MAT JOURNALS

Journal of Data Mining and Management
e-ISSN:2456-9437
Volume 3 Issue 3

A Comparative Detailed Study of Data Mining Methods and Tools in Data Warehouse

[1]Anil Kumar, [2]Amit Kumar Kar, [3]Shailesh Kumar Patel, [4]Roop Ranjan
[1,2,3,4]Assistant Professor
Department of Computer Science & Engineering, Institute of Technology and Management, Integrated Technical Campus, GIDA, Gorakhpur, UP, India
Email:*[1]anil.itmgkp@gmail.com, [2]amit.kar1983@gmail.com, [3]shaileshkumarpatel2008@gmail.com, [4]roop.ranjan@gmail.com*
DOI: *http://doi.org/10.5281/zenodo.1456480*

Abstract

Now-a-days we have a tendency to be in modern era. There are immense quantity of knowledge and knowledge, that to be collected from completely different sources and analyzed to urge the information. Once grouping the info from numerous sources, it's keep in huge repositories, that is thought as knowledge warehouse. There is variety of techniques wont to extract the knowledge from data warehouse and analyze to urge the important information that is thought as data processing. For this purpose, we have a tendency to use completely different data processing tools like wood hen, KEEL, R, KNIME, ORANGE etc. During this paper we are going to compare completely different data processing technique and tools for maintenance in knowledge warehouse.

Keywords: *Data Warehouse, Data Mining, Weka, KEEL, R, KNIME, ORANGE.*

INTRODUCTION

Data mining is outlined because the follow of examining massive pre-existing databases so as to come up with new data. In different words it is outlined because the method of analysing knowledge from completely different views and summarizing it into helpful data. It's generally additionally called knowledge or information discovery.

There are primarily two techniques employed in data processing i.e. association mining and cluster. In association data processing the co-occurrence of one knowledge item with the info item is noticed. Association may be a data processing perform that finds the chance of the co-occurrence of things in a very assortment of knowledge. The relationships between co-occurring things are expressed as association rules. Cluster may be a technique by that the hierarchy of the info things is fashioned, so one set of knowledge things is differentiated from the others. Organization and account of knowledge is given by abstracting the underlying structure of cluster analysis [1].

Data mining tools can be used to predict the future trend, co-occurrence of data items, knowledge pattern and other decisions. Now-a-days, there are number of data mining tools used. In this paper we will discuss five open source tools i.e. Weka, KEEL, R, KNIME and ORANGE.

Data mining tools is accustomed predict the longer term trend, co-occurrence of knowledge things, information pattern and different choices. Now-a-days, there is variety of knowledge mining tools used. During this paper we are going to discuss 5 open supply tools i.e. Weka, KEEL, R, KNIME and ORANGE. Data reposition is outlined as a central repositories system used for coverage and knowledge analysis. It stores current and historical knowledge

and used for making analytical reports for information employees throughout the enterprise. Therefore, it's additionally called enterprise knowledge warehouse. It Integrates knowledge from completely different supply systems, making a central read across the enterprise.

DATA MINING TECHNIQUES

There are mainly two data mining techniques used i.e. association mining and clustering [3].

a) Association Mining: In association mining there is use of different association rules to find the frequent item sets. It means to find the association of one item with the other items. AssociationrulesareoftheformifPthenQ.For example:75%of thosewhobuycarinsurancealsobuyhomeinsurance;80%of thosewhobuyclothon-linealsobuyelectronic itemson-line;35% of thosewhohavemore than one carandarehaving more than one houses.

The association of items is shown in the table given below.

Table: 1.*Market basket analysis*

S.No.	Items
1	{Bread, Butter}
2	{ Bread, Egg, Milk}
3	{Butter, Paneer, Coke}
4	{Bread, Butter, Egg, Coke}

For example

{Bread}⟶{Butter}

It means when customer will purchase bread, there may be chances of purchasing of butter. So there is a strong relationship between these two items.

