systems and a three tier data warehouse architecture.

REFERENCES

1) Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules (1994) Proc. 20th Int. Conf. Very Large Data Bases, VLDB-94.

2) Bayardo, R. and R. Srikant, Technological Solutions for Protecting Privacy, IEEE Computer, Sep 2003.

3) Chernoff, H. (1973). Using faces to represent points in k-dimensional space graphically. Journal of American Statistical Association, 68, 361-368.

4) Domingos, MetaCost: a general method for making classifiers cost-sensitive, KDD-99, Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, 1999.

5) Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, Advances in Knowledge Discovery and Data Mining, (Chapter 1), AAAI/MIT Press 1996.

6) Mounika Reddy, Avula Deepak, Ekkati Kalyani Dharavath, Kranthi Gande, Shoban Sriramoju, "Risk-Aware Response Answer for Mitigating Painter Routing Attacks" in "International Journal of Information Tech. and Management", Vol VI, Issue I, Feb 2014 [ISSN : 2249-4510].

7) Mounica Doosetty, Keerthi Kodakandla, Ashok R, ShobanBabu Sriramoju, "Extensive Secure Cloud Storage System Supporting Privacy-Preserving Public Auditing" in "International Journal of Information Technology and Management", Volume VI, Issue I, Feb 2012 [ISSN : 2249-4510].

8) Shoban Babu Sriramoju, "An Application for Annotating Web Search Results" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2,Issue 3,Mar 2014 [eISSN : 2320-9801, pISSN: 2320-9798.

9) E. Raju and K. Sravanthi, "Analysis of Social Networks Using the Techniques ofWeb Mining," in International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, issue 10, pp. 443-450, 2012.

10) Shoban Babu Sriramoju, Madan Kumar Chandran, "UP-Growth Algorithms for Knowledge Discovery from Transactional Databases" in "International Journal of Advanced Research in Computer Science and Software Engineering", Vol 4, Issue 2, February 2014 [ISSN : 2277 128X].

11) Shoban Babu Sriramoju, Azmera Chandu Naik, N.Samba Siva Rao, "Predicting The Misusability Of Data From Malicious Insiders" in "International Journal of Computer Engineering and Applications" Vol V, Issue II, Febrauary 2014 [ISSN : 2321-3469].

12) Ajay Babu S, Shoban Sriramoju "Analysis on Image Compression Using Bit-Plane Separation Method" in "International Journal of Information Technology and Management", Vol VII, Issue X, November 2014 [ISSN: 2249-4510].