iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** X **Month of publication:** October 2020

DOI: https://doi.org/10.22214/ijraset.2020.32045

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue X Oct 2020- Available at www.ijraset.com

Knowledge Discovery and Data Mining

Siddharth Nandakumar Chikalkar
Bachelor of Computer Application, Vivekanand collage, Kolhapur

Abstract: knowledge discovery in databases (KDD) plays an important role in large organisation where data is store in large base.it help with exploring and understanding very large data set and building predictive model. This is the task-oriented process it been to identifying valid useful and understandable pattern from large and complex data set .data mining is the core of KDD process in KDD process interring the algorithm for extracting useful information the model purpose is understanding analysis and prediction. Increasing growth of every sector produce data and helping of this model we recognize the pattern and trend in large data sets in sector.

I. INTRODUCTION

Data science is the field which every field is needed. every day data is producing rapidly and this data have to handle in every day for increasing productivity .data mining is the incorporation of quantitative methods or mathematical method that may include mathematical equation algorithms some your prominent methodologies are tradition logistic regression neural network segmentation classification clustering those are all method that utilize mathematics .data mining is applicable across industry sectors generally wherever you have processes wherever you have data it is the application of those powerful mathematical techniques in core incorporation with some statistical type of inference testing they call it that will extract trends and patterns there data mining is use. Basically, data mining is the process where the raw data turn into useful information .it has many phares to analyse data and extract useful information .in this paper we see all of those steps in KDD means knowledge discovery in database. KDD is the process of finding knowledge in large data base it is the procedure of the data mining.

II. WHY WE NEED DATA MINING

Everyday volume of information is increasing rapidly and we handle business transaction, sensor data, scientific data videos picture etc. evolution of technology increasing production of data every day. That's why the explosive growth of data from terabytes to petabytes. Data availability has been easily like form automated data collection tool, data base system, web computerised society data can available in large amount. And we have data form business like web e-commerce, transaction, stock etc. form science we also got remote sensing data bioinformatic scientific simulation etc. and mainly we got a lot of data from society and everyone. like news, cameras, YouTube, social media platform like Instagram, Facebook, twitter, snapchat any many more. So, we need some kind of system that will capable of extracting essence of information available and that can automatically generate report. views or summery of data for decision making.

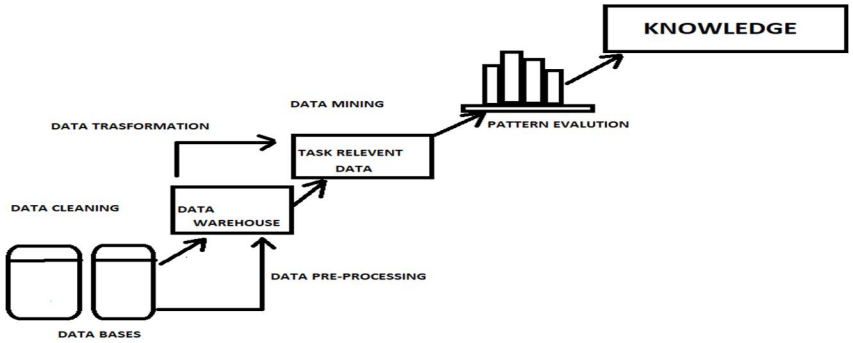


Figure: knowledge discovery in Database

III. KNOWLEDGE DISCOVERY PROCESS

The KDD process is used in large data set to identify pattern and trend there is many phrases of this process it is the traditional method of turning data into knowledge relies on manual analysis and interpretation. The process starts with the KDD goals, and ends with the implementation of the discovered knowledge. the process is iterative at each step, meaning that any step can moving back to previous steps may be required. First in this process understand the goal of end-user. And then the process is beginning with data cleaning.