Associationrefersto miningfrequentpatterns,frequent items and correlationsamongthedatain largetransactional or relationaldatasets.

b) Clustering: Cluster analysis makes and add up information by abstracting underlying structure either as a grouping of people information or as a hierarchy of teams. In cluster analysis the info objects owned by a cluster area unit similar, area unit known as undiversified cluster and also the objects owned by totally different clusters, are known as heterogeneous cluster. This definition shows that cluster can't be a ballroom dance method [1].There are some different types of clustering used, which are given below:

i) Hierarchical-based clustering
ii)Partitioning-based clustering
iii) Grid-based clustering, and
iv) Density-based clustering.

i) Hierarchical-based clustering: Hierarchical-based clustering algorithmsmakeahierarchicaldecompositiono f thearticles. Hierarchical clustering creates a cluster. Hierarchymeans atree structure ofclusters. Each and each cluster node consists of kid clusters. Such an approach permits exploring information on levels of various roughness. Hierarchical-based cluster ways area unit divided into agglomerated (bottom-up) and divisive (top-down).

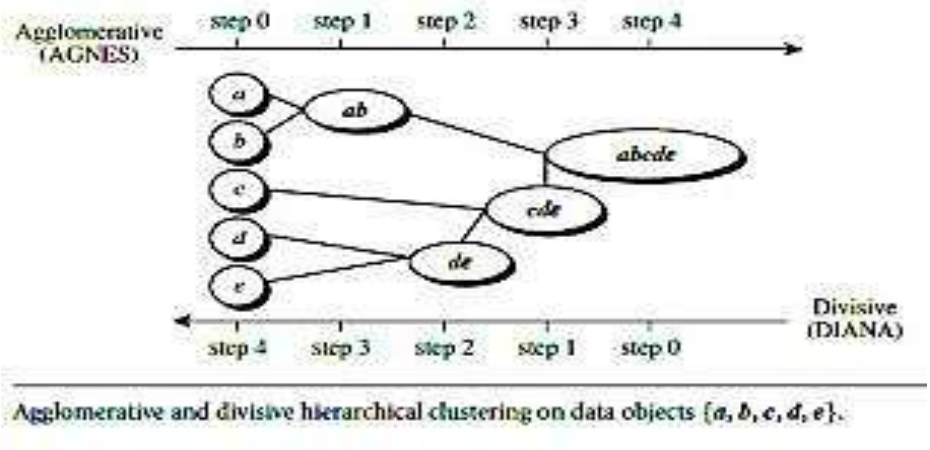

Fig: 1. *Hierarchical-basedclustering algorithms*

ii) Partitioning-based clustering: Data partitioning-based algorithms create the info into variety of subsets. Specifically, it means that totally different relocation ways that iteratively delegate points between the k numbers of clusters. Relocation algorithms bit by bit improve clusters with applicable information, this end in top quality clusters.

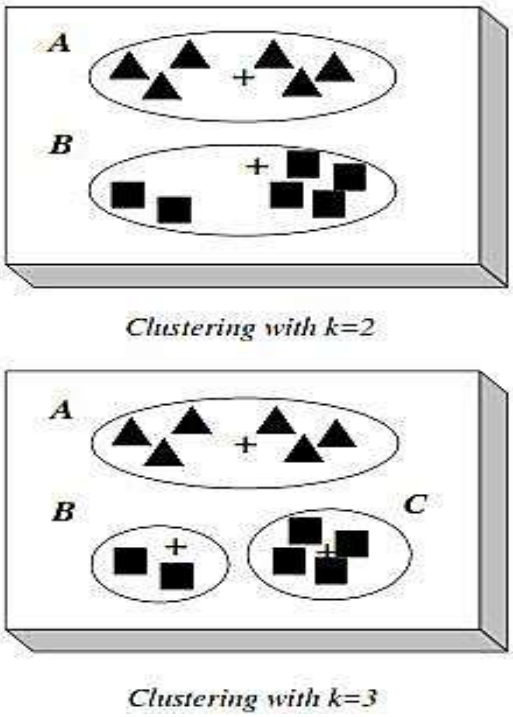

Fig: 2. *Partitioning-based clustering algorithms*

iii) Grid-based clustering: The main principleof thesealgorithmsistoquantize and quantify thedatasetinto the different cellsandthen the workwithobjects containingtothesecells.Theydonotrelocatepoi nts. They are nearer to hierarchical algorithms however the concatenation of grids, and consequently clusters, doesn't build a distance live however it's determined by some predefined parameter.

