ISSN: 2454-9940

# INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 1, 2025

# DATA MINING SYSTEM AND APPLICATIONS: A REVIEW

*K Radha[1], Dr. B LaxmiKantha[2]

[1]Assistant Professor, St. Martin's Engineering College, Secunderabad, Telangana – 500100

[2]Associate Professor, St. Martin's Engineering College, Secunderabad, Telangana – 500100

*Corresponding Author

Email: kradhait@smec.ac.in

***ABSTRACT:***

*In the Information Technology era information plays vital role in every sphere of the human life. It is very important to gather data from different data sources, store and maintain the data, generate information, generate knowledge and disseminate data, information and knowledge to every stakeholder. Due to vast use of computers and electronics devices and tremendous growth in computing power and storage capacity, there is explosive growth in data collection. The storing of the data in data warehouse enables entire enterprise to access a reliable current database. To analyze this vast amount of data and drawing fruitful conclusions and inferences it needs the special tools called data mining tools. This paper gives overview of the data mining systems and some of its applications.*

***Keywords:***

*Data mining system architecture, Data mining application*

## 1. INTRODUCTION

To generate information it requires massive collection of data. The data can be simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is 'Data Mining'.

Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were too time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

## 2. THE DATAMINING TASKS:

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as[1,2]:

1. Exploratory Data Analysis: It is simply exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.

2. Descriptive Modeling: It describe all the data, It includes models for overall probability distribution of

the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

3. Predictive Modeling: This model permits the value of one variable to be predicted from the known values of other variables.

4. Discovering Patterns and Rules: It concern with pattern detection, the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest.

5. Retrieval by Content: It is finding pattern similar to the pattern of interest in the data set. This task is most commonly used for text and image data sets.

## 3. TYPES OF DATAMINING SYSTEMS:

Data mining systems can be categorized according to various criteria the classification is as follows:

- Classification of data mining systems according to the type of data source mined: This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

- Classification of data mining systems according to the data model: This classification based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc.

- Classification of data mining systems according to the kind of knowledge discovered: This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

- Classification of data mining systems according to mining techniques used: This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data .

## 4. DATA MINING LIFE CYCLE:

The life cycle of a data mining project consists of six phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase .The main phases are:

1. Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

2Data Understanding: It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. Data Preparation: It covers all activities to construct the final dataset from the initial raw data.

4. Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

5. Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

6. Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as

simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

## 5. THE DATA MINING MODELS:

The data mining models are of two types: Predictive and Descriptive.

The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc.

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc.

Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of itsfuturestate.Classificationisatechniqueofmappingthetargetdatatothepredefinedgroupsor classes, this is a supervise learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that map data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data in to groups or clusters. The clusters are defined by studying the behavior of the data by the domain experts. The term segmentation is used in very specific context; it is a process of partitioning of database into disjoint grouping of similar tuples. Summarization is the technique of presenting the summarize information from the data. The association rule finds the association between the different attributes. Association rule mining is a two-step process: Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend.

## 6. THEKNOWLEDGEDISCOVERYPROCESS:

Data mining is one of the tasks in the process of knowledge discovery from the database. The steps in the KDD process contains:

1. Data cleaning: It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collection.

2. Data integration: In this stage, multiple data sources, often heterogeneous, are combined in a common source.

3. Data selection: The data relevant to the analysis is decided on and retrieved from the data collection.

4. Data transformation: It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure.

5. Data mining: It is the crucial step in which clever techniques are applied to extract potentially useful patterns.

6. Pattern evaluation: In this step, interesting patterns representing knowledge are identified based on given measures.

7. Knowledge representation: It is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

## 7. DATAMINING METHODS:

The data mining methods are broadly categories as: On-Line Analytical Processing (OLAP), Classification, Clustering, Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining, Web Mining etc. These methods use different Types of algorithms and data. The data source can be data warehouse, database, flat file or text file. The algorithms may be Statistical Algorithms, Decision Tree based, Nearest Neighbor, Neural Network based, Genetic Algorithms based, Ruled based, Support Vector Machine etc. The selection of data mining algorithm is mainly depends on the type of data used for mining and the expected outcome of the mining process. The domain experts play a significant role in the selection of algorithm for data mining.

A knowledge discovery (KD) process involves preprocessing data, choosing a data- mining algorithm, and post processing the mining results. There are very many choices for each of these stages, and non-trivial interactions between them. Therefore both novices and data-mining specialists need assistance in knowledge discovery processes.

## 8. DATAMINING APPLICATION:

The data mining applications can be generic or domain specific. The generic application is required to be an intelligent system that by its own can takes certain decisions like: selection of data, selection of data mining method, presentation and interpretation of the result. Some generic data mining applications cannot take its own these decisions but guide users for selection of data, selection of data mining method and for the interpretation of the results. The multi agent based data mining application has capability of automatic selection of data mining technique to be applied. The Multi Agent System used at different levels: First, at the level of concept hierarchy definition then at the result level to present the best adapted decision to the user. This decision is stored in knowledge Base to use in a later decision-making. Multi Agent System Tool used for generic data mining system development uses different agents to perform different tasks.

A multi-tier data mining system is proposed to enhance the performance of the data mining processIt has basic components like user interface, data mining services, data access services and the data. There are three different architectures presented for the data mining system namely One-tire, Two-tire and Three-tire architecture.

Generic system required to integrate as many learning algorithms as possible and decides the most appropriate algorithm to use. CORBA (Common Object Request Broker Architecture) has features like:

Integration of different applications coded in any programming language considerably easy.

It allows reusability in a feasible way and finally it makes possible to build large and scalable system. The data mining system architecture based on CORBA is given by Object Management Group has all characteristics to accomplish a distributed and object oriented computation.

A data-centric focus and automated methodologies makes data mining accessible to non- experts. The use of high-level interfaces can implement the automated methodologies that hide the data mining concepts away from the users. A data-centric design hides away all the details of mining methodology and exposes them through high-level tasks that are goal-oriented. These goal-oriented tasks are implemented using data-centric APIs. This design makes data mining task like other types of queries that users perform on the data.

In data mining better results could be obtained if large data is available. It leads to the merging and linking of local databases. A new data-mining architecture based on Internet technology addressed this problem.Details database is thus the main issue. The text mining application for extraction of implicit attributes and explicit attributes from product descriptions documents is the main task in such applications. Naive Bayes and Expectation-Maximization these two methods of data mining are used in this context.

In another application to design effective user interfaces for consumer information system data mining can be used effectively. Consumers use compensatory and non-compensatory decision strategies when formulating their purchasing decisions. Compensatory decision-making strategies are used when the consumer fully rationalizes their decision outcome whereas non- compensatory decision-making strategies are used when the consumer considers only that information which has most meaning to them at the time of

decision. These decision-making strategies are considered while designing online shopping support tools, and personalizing the design of the user interface. The data mining methods cluster analysis and rough sets, are used to obtain consumer information needed in support of designing customizable and personalized user interface enhancements.

## 9. CONCLUSION:

Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation.

## References:

[1]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.*

[2] Larose,D.T.,"DiscoveringKnowledgeinData:AnIntroductiontoDataMining",*ISBN0-471-66657-2, John Wiley & Sons, Inc, 2005.*

[3] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", *Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.*

[4].Chapman, P.Clinton,J., Kerber, R.,Khabaza, T.,Reinartz,T., Shearer,C.andWirth,R.."CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000".

[5]. Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," *AI Magazine, American Association for Artificial Intelligence, 1996.*

[6]. Tan Pang-Ning, Steinbach, M., Vipin Kumar. "Introduction to Data Mining", *Pearson Education, New Delhi, ISBN: 978-81-317-1472-0, 3rd Edition, 2009.*

[7]. Bernstein, A. and Provost, F., "An Intelligent Assistant for the Knowledge Discovery Process", *WorkingPaperoftheCenterforDigitalEconomyResearch,NewYorkUniversityandalsopresentedatthe IJCAI 2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases.*

[8]. Baazaoui, Z., H., Faiz, S., and Ben Ghezala, H., "A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data, volume 5, ISSN 1307-6884," *Proceedings of World Academy of Science, Engineering and Technology, April 2005.*

[9]. Rantzau, R. and Schwarz, H., "A Multi-Tier Architecture for High-Performance Data Mining**,** A Technical Project Report of ESPRIT project, The consortium of CRITIKAL project, Attar Software Ltd. (UK), Gehe AG (Denmark); Lloyds TSB Group (UK), Parallel Applications Centre, University of Southampton (UK), BWI, University of Stuttgart (Denmark), IPVR, University of Stuttgart (Denmark)".

[10]. Botia, J. A., Garijo, M. y Velasco, J. R., Skarmeta, A. F., "A Generic Data mining System basicdesign and implementation guidelines*", A Technical Project Report ofCYCYT project of Spanish Government. 1998. Web Site: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.1935*

[11]Campos,M.M.,Stengard,P.J.,Boriana,L.M.,"Data-CentricAutomatedDataMining",,WebSite.:

www.oracle.com/technology/products/bi/odm/pdf/automated_data_mining_paper_1205.pdf

[12].Sirgo, J., Lopez, A., Janez, R., Blanco, R., Abajo, N., Tarrio, M., Perez, R., "A Data Mining Engine based on Internet, Emerging Technologies and Factory Automation," *Proceedings ETFA '03, IEEE Conference, 16-19 Sept. 2003. Web Site* :www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.8955

[13]BiancaV.D.,PhilippeBouladeMareüiland MartineAdda-Decker,"Identificationofforeign-accented French using data mining techniques, Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI)". Web Site :*www.limsi.fr/Individu/bianca/article/Vieru&Boula&Madda_ParaLing07.pdf*

[14]Halteren, H. van, "Linguistic Profiling for Author Recognition and Verification", *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics USA, Barcelona, Spain, Article No. 199, Year of Publication: 2004.*

[15]. Halteren, H. V., Oostdijk N., "Linguistic profiling of texts for the purpose of language verification, The ILK research group, Tilburg centre for Creative Computing and the Department of Communication and Information Sciences of the Faculty of Humanities, Tilburg University, The Netherlands." WebSite: *www.ilk.uvt.nl/~antalb/textmining/LingProfColingDef.pdf*