WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Data Mining and Warehousing

Mohmmed Asharaff

Research Scholar, Universidad Azteca -Azteca University,Chalco, Mexico
byjumannar@gmail.com

ABSTRACT

The purpose of this study is to demonstrate the importance of data warehouses and data mining. It also tries to show how the data mining process can help decision-makers make more informed decisions. This study was founded on a review of the literature on data mining and data warehousing. Information from a literature review and a real-world application was used to develop the models. The most important discoveries are the phases of data mining approaches, which are demonstrated by the established model, and the usefulness of data warehousing and data mining.It can help users get more specific responses, helping both technical and nontechnical users to make better decisions. In practice, any organization with a huge amount of data will benefit greatly from data warehousing and data mining. Data warehousing and data mining can help standard (operational) databases. They also help to save millions of dollars and increase profit by making the right judgments with the help of data mining. This paper describes how data mining works and how any organization can use it to help users receive better answers from enormous amounts of data. It demonstrates an alternative approach to data querying. Data mining takes one step further than ordinary database queries by acquiring more important data.

Keywords—component; Data Mining; Data Warehousing; Operational Database

I. INTRODUCTION

Have you ever thought about the suggestions you get when you shop online? When you buy a refrigerator online, for example, the website may offer additional products to consider. Have you ever thought about your bank's warnings when you use your credit card in a different city? These are examples of data mining, which is the process of uncovering useful patterns in a huge data set. This vast volume of data is created by collecting current and historical data from numerous sources and centrally storing it in a Data Warehousing (DW) [1] repository.

A critical store for historical data and non-routine transactions is the data warehouse. Consider the historical statistics on client buy transactions at modern supermarkets, for example. Keeping this type of data in a traditional database would make it extremely huge, resulting in delayed outcomes. For these reasons, non-routine transactions and historical data should be maintained in a data warehouse for data mining [2].
Standard databases and data warehousing are developed in a variety of ways. In data warehousing, dimensional modeling approaches are used, whereas in normal database architecture, an Entity Relationship Model is used [3]. Performance is improved by multidimensional modeling (for example, star schema) [4]. Data Mining (DM) is a database and artificial intelligence combination that provides valuable information to both technical and non-technical users to assist them in making better decisions. It is frequently employed as a decision-making tool [5].
The DM procedure is not simple. There are several different sorts of feedback, and the entire process may need to be repeated at times. As a result, data mining is widely recognized as an iterative process[6]. There are six steps to it: 1) define the problem, 2) gather data, 3) explore data, 4) model, 5) evaluate, and 6) deploy [7].

Data mining can aid in the automation of knowledge extraction. This is why it's employed in so many industries, including science and business, where enormous amounts of data must be examined [8]. Web mining is one of the most common data mining applications. The Internet is becoming more and more important in our everyday life. Data mining approaches for retrieving information are becoming increasingly relevant as terabytes of data are added every day [6].

II. RESEARCH METHOD

I combined a literature review [9] on data mining [7, 10] and data warehousing [4, 11-13] with real-world findings from a case study [14] in this paper. I was able to develop a theoretical foundation for this issue because to previous research on data mining and data warehousing. My academic literature review helped me better grasp data warehousing and data mining, as well as identify essential aspects in the data mining process. The real-world findings of the case study frequently illustrate the knowledge gathered from the literature review.

I. DATA, INFORMATION AND KNOWLEDGE

- Data: facts or a summary Data and messages make up data. As input, data is fed into computers. Computers can process three categories of data: operational, non-operational, and meta data [7].
- Information:Establishing a link or connection between data can provide it. Data is transformed into information by computers [7].
- Knowledge: Patterns or links between previous and future data can be quite useful. Combining historical sales data with consumer data, for example, can give information about customers' purchase behavior [7].

II. DATA WAREHOUSING

In order to give valid information, both current and historical data need be accessible for the data mining process, yet storing historical data in a typical database will have a negative influence on the database itself. Data from the past is typically not used in day-to-day transactions, but it is used for data processing and monitoring. Storing historical data in a conventional database would result in a considerable increase in the size of the database, resulting in decreased efficiency. It's a good idea to move old data from multiple locations and combine it all in a new archive known as a data center [12].