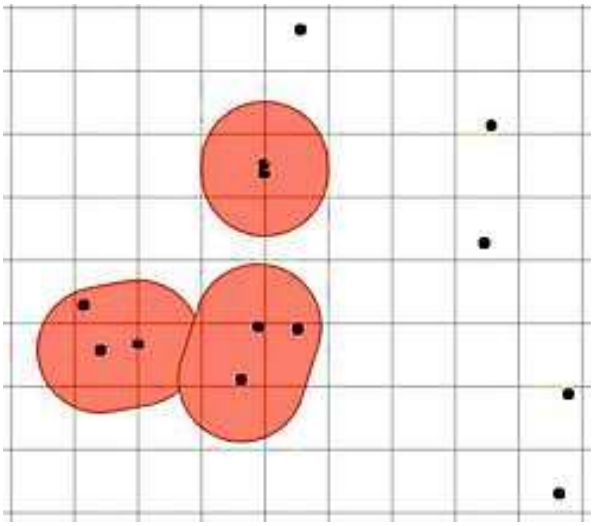

Fig: 3. *Grid-based clustering algorithms*

iv) Density-based clustering: In density-based clusteringtechniques, there are some different densities exist on the basis of presence of different data objects in a specified area. Thisalgorithmcreates groupof data objectsaccordingto their densityobjectivefunctions.Densityisnormallydefinedasthecollection ofdata objectsinaparticularneighbourhoodofadataobjects.Inthistechniqueagivenclustercontinuesmove to a higher levelas longasthe numberofobjectsinthe neighbourhoodexceedssomespecific parameters.It is thought of to diverge from the concept in divided algorithms that use unvaried relocation of points given a particular range of clusters.

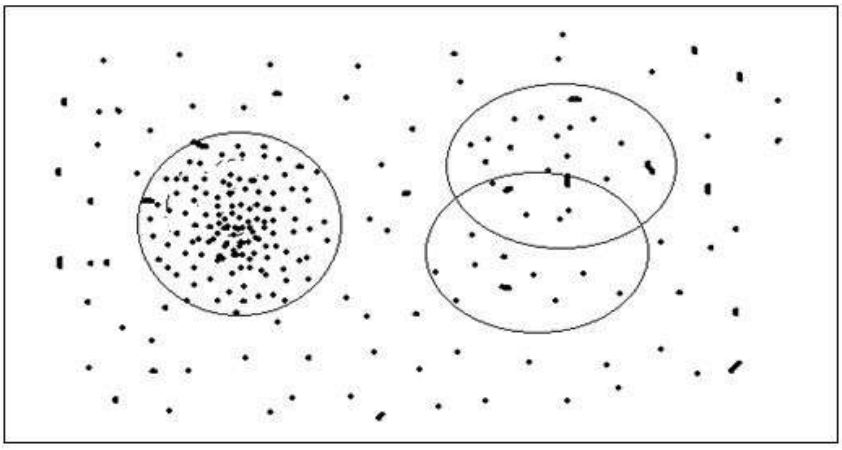

Fig: 4. *Density-based clustering algorithms*

The following table 2 having the comparison of different data mining techniques on the basis of different methodology used in mining.

Table: 2.*A comparative study of different data mining techniques*

S. No.	Data mining techniques	Algorithm	Methodology
1	Association mining	Frequent mining	Using domain creation clusters are formed and efficient merging process is used
2	Clustering	Hierarchical	Use of top-down and bottom-up approach, the hierarchy among the different clusters
		Partitioning	Uses similarity, function, view, replacement function to dynamic materialized view selection
		Grid-based	Formation of grid by the use of different combination of rows and columns
		Density-based	On the basis of number of elements present in specific area

DATA MINING TOOLS

Data mining has a different application like advertising of different goods and services. Complexity involved in building data mining applications, a large number of data mining tools created over past some years. Within data processing, there's a bunch of an oversized range of tools that are developed by analysis community and information analysis teams. They're offered freed from price mistreatment one in all the present open supply tools [2].

Data mining givesdifferent mining techniques to extract data. The development and application of data mining algorithm needs use of a large number of data mining tools, which are discussed below. Every data mining tool has its own pros and cons.

i) Weka: Weka (Waikato environment for knowledge Analysis) could be an assortment of machine learning algorithmic program and huge range of tools for image for analytics of knowledge and prognostic modelling for acting variety of knowledge mining tasks.These algorithms are often used directly on a collection of knowledge things or with the assistance of java code. It includes pre-processing on information objects, classification of knowledge objects, bunch of knowledge objects and association rule extraction. Weka provides following three types of graphical user interface:

(a) Data analysis to support pre-processing on data, attribute selection, learning, visualization

(b) Provide environment for testing and evaluating machine learning algorithms and

(c) The knowledge flow for new process.

