Qubahan Techno Journal

The Concept of Data Mining and Knowledge Extraction Techniques

Revink Masoud Abdulhakim
Department of Computer Science

Duhok, Kurdistan Region of Iraq

Abstract- Recent days, the concept of data mining and the need for it, its objectives and its uses in various fields, explain its procedures and tools, the type of data that is mined, and the structural structure of that data while simplifying the concept of databases, relational databases and the query language. Explain the benefits and uses of mining or mining data stored in specialized databases in various vital areas of society. Also, it is the process of analyzing data from different perspectives and discovering imbalances, patterns and correlations in data sets that are insightful and useful for predicting results that help you make a good decision. Let's bring back our mining example, when you plan to prospect for gold or any valuable minerals you first have to determine where you think the gold is to start digging. In the process of data mining we have the same concept. To mine data, you must first collect data from various sources, prepare it, and store it in one place, as nothing from data mining is related to the process of searching for the data itself. Currently, the company is storing data in what is called a Datawarehouse which we will talk about in a later stage in detail.

Keywords: Data Mining, Classification, Knowledge Extraction, Artificial Intelligence.

I. INTRODUCTION

Data mining (or data mining or data mining) aims to extract information that is hidden in large data blocks. Data mining is a modern technology that has imposed itself strongly in the era of informatics, and its use provides companies and institutions in all fields, civil and governmental, the ability to explore, and / or focus on, the most important information in large data blocks. Exploration techniques also focus on sensing and building.[1]

The future exploration of patterns, correlations, behavior and trends, allowing the right decisions to be assessed and taken at the right time, and appropriate solutions to problems, planning, development and modernization in all fields. [2]

Exploration techniques answer many questions, and in record time, especially those types of questions that were difficult to answer, if not impossible, using classical statistical analysis techniques, which, if any, take a long time and many analysis procedures.[1,3]

II. LITERATURE REVIEW

A. DATA MINING PROCEDURES AND TOOLS

1. **Classification**: Classification is based on machine learning, whereby each element in a group of data is classified into one of a predefined set of classes or groups. The program is developed so that it can learn how to classify data elements into groups. For example, the classification can be applied to records of employees who have left the company. In this case, the employee records are divided into two groups called "left" and "rest" and then data mining methods can classify employees into these two predefined groups.[4,5]
2. **Clustering**: This technique groups a number of objects that have similar characteristics into one group. This technique defines the groups and places the elements in each group depending on the common characteristics between the elements, while the classification methods determine the elements into predefined groups. For example, in the library, there are different topics from books, and to make the reader reach the book collection within a specific topic, grouping methods are used. So that the collection of books belonging to similar topics is kept on one shelf.[5]
3. **Prediction**: The prediction or prediction of data mining techniques that reveal the relationship between the independent and dependent variables. That is, if we want to use sales forecasting methods to predict future profits, if we consider that sales are an independent variable, then profit can be a dependent variable.[4,5]

4. **Association**: It is one of the best techniques for data mining where a pattern is discovered based on the relationship between elements in the same business process. This technique is used in market basket analysis to identify a group of products that customers buy together frequently. People based on historical sales data.

B. DATA MINING AND BUSINESS INTELLIGENCE

Data mining is a relatively modern science, and it is an extension of the science of statistical analysis and the main nerve of business intelligence science, or in all its forms, used mainly in the field of business. Prospecting science arose as a natural result of the great development that took place in the field of information systems and the great inflation of information that grows exponentially, especially after the widespread use of information systems and the accumulation of the huge amount of data that has become daily in circulation in many fields, which led to an urgent need To answer many questions and explore future knowledge, estimates and predictions.[6]