Moving data from operational databases to a data warehouse involves three steps: The three steps are cleaning, transformation, and integration [11]. A data warehouse can have a variety of meanings. "A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile repository of data in support of management's decision-making process," according to Bill Inmon's description. [1] The following are the characteristics of a data warehouse:

- DW is subject-oriented, which means it may be used to explore any topic.
- DW incorporates current and historical data from a variety of sources.
- time-variant: DW stores historical data from various periods of time.
- DW content should not be modified because it is a non-volatile collection of data. It is information from the past.

Data warehousing is built utilizing dimensional modeling approaches, rather than the Entity Relationship Model, which is used to build ordinary databases [3]. Data warehousing modeling is a complex task. It requires 1) knowledge of market processes, 2) comprehension of the structural and behavioral system's conceptual model, and 3) knowledge of data warehousing techniques [15]. All of the data is organized into two types of tables using the dimensional modeling technique: fact tables and dimension tables (Fig 3). The process of accessing data from a data warehouse is made easier and faster with this technology [4]. In a dimensional model, there are three types of architectures based on the representation of the fact table and dimension tables: 1) star schema, 2) snowflake schema, and 3) galaxy schema [16].
Without meta data, the architecture of data warehousing is incomplete. Data is what defines a data warehouse. It is used to build, maintain, and teach users how to use a data warehouse. This information is essential to any data mining procedure [17].

III. DATA MINING

Data Mining (DM) is a database and AI technique for collecting important information from massive databases and supporting users in making better decisions. It's frequently employed as a decision-making tool [5].

A. Data Mining Usage

Humans find it impossible to interpret and extract useful information from such a massive volume of data. As a result, it is critical to employ Data Mining techniques. DM is utilized in a multitude of fields to aid in the extraction of useful data and subsequent decision-making. For example, DM could be used to market a product. It could help by providing useful information on the best media to use and when to air an ad in order to increase a product's sales. DM techniques (for example, correlation analysis) analyse all previous relevant campaign data and compare sales to produce analytical reports for policymakers to employ in order to increase potential sales [18]. The most common application of data mining is web mining [19]. With terabytes of data being added to the internet every day, there is a need for a better way to evaluate web sites and collect relevant information [6].

B. Data Mining Process

Data mining is a difficult process. It's difficult to grasp and has feedback loops, making it an iterative process. Figure 1 [7] depicts the phases of the data mining approach. It also shows that procedures can be reversed and that the entire process can frequently be repeated from the beginning. In fact, the data mining procedure is divided into six steps: (1) Define the problem, (2) Prepare the data, (3) Explore the data, (4) Model, (5) Evaluate, and (6) Deploy the solution.

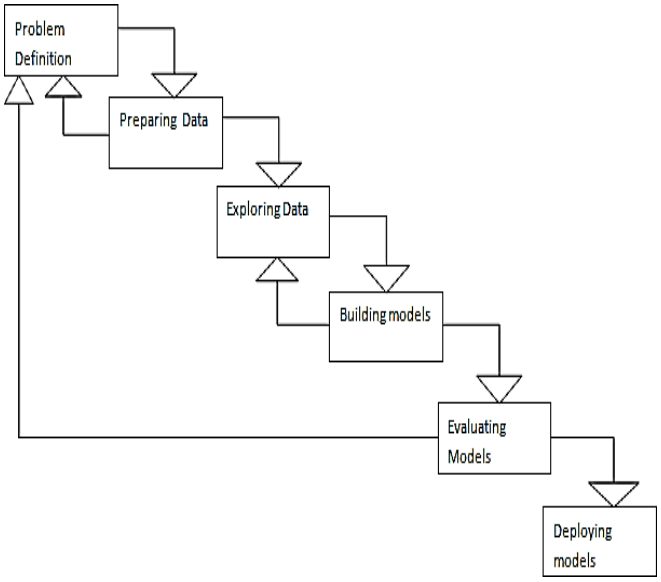


Figure 1: The Data Mining process model

C. *Explanation/Discussion of Model*

1) Data Mining Process - Goal

Data mining is a distinct type of data querying than regular querying. Data mining goes a step farther than normal database searches by obtaining more important information from a huge number of datasets. It's not an easy procedure. It comprises numerous phases of ideas between them, and the entire process can be reproduced from the start in order to get clearer responses that match the decision makers' needs [20].

2) Problem Definition

The initial step in the data mining procedure is to comprehend the market difficulty. The working group should identify the project goals and company requirements at this time [10]. This move can also be used to explain the model parameters that will be utilized to test it. The team should collaborate on a description of the data mining conundrum [21].

3) Data Exploration

In order to properly interpret the metadata, domain specialists gather, identify, and explore data, as well as communicate with data mining and industry experts from the previous step [10]. Grasp the data will enable you to gain a better understanding of the market, which will aid in the development of the mining model [22].

4) Data Preparation

Before constructing mining models, domain specialists should create data models and fix any data-related difficulties. All bad data should be destroyed before moving on to the next phase, and any missing data should be added [23]. During this procedure, the final dataset should be prepared [10].

5) Modeling

After finishing the data discovery and planning phases, data mining professionals will begin the modeling process by selecting modeling approaches and identifying the columns of data needed to build a mining framework, which will be followed by the mining models [10]. The created model should be able to meet the specifications [24].

6) Evaluation

Data mining professionals construct numerous models, test them, and choose the best one in this step. The chosen model should be thoroughly tested before being deployed into the production area before moving further with the implementation process. If no model performs as expected, the entire approach, which is described in the problem description [10], should be repeated from the start.

7) Deployment

After it has been evaluated, the best model will be implemented into the manufacturing system. This procedure may lead to the creation of a data mining study [10]. After the deployment, data mining tasks can be done. Prediction activities, for example, can help organizations make better

decisions. The implemented paradigm should add value to the firm [25].

IV RESEARCH OUTCOME

The proposed model aids in the identification of essential aspects in the data mining process, resulting in a better knowledge of the process (steps). It also emphasizes how challenging it is. Correct outcomes from reviews and iterations are crucial. This understanding facilitates in the creation of a real-world data warehouse and the subsequent usage of data mining tools. I chose a data-intensive system for the installation and observations.The chosen system is a University Housing System, which typically contains a large quantity of data pertaining to objects such as buildings, flats, furnishings, students, and a variety of other items such as maintenance requests. I started by setting up the system's operating database. Then, in orderto leverage data mining techniques, we established a data warehouse.

A. Operational Database

1) Brief

Because university housing data must be retained in a defined order, it is advisable to use the database's aid, which makes dealing with the data relatively simple. The Housing Department's website also helps to speed the procedure. One of the most significant services that the Housing Department can give in a timely manner is the repair of any malfunction or injury that may occur. As a result, constructing a database will aid in the process's acceleration. Students can submit queries by completing an electronic form with the necessary information to be kept in the database.The requested information would subsequently be forwarded to the employee in charge of student maintenance requests. When the workers have completed their tasks, thedatabase should be double-checked for all details pertaining to the reconstruction project.

2) Functions of the System

This system's main features are as follows:

- Students can submit requests to have their goods fixed.
- The requests made by the pupils will be forwarded to an employee who is in charge of receiving them.
- The price of the repair work and the status of the student's repair request are both available to them.
- Employees can create a list of all new requests.
- Employees have the ability to display the status of the repair procedure.
- Employees have the ability to amend the mending information.

3) Design

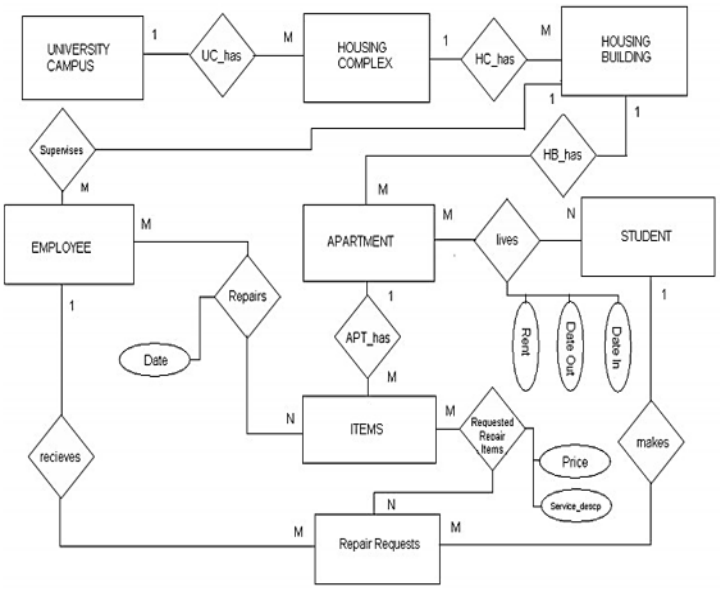


Fig 2: Developed Entity-Relationship Diagram for the proposed Operational Database (ERD)

B. Data Warehouse and Data Mining

1) Brief

This Data Warehouse focuses on "repair Request Information." It focuses on requests for repairs made by students who live in one of the university's residence halls.

This data warehouse holds the information submitted by students who request that broken objects be repaired. The following details are included:

- Information about the students.
- Information about the item.
- Location (name of the campus, complex, building number, and apartment number).
- Inquire about something.
- Information about the personnel who will be responsible for the damaged item.

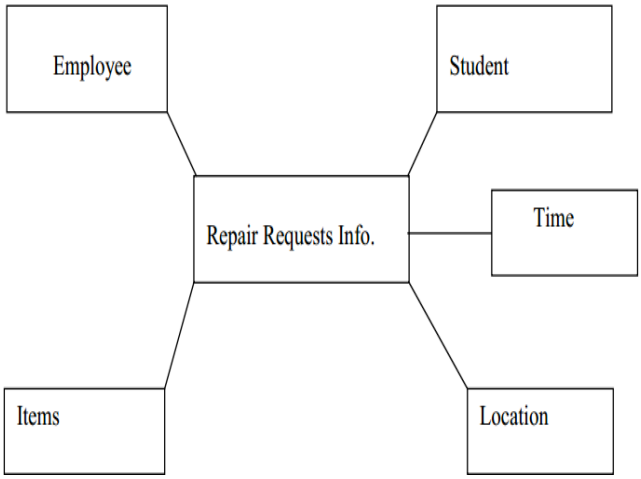
2) Proposed results

With the help of data warehousing and data processing tools, managers would be able to make better decisions. This strategy addresses a few crucial questions, allowing managers to make more informed decisions.

For example:

- What are the items that have been ordered to be fixed together?
- What objects are being requested to be repaired in order?
- What are the most popular items for which students submit requests for maintenance?
- What were the most common items that needed to be fixed last year?
- Last month, what was the name of the building with the highest maintenance requests?

3) Design



V.CONCLUSION

We now have a massive amount of data, which necessitates the usage of data warehousing and data mining. A data warehouse (also known as an operational database) is a central storage location for a subject-oriented, integrated, time-variant, and non-volatile collection of data from numerous sources [1]. For improved performance, data warehousing divides data into two architectures: fact tables and dimension tables [4]. As a result, designing a data warehouse is not the same as designing an operating database. The data warehouse is modelled using dimensional modeling (star schema, snowflake schema, or galaxy schema), whereas the operational database is modelled using entity relationships diagram [3].

Data mining has become a vital approach for obtaining useful information from today's huge volumes of data. It could also help with data extraction from the Internet, which has become an indispensable part of our daily lives. (1) Problem definition, (2) Data preparation, (3) Data exploration, (4)Modeling, (5)Evaluation, and (6)Deployment are the six steps [7]. It's an iterative process with feedback at each stage and the necessity to restart from the beginning on occasion. During the mining process, iterations are required to deliver clearer answers that consumers may use to make better judgments.

Data mining techniques are appealing for usage in a range of sectors, especially science and industry with vast amounts of data, due to their potential

to automate them and the extra value of employing them [8]. Data mining is a complex method of analyzing and searching data. It takes things a step further by finding relevant links in records, such as hidden trends or relationships [26]. Web mining [6] is one of the most prominent data mining applications. It assists in the extraction of useful information from the massive amount of information available on the internet.

REFERENCES

[1] Chen, Y. and L.-l. Qu. The Research of Universal Data Mining Model SYSTEM BASED on Logistics Data Warehouse and Application. InManagement Science and Engineering, 2007. ICMSE 2007. International Conference on. 2007.

[2] Viqarunnisa, P., H. Laksmiwati, and F.N. Azizah. Generic data model pattern for data warehouse. in Electrical Engineering and Informatics (ICEEI), 2011 International Conference on. 2011.

[3] ElDahshan, K.A. and H.M.S. Lala. Mining uncertain data warehouse. in Internet Technology and Secured Transactions (ICITST), 2010 International Conference for. 2010.

[4] Nimmagadda, S.L. and H. Dreher. On designing multidimensional oil and gas business data structures for effective data warehousing and mining. in Digital Ecosystems and Technologies, 2009. DEST '09. 3rd IEEE International Conference on. 2009.

[5] Trifan, M., et al. An ontology based approach to intelligent data mining for environmental virtual warehouses of sensor data. in Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2008. VECIMS 2008. IEEE Conference on. 2008.

[6] Dongkwon, J. and M. Songchun. Scalable Web mining architecture for backward induction in data warehouse environment. in TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. 2001.

[7] Bora, S. Data mining and ware housing. in Electronics Computer Technology (ICECT), 2011 3rd International Conference on. 2011. IEEE.

[8] Yi, L. and P. Yongjun. Application of Digital Content Management System Based on Data Warehouse and Data Mining Technology. in Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on. 2012.

[9] LePine, J.A. and A. Wilcox-King, EDITORS'COMMENTS: DEVELOPING NOVEL THEORETICAL INSIGHT FROM REVIEWS OF EXISTING THEORY AND RESEARCH. Academy of Management Review, 2010. **35**(4): p. 506-509.

[10] Huifang, Z. and P. Ding. A knowledge discovery and data mining process model in E-marketing. in Intelligent Control and Automation (WCICA), 2010 8th World Congress on. 2010.

[11] Zhen, L. and G. Minyi. A proposal of integrating data mining and on- line analytical processing in data warehouse. in Info-tech and Info-net, 2001. Proceedings. ICII 2001 - Beijing. 2001 International Conferences on. 2001.

[12] Nimmagadda, S.L., et al. On new emerging concepts of modeling petroleum digital ecosystems by multidimensional data warehousing and mining approaches. in Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on. 2010.

[13] Krippendorf, M. and S. Il-Yeol. The translation of star schema into entity-relationship diagrams. in Database and Expert Systems Applications, 1997. Proceedings., Eighth International Workshop on. 1997.

[14] Eisenhardt, K.M., Building Theories from Case Study Research. The Academy of Management Review, 1989. **14**(4): p. 532-550.

[15] Usman, M. and R. Pears. A methodology for integrating and exploiting data mining techniques in the design of data warehouses. in Advanced Information Management and Service (IMS), 2010 6th International Conference on. 2010.

[16] Sung Ho, H. and P. Sang-Chan. Data modeling for improving performance of data mart. in Engineering and Technology Management, 1998. Pioneering New Technologies: Management Issues and Challenges in the Third Millennium. IEMC '98 Proceedings. International Conference on. 1998.

[17] Yuekun, M., et al. Implementation of Metadata Warehouse Used in a Distributed Data Mining Tool. in Challenges in Environmental Science and Computer Engineering (CESCE), 2010 International Conference on. 2010.

[18] Yun, Z. and L. Weihua. AHP Construct Mining Component strategy applied for data mining process. in Information Science and Technology (ICIST), 2012 International Conference on. 2012.

[19] Xujuan, Z., et al. Using Information Filtering in Web Data Mining Process. in Web Intelligence, IEEE/WIC/ACM International Conference on. 2007.

[20] Bianchi-Berthouze, N. and T. Hayashi. Subjective interpretation of complex data: requirements for supporting the mining process. in Systems, Man and Cybernetics, 2002 IEEE International Conference on. 2002.

[21] Chieh-Yuan, T. and T. Min-Hong. A dynamic Web service based data mining process system. in Computer and Information Technology, 2005. CIT 2005. The Fifth International Conference on. 2005.

[22] Ding, P. A formal framework for Data Mining process model. in Computational Intelligence and Industrial Applications, 2009. PACIIA 2009. Asia-Pacific Conference on. 2009.

[23] Tsumoto, S., et al. Exploratory temporal data mining process in hospital information systems. in Cognitive Informatics & Cognitive Computing (ICCI*CC), 2012 IEEE 11th International Conference on. 2012.

[24] Ding, P. Data mining process model for marketing and CRM. in Machine Learning and Cybernetics (ICMLC), 2010 International Conference on. 2010.

[25] Gang, K. and P. Yi. A Standard Process for Data Mining Based Software Debugging. in Networked Computing and Advanced Information Management, 2008. NCM '08. Fourth International Conference on. 2008.

[26] Brohman, M.K. Knowledge Creation Opportunities in the Data Mining Process. in System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on. 2006.