Weka is best suited for mining association rules. It also has poorlinked to excel spreadsheet and non java based databases.

ii) KEEL: It, Knowledge Extraction based on Evolutionary Learning (KEEL), is an application package of machine learning software tools. It is designed for providing solution to different data mining problems. It conjointly provides a group of libraries for pre and post-processing methodology for information manipulating, soft computing strategies in data of extracting and learning, and supply scientific and analysis strategies for data processing techniques. It considers regression, classification, bunch and pattern mining then on. It conjointly has process intelligence based mostly learning algorithms for hybrid model like genetic fuzzy system, neural networks etc.

iii) R: Revolution (R) could be a free artificial language and provides code setting for applied math computing and graphics. The R language is employed between statisticians and information miners for developing applied math code and information analysis.It has totally completely different blessings of publication of quality plots together with symbol's of arithmetic and different formula. Numerical programming is healthier integrated in R. Import and export of knowledge from programme is simpler in R. It's a strong language in APL, MATLAB and LISP. Less specialised towards data processing is one among the restrictions of R.

iv) KNIME: It, Konstanz data mineworker, is an open supply information analytics, coverage and integration platform for data processing. It's been used in analysis together with totally different areas of intelligence and money analysis. KNIME is predicated on Eclipse platform. KNIME is written in java. It's a standard information exploration technique that allows the user to visually produce information flows, by selection executes some or all analysis steps and later investigates the result through mutual perspective on information and models. It makes all analysis modules of the acknowledge. It's solely restricted error mensuration strategies. It doesn't have automatic facility parameter optimisation of machine learning. It's designed for enterprise coverage, business intelligence and data processing.

v) ORANGE: This tool has different components for artificial intelligence and in bio-medicals & text mining and it is linked withfeatures for data analytics.
The following table 3 having the comparison of different data mining tools.

Table: 3.*A comparative study of different data mining tools*

S. No.	Name of Tool	Language	Operating System	Area
1	Weka	JAVA	Cross platform	Machine learning
2	KEEL	JAVA	Cross platform	Machine learning
3	R	C, Fortran	Cross platform	Statistical computing
4	KNIME	JAVA	Linux, Windows	Enterprise reporting, Data mining, Business intelligence
5	ORANGE	C, C++, Python	Cross platform	Machine learning, Data mining, data visualization

CONCLUSION

In this paper, we tend to are comparison completely different data processing techniques on the idea of various methodology used. Conjointly there's comparison of completely differentdata processing tools on the idea of supporting different package and language.

There are in the main 2 data processing techniques mentioned i.e. association mining and bunch. In association mining, there's use of frequent item sets on the idea of various association rules. There are such a lot of kinds of bunch i.e. hierarchical-based, partitioning-based, grid-based and density-based. Each clustering algorithms are used in different data sets to find out the different clustersfor the specific area.

In this paper, we have discussed different open source data mining tools i.e. Weka, KEEL, R, KNIME and ORANGE. Each tools used in different applications, which gives their own maximize result. Like Weka is mainly used in visualization analysis, which is more sophisticated. KEEL is mainly used in scientific and research analysis. R is mainly used in statistical computing and graphics analysis.

KNIME is generally employed in business intelligence method, money information analysis. In KNIME, there is use of pipelining concept in data analysis. ORANGE is used in bioinformatics and text mining. Each data mining tools having their own merit in data analysis.

REFERENCES

1. Amit Kumar Kar, Shailesh Kumar Patel and Rajkishor Yadav, "A Comparative Study and Performance Evaluation of Different Clustering Techniques in Data Mining",IEEE sponsored ACEIT Conference Proceeding 2016, IJCSIT, March 2016.
2. KalpanaRangara andDr. K. L. Bansal, "Comparative Study of Data Mining Tools", IJARCSSE, Volume 4, Issue 6, June2014, ISSN: 2277 128X.
3. P. R. Vishwanath,Dr. D. Rajyalakshmi andDr. M. Sreedhar Reddy, "A Comparative Study of Data Mining Techniques in Maintenance of Data warehouse", IJSER, Volume 4, Issue 11, November 2013, ISSN: 2229-5518.

Cite as: Anil Kumar, Amit Kumar Kar, Shailesh Kumar Patel, & Roop Ranjan. (2018). A Comparative Detailed Study of Data Mining Methods and Tools in Data Warehouse. Journal of Data Mining and Management, 3(3), 18–24. http://doi.org/10.5281/zenodo.1456480