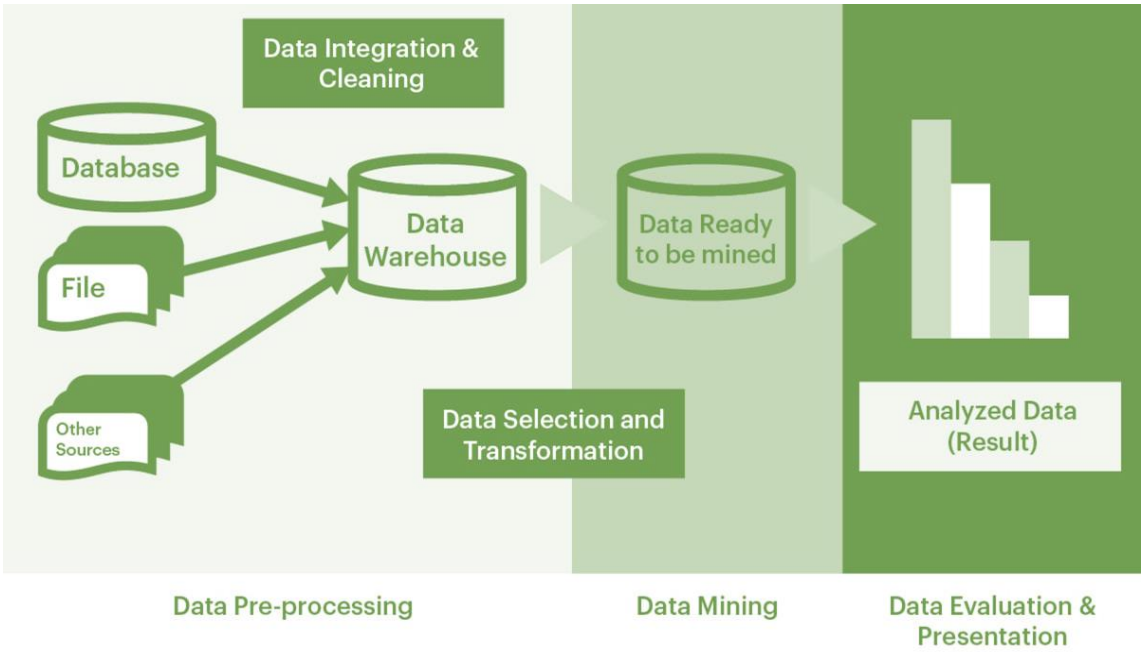

Fig 1: Data Integration and Cleaning

The analysis and exploration operations are considered one of the priorities of the planning departments in companies and institutions in the world at the present time and the ideal tool for the higher levels of management that aspire to success and ensure its continuation strategically, given the ability to produce real knowledge hidden in the large data blocks of the activities of each company or institution which It is stored and accumulated daily.[7]

C. DATA MINING NAMING ETYMOLOGY

It should be noted that the origin of the name of data mining is due to the fact that the exploration operations of huge quantities of data aim at extracting the knowledge hidden in it, a process that is very similar to the process of mining for gold and precious metals hidden in the ground, considering that knowledge is something Precious that is searched and explored in big data blocks. In short, the scientific term for data mining operations is data mining or data mining.[8]

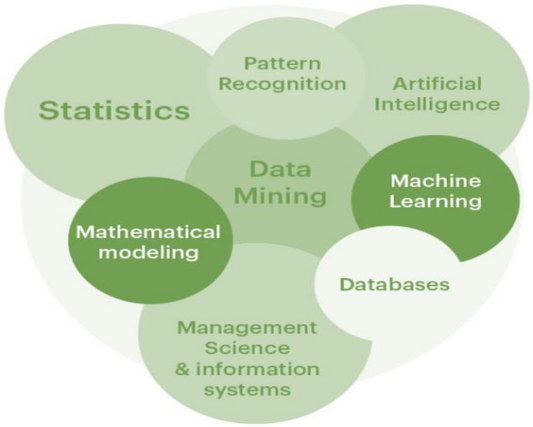

Fig 2: Data Mining Naming Etymology

III. History Of Data Mining

Data mining companies are using; To increase revenue, improve spending, target new customers, provide better customer service, and listen to what others have to say and do with competitive intelligence. And these are just some of the uses. Although the origin of data mining dates back to the 1990s. But the process of mining data to discover hidden patterns and anticipate results has a long history. Sometimes referred to as "knowledge discovery", the term "data mining" was not coined until the 1990s. But its foundation has an intertwining of several scientific disciplines: statistics (the numerical study of data relationships), artificial intelligence (human-like intelligence exposed by machines), machine learning (algorithms that learn from data to make predictions) and the cognitive scope of business. As a result of the growth in this area, in 1999 several large companies began working together to establish formal standards and forms of approach to data mining. The result of their work was CRISP-DM, which stands for the industry-standard process of data mining.[9]

IV. Stages Of Knowledge Discovery:

Knowledge Discovery in Database (KDD) is not an easy process that some people may think stops when data is collected and managed. Rather, we see it as extending to analysis, anticipation and prediction of what will happen in the future. Data mining is part of knowledge discovery, and this process is the most comprehensive. The knowledge discovery process includes the following steps:[10,11]

i. Data Discovery:

It is the stage of data collection and includes the detection, identification and characterization of available data

.

ii. Data Cleaning And Cleaning:

Noise and unimportant data are removed at this stage, and conflicting data and inconsistent data are deleted.

iii. Data integration:

In this stage, similar and related data are collected from multiple data sources and merged together.

iv. Data selection:

In this stage, the appropriate data are identified and retrieved from the dataset.

v. Data transformation:

In this stage, the data is converted into custom forms that are suitable for search and retrieval procedures by means of the achievement feed or grouping operations.

vi. Data mining:

That is, the use of intelligent methods applied to extract patterns of data and the extraction of useful models as possible.

vii. Pattern evaluation:

It is at this stage that the really important patterns that represent the knowledge base for the use of some important metrics are identified.

viii. Knowledge presentation:

It is the last stage of knowledge discovery in databases, and it is the stage that the user sees. This basic stage uses the visual method to help the user understand and interpret the results of data extraction.

V. Steps in Crisp-DM: Standard Data Mining Process:

1. **Understand the Business.**
2. **Understanding the Data.**
3. **Data Preparation.**
4. **Evaluation.**

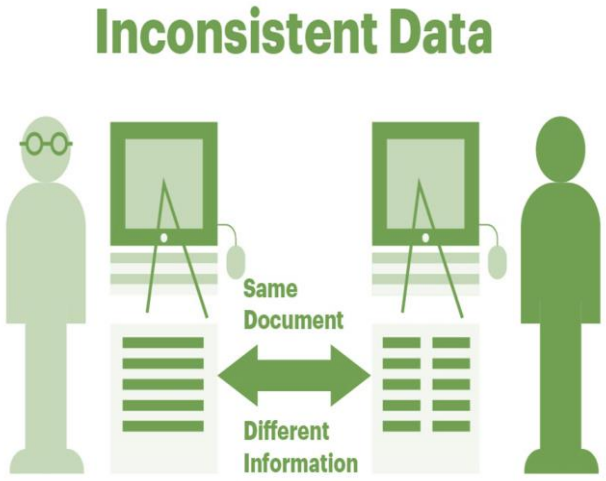

Fig 3: Inconsistent Data

5. Modeling

Model: In data mining, it is a computational representation of real word observation. Models are applications of an algorithm to search for, identify, and display any patterns or message in your data. There are two types of models in data mining:[12]

- Taxonomic or descriptive
- Predictive

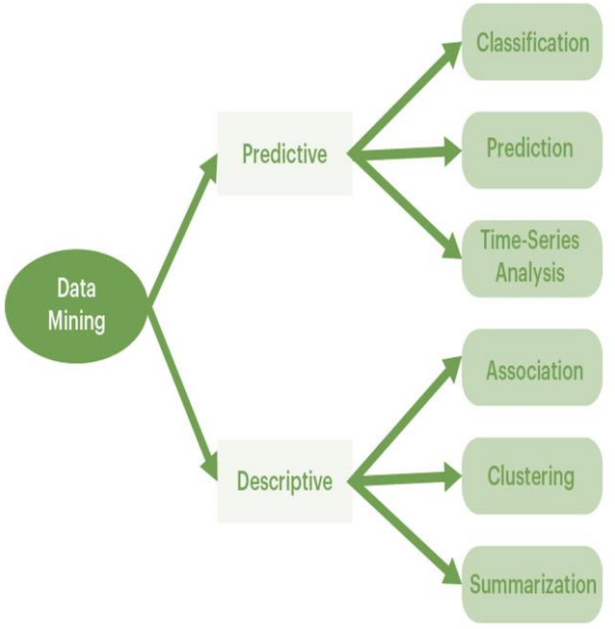

Fig 4: Data Modeling

6. Deployment

Here we need to specify how the results will be used. The knowledge gained must be organized and presented in a way that stakeholders can use. Depending on the requirements, the publishing phase can be as simple as generating a report, or as complex as implementing a data mining process that is reproducible across the organization. CRISP-DM offers a unified framework for documenting expertise and general guidance. In addition, CRISP-DM can be applied in different industries with different kinds of data.

Fig 5: CRISP-DM in Data Mining

VI. Data mining applications

1. Voice recognition:

For example, Siri on your iPhone. How does he understand your commands? Obviously, it's definitely not programmable because everyone has a different voice, tone and accent, and it not only understands but adapts better to your voice as you continue to use it more and more.[12]

2. Classification of DNA strings:

DNA sequences contain biological information. One of the many methods of DNA sequencing is through sequencing, that is, applying data mining techniques to find the patterns and then comparing them with previously studied sequences to understand the given sequence.[11,12]

3. Natural Language Processing (NLP):

Machines do not understand natural language, which is a challenge. But in recent times, real progress has been made in natural language processing after implementing data mining. So the machine learning model no longer uses manual coding for large sets of rules, but rather uses general learning algorithms to learn these rules automatically by analyzing a large set of typical examples in the real world.[12]

VII. Conclusion

Using database mining tools, we can go further, to search for, mine, and explore trends and patterns to find patterns of behavior and trends, or by using categorization and prediction that predicts what might happen in the future. For example, the customers' preferences are explored when they buy some products over other products, and the probability that customers will buy specific products is predicted according to the availability of data that is studied from historical data in the database, or, for example, the behavior of customers is explored when they buy certain products with other products. Others in the same shopping journey, and other patterns and trends that show customers' buying behavior and help build future expectations for their buying behavior. Data mining techniques and algorithms are also used extensively in the field of bioinformatics, and this scientific discipline has played a prominent role in developing specialized algorithms to serve this branch of science without other sciences. All data collected by companies will not serve any purpose without knowledge discovery. Data mining helps companies visualize hidden patterns and trends in data sets that may not have been previously visible. Whatever insights are revealed, they will lead to clear decision-making, benefiting both the companies and the customers they serve and the stakeholders.

References

1. *Amin, Mohammad Shafenoor, Yin Kia Chiam, and Kasturi Dewi Varathan. "Identification of significant features and data mining techniques in predicting heart disease." Telematics and Informatics 36 (2019): 82-93.*
2. *Alasadi, Suad A., and Wesam S. Bhaya. "Review of data preprocessing techniques in data mining." Journal of Engineering and Applied Sciences 12.16 (2017): 4102-4107.*
3. *Gorunescu, Florin. Data Mining: Concepts, models and techniques. Vol. 12. Springer Science & Business Media, 2011.*
4. *Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining concepts and techniques third edition." The Morgan Kaufmann Series in Data Management Systems 5.4 (2011): 83-124.*
5. *Rajman, Martin, and Romaric Besançon. "Text mining-knowledge extraction from unstructured textual data." Advances in data science and classification. Springer, Berlin, Heidelberg, 1998. 473-480. Available at: http://hdl.handle.net/11129/1734 .*

6. *Asaad, Renas Rajab. (2014). An Investigation of the Neuronal Dynamics Under Noisy Rate Functions. Thesis (M.S.), Eastern Mediterranean University, Institute of Graduate Studies and Research, Dept. of Computer Engineering, Famagusta: North Cyprus.*
7. *Asaad, R. R., Abdurahman, S. M., & Hani, A. A. (2017). Partial Image Encryption using RC4 Stream Cipher Approach and Embedded in an Image. Academic Journal of Nawroz University, 6(3), 40–45. https://doi.org/10.25007/ajnu.v6n3a76*
8. *Rajab Asaad, R., & Masoud Abdulhakim, R. (2021). The Concept of Data Mining and Knowledge Extraction Techniques. Qubahan Academic Journal, 1(2), 17–20. https://doi.org/10.48161/qaj.v1n2a43*
9. *Asaad, R. R., Ahmad, H. B., & Ali, R. I. (2020). A Review: Big Data Technologies with Hadoop Distributed Filesystem and Implementing M/R. Academic Journal of Nawroz University, 9(1), 25–33. https://doi.org/10.25007/ajnu.v9n1a530*
10. *Asaad, R. R. (2019). Güler and Linaro et al Model in an Investigation of the Neuronal Dynamics using noise Comparative Study. Academic Journal of Nawroz University, 8(3), 10–16. https://doi.org/10.25007/ajnu.v8n3a360*
11. *Asaad, R. R. (2021). Penetration Testing: Wireless Network Attacks Method on Kali Linux OS. Academic Journal of Nawroz University, 10(1), 7–12. https://doi.org/10.25007/ajnu.v10n1a998*
12. *Almufti, S., Marqas, R., & Asaad, R. (2019). Comparative study between elephant herding optimization (EHO) and U-turning ant colony optimization (U-TACO) in solving symmetric traveling salesman problem (STSP). Journal Of Advanced Computer Science & Technology, 8(2), 32.*
13. *Asaad, R. R., & Abdulnabi, N. L. (2018). Using Local Searches Algorithms with Ant Colony Optimization for the Solution of TSP Problems. Academic Journal of Nawroz University, 7(3), 1–6. https://doi.org/10.25007/ajnu.v7n3a193*
14. *Almufti, S., Asaad, R., & Salim, B. (2018). Review on elephant herding optimization algorithm performance in solving optimization problems. International Journal of Engineering & Technology, 7, 6109-6114.*
15. *Asaad, R. R., & Ali, R. I. (2019). Back Propagation Neural Network(BPNN) and Sigmoid Activation Function in Multi-Layer Networks. Academic Journal of Nawroz University, 8(4), 216–221. https://doi.org/10.25007/ajnu.v8n4a464*
16. *Asaad, R. R., Abdulrahman, S. M., & Hani, A. A. (2017). Advanced Encryption Standard Enhancement with Output Feedback Block Mode Operation. Academic Journal of Nawroz University, 6(3), 1–10. https://doi.org/10.25007/ajnu.v6n3a70*
17. *Abdulfattah, G. M., Ahmad, M. N., & Asaad, R. R. (2018). A reliable binarization method for offline signature system based on unique signer's profile. INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING INFORMATION AND CONTROL, 14(2), 573-586.*
18. *Almufti, S. M., Ahmad, H. B., Marqas, R. B., & Asaad, R. R. (2021). Grey wolf optimizer: Overview, modifications and applications. International Research Journal of Science, Technology, Education,and Management, 1(1),1-1.*
19. *Asaad, R. R., Sulaiman, Z. A., & Abdulmajeed, S. S. (2019). Proposed System for Education Augmented Reality Self English Learning. Academic Journal of Nawroz University, 8(3), 27–32. https://doi.org/10.25007/ajnu.v8n3a366*
20. *Asaad, R. R. (2020). Implementation of a Virus with Treatment and Protection Methods. ICONTECH INTERNATIONAL JOURNAL, 4(2), 28-34. https://doi.org/10.46291/ICONTECHvol4iss2pp28-34*
21. *Ihsan, R. R., Almufti, S. M., Ormani, B. M., Asaad, R. R., & Marqas, R. B. (2021). A survey on Cat Swarm Optimization algorithm. Asian J. Res. Comput. Sci, 10, 22-32.*
22. *Asaad, R. R., & Segerey, R. I. (2018). School Management Application Using iOS. Academic Journal of Nawroz University, 7(4), 38–44. https://doi.org/10.25007/ajnu.v7n4a269*
23. *Asaad, R. R., Mustafa, R. F., & Hussien, S. I. (2020). Mortality Statistics and Cause of Death at Duhok City from The Period (2014-2019) Using R Language Data Analytics. Academic Journal of Nawroz University, 9(3), 1–7. https://doi.org/10.25007/ajnu.v9n3a699*
24. *Asaad, R. R. (2021). A Study on Instruction Formats on Computer Organization and Architecture. ICONTECH INTERNATIONAL JOURNAL, 5(2), 18-24. https://doi.org/10.46291/ICONTECHvol5iss2pp18-24*
25. *Asaad, R. R. (2021). Virtual reality and augmented reality technologies: A closer look. Virtual reality, 1(2).*
26. *Asaad, R. R. A Review: Emotion Detection and Recognition with Implementation on Deep Learning/Neural Network.*
27. *Asaad, R. R., Saeed, V. A., & Abdulhakim, R. M. (2021). Smart Agent and it's effect on Artificial Intelligence: A Review Study. ICONTECH INTERNATIONAL JOURNAL, 5(4), 1-9.*
28. *Asaad, R. R. A Asaad, R. R. A Review: Emotion Detection and Recognition with Implementation on Deep Learning/Neural Network.*