A. *Data Cleaning*

This is the first process of data mining the Data cleaning is defined as removal of noisy and irrelevant data from collection. We got the data for data mining from multiple sources so some kind of data may irrelevant to the data mining process so in this step we clean the data and extract the relevant data from all source we have. There Is different type of source of data that are used in data mining process. The data from multiple sources are integrated into a common source known as Data Warehouse**.** and the data which we got for mining it would be flat files means the data file in text form or binary form which easily extract by data mining algorithm, relational databases in this type we got data in rows and column physical schema in relational data base define the structure of the table and logical schema define the relationship among the table. And next is transaction data bases in this type of source we can get the data organize by time date and stamps to represent the transaction in data base. This type of data base capable to roll back or undo operation when a transaction is not complected or committed. Next is multimedia data base this type of data base consist audio, video image and text media. they can store in object-oriented data base. Next is spatial data bases in this type of data base we can get the geographical information. Next is time series data where we can get the stock exchange data user logged data .and last is www means world wide web is the collection of audio video text etc which is identified uniform resource locator through web browser. This is the all type of source we gather the data for data mining.

B. *Data Pre-processing*

In this stage of data mining where multiple data source is combined. After data cleaning we got the data from various source so here we integrate those data. then the only those data will be retrieved from data base which is relevant from analysis task. Then we got the data in consist state for applying algorithm.in this data pre-processing we arises problem which is some data is missing from data .so we have to fill missing value there are the various way to do this task. We can choose to fill value manually, by attribute mean and most probable value. And regression Here data can be made smooth by fitting it to a regression function. The regression used may be (having one independent variable) or multiple (having multiple independent variables). after data pre-processing the data which is extracted, this data is also important is describe useful information this data is used to help an organization to decision making because this data is integrated data from one or more disparate source.

C. *Data Transformation*

In the data transformation data transformed to appropriate form for data mining. there is different step for data transformation the first step is smoothing, in this process the noise of data will be eliminate by some algorithm. and we can highlight some important features in the data set.it help in predicting pattern. And after smoothing the data we can identify the simple change to predict different trend and pattern. The next step in data transformation is aggregation. here the data is store in summery format. the data integrate into data analysis discerption. this collection of data is useful from everything for decision concerning, strategy, product prising, operations and marketing strategy. after that the discretization process has been proceed, here transforming the continuous data into set of small intervals. Because the data mining activities required the continues attribute .data mining task can manage the continuous attribute.it can improve efficiency by replacing the constant quality attribute with discrete values so its transformed data in set of small intervals like (1-10,11-20) .one of data transformation procedure is normalization. this procedure involves converting all data variable into given range. It generally required when we are dealing with attribute with different scale. There is some method for data normalization which is decimal scaling method, min-max normalization and z-score normalization. All of this the data ready to data mining

D. *Data Mining*

This process is important now we have to decide which type of data mining to use for example regression or clustering .in this process the useful pattern been extracted from data it is intelligent method are applied in order to extract useful information from transformed data .and the pattern are extracted by algorithm .in data mining the algorithm use like c4.5 ,k-mean, algorithm ,expectation-maximization this kind of algorithm used in data mining .k-mean and expectation maximization generally use in data mining process of KDD.

IV. PATTERN EVALUATION

in this stage is identify patterns obtain in data mining pattern been convert in knowledge here use summarization and visualization techniques to make data understand by user.in this stage of knowledge discovery the pattern and trend are have been identified. And this useful information has been representing for strategy and prediction.

V. KNOWLEDGE REPRESENTATION

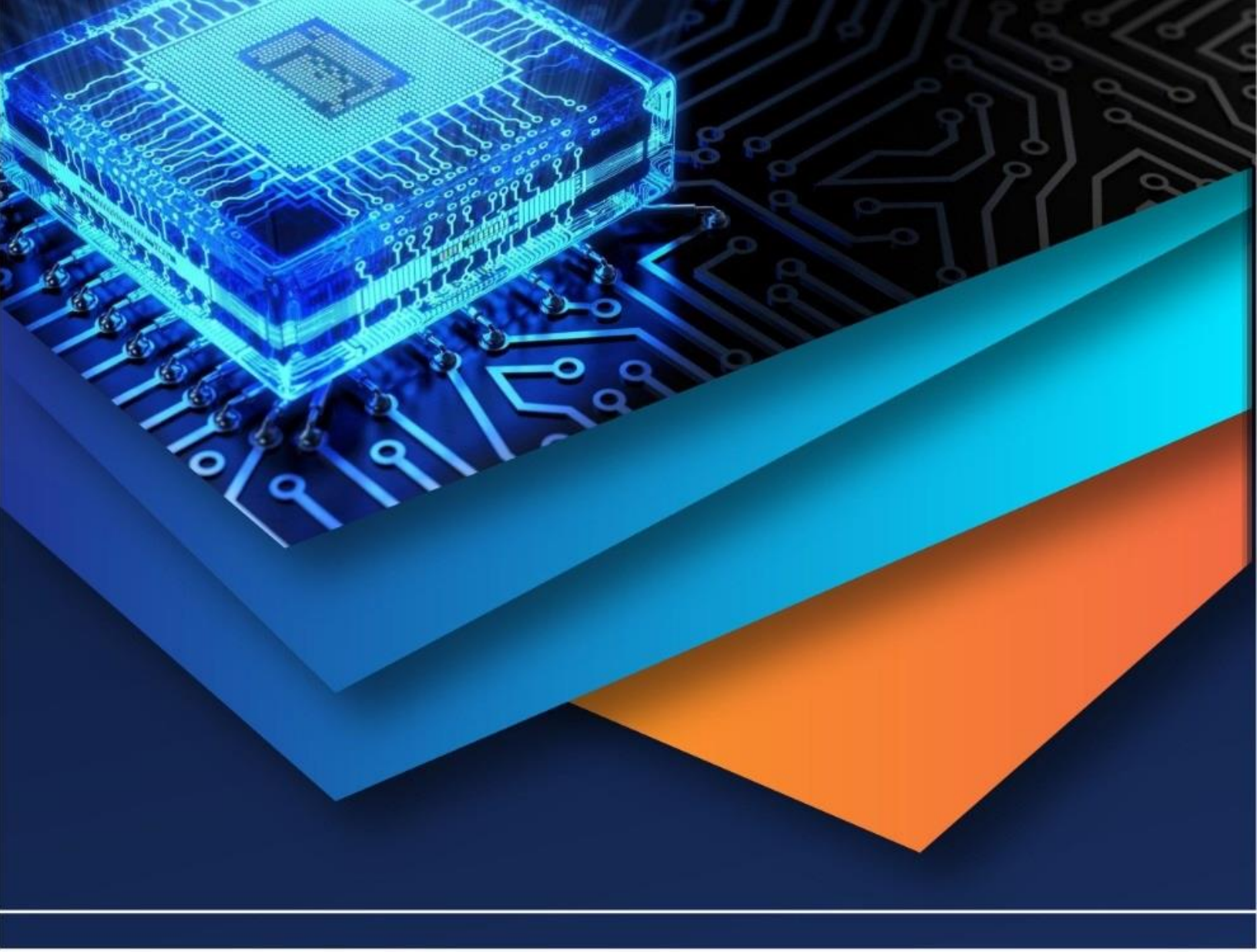
Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results. from generating report generating tables generate discriminate rules and classification rules or characterization rules etc.

VI. CONCLUSION

The object of this research paper is to study the KDD process. in this paper we present the different phrase in knowledge discovery process. The KDD process is one of the best way to finding trend and pattern in large data set. we provide which type of algorithm were used in data mining which is core of KDD process. And how data transformation process happened. The main advantage of the integrated approach is that the pre-processing steps are much easier and more convenient for data mining. Data pre-processing and data transformation is also important phrase in KDD and this phrase are very challenging to extract task relevant and useful data

REFERENCE

[1] Dehaspe, L., Toivonen, H., Discovery of frequent Datalog patterns. Data Mining and Knowledge Discovery, 3:7-36, 1999.
[2] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.
[3] R·Groth·HouDi.Data Mining - Building Competitive Advantages of Enterprises[M]. Xi'an:Xi'an Jiaotong University press,2001.
[4] YangJingfang.The application of machine learning algorithm in data mining[J].Electronic Technology & Software Engineering,2018(04):1
[5] ChenXiao.Application of machine learning algorithm in data mining[J].Modern Electronics Technique,2015,38(20):11-14
[6] L. Soibelman, M. Asce, K. Hyunjoo, Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases, J. Computing In Civil Engineering : (January 2002).
[7]] M. J. A. Berry, G. Linoff, Data mining: for Marketing, sales, and customer Support, John Wiley and Sons (Publish.): (1997).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY