ASEE 2014 Zone I Conference, April 3-5, 2014, University of Bridgeport, Bridgpeort, CT, USA.

Data Mining and Warehousing

Ali Radhi Al Essa
School of Engineering
University of Bridgeport
Bridgeport, CT, United States
aalessa@my.bridgeport.edu

Bach, Christian
School of Engineering
University of Bridgeport
Bridgeport, CT, United States
cbach@bridgeport.edu

***Abstract*—The aim of this paper is to show the importance of using data warehousing and data mining nowadays. It also aims to show the process of data mining and how it can help decision makers to make better decisions. The foundation of this paper created by doing a literature review on data mining and data warehousing. The models developed based on the knowledge gained from the literature review and a real case implementation. The most important findings are the phases of data mining processes, which are highlighted by the developed model, and the importance of data warehousing and data mining. It can help to get better answers which allow both technical and nontechnical users to make much better decisions. Practically, data warehousing and data mining is really useful for any organization which has huge amount of data. Data warehousing and data mining help regular (operational) databases to perform faster. They also help to save millions of dollars and increase the profit, because of the correct decisions made with the help of data mining. This paper shows the process of data mining and how it can be used by any business to help the users to get better answers from huge amount of data. It shows an alternative way of querying data. Instead of doing regular queries from regular databases, data mining goes further by extracting more useful information.**

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

I. INTRODUCTION

Have you ever think about the recommendations you get when you shop online. If you purchase for example a TV online, the website recommends you another products that you really need to get. Also have ever think about the alerts you get from your bank when you do a sudden use of your credit card in a different city. Actually these are examples of data mining which is the process of discovering useful patterns in a huge data set. This huge data is created by integrating current and historical data from different sources and store them centrally in a special repository called Data Warehousing(DW) [1].

DW is a very important repository especially for the historical data and non-every-day transactions. For example, the old data about the purchase transactions made by customers at a modern supermarkets. Keeping this kind of data in a regular database will make it very huge and then slower performance. For those reasons the historical data and non-every-day transactions should be archived in a data warehouse for data mining purposes [2]. The ways of designing data warehousing and regular databases are different. Data warehousing design depends on a dimensional modeling techniques and a regular database design depends on an Entity Relationship model [3]. The multidimensional modeling (e.g. star schema) provides faster performance [4].

Data Mining (DM) is a combination of Database and Artificial Intelligent used to provide useful information to both technical and non-technical users which will help them to make better decisions. It is usually used as a decision support system [5].

DM is not an easy process. It has several feedback and sometimes the whole process needs to be repeated. For that reason the data mining process is considered as an iterative process[6]. It involves six phases: 1) problem definition, 2) data preparation, 3) data exploration 4) modeling, 5) evaluation, and 6) deployment [7].

Data Mining can automate the process of extracting information. This is why it is used in different areas, especially science and business where it is important to analyze huge amount of data [8]. One of the most common use of data mining is web mining. The Internet becomes more important and part of our life. While terabytes of data being added every day, extracting the information using the data mining techniques becomes very important [6].

II. RESEARCH METHOD

The research method used in this paper is a combination of a literature review [9] on data mining [7, 10] and data warehousing [4, 11-13] and real world observations of a case study [14] made by myself. The previous studies done on the data mining and data warehousing helped me to build a theoretical foundation of this topic. The academic literature review, that I have done, improve my understanding of the data warehousing and data mining and then help me to identify the main factors of the data mining process. Also the real world observations of the case study emphasize the understanding gained from the literature review.

III. DATA, INFORMATION AND KNOWLEDGE

- Data: facts or description. Numbers and texts are considered data. Computers take data as an input. There are different formats of data which can be processed by computers:

operational, non-operational and Meta data [7].

- Information: It can be provided by having relationship or association among data. Computers process data into information [7].
- Knowledge: useful patterns or relationship between historical and future information. For example, historical sales information with information about customers can provide knowledge of customer's buying behaviors [7].

IV. DATA WAREHOUSING

To get accurate results from data mining process, current and historical data should be available for the process but keeping the historical data in a regular database would cause a negative effect on the database itself. Usually old data is not used for everyday transactions but it is used for the data mining and reporting issues. Storing historical data in everyday database will cause a huge increase of its size which leads to a slower performance. A good practice is to move the old data from different sources and integrate the whole in another repository called data warehouse [12]. Moving the data from operational databases to a data warehouse involves three steps: 1) cleaning, 2) transformation, and 3) integration [11].

Data warehouse has more than one definitions. The most common one is defined by Bill Inmon who defined it as the following : "A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process" [1].
As defined , any data warehouse (DW) should have the following characteristics :

- subject-oriented: DW can be used to analyze any subject.
- integrated: DW integrates current and historical data from different sources.
- time-variant: DW keeps historical data of different time.
- non-volatile collection of data: content of DW should not be changed. It is historical data.

Unlike the modeling techniques used to design regular databases - Entity Relationship model, data warehousing is designed by using dimensional modeling techniques [3]. Data warehousing modeling is complex. It needs: 1) knowledge of the business processes , 2) Understanding the structural and behavioral system's conceptual model, and 3) being familiar with data warehousing techniques [15]. The dimensional modeling technique organizes all the data into 2 types of tables – fact table and dimension tables (Fig 3). This technique makes the process of retrieving data from data warehouse easier and faster [4]. According to the representation of the fact table and the dimension tables, there are 3 types of architectures in dimensional model: 1) star schema, 2) snowflake schema, and 3) galaxy schema [16]. Meta data is an important part of the data warehousing architecture. It is data that describes the data warehouse. It is used to build, manage and tell how to use the data warehouse. This data is the basic for any data mining process [17].

V. DATA MINING

Data Mining (DM) is a combination of Database and Artificial Intelligent used to extract useful information from huge amount of datasets to help the users to make better decisions. It is usually used as a decision support system [5].

A. Data Mining Usage

Having enormous volume of data, makes it very difficult for human to analyze and get useful information. This causes the importance of using Data Mining techniques. DM is used in different areas to help to extract useful information then make better decisions. For example, DM can be used for marketing purposes. It can help by giving useful information about the best media and time to publish an advertisement which would help to increase the sales of a product. DM techniques (e.g. association analysis) check all the historical related marketing data and compare the sales to provide informative reports to be used by the decision makers then increase the future sales [18].

The most common use of data mining is the web mining [19]. As terabytes of data added every day in the internet , makes it necessary to find a better way to analyze the web sites and to extract useful information [6].

B. Data Mining Process

Data mining process is not an easy process. It is complicated and has feedback loops which make it an iterative process. Figure 1 [7] shows the steps of data mining process. It also shows that the steps might be repeated and sometimes it is possible to restart the entire process from the beginning. Actually , the data mining process involves six steps: (1) Problem definition, (2) Data Preparation, (3) Data Exploration, (4) Modeling, (5) Evaluation, and (6) Deployment.

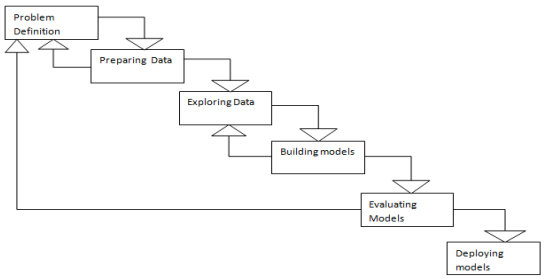

Figure 1 : The Data Mining process model

C. Explanation/Discussion of Model

1) Data Mining Process - Goal

"The Data Mining process is not a simple function, as it often involves a variety of feedback loops since while applying a particular technique, the user may determine that the selected data is of poor quality or that the applied techniques did not produce the results of the expected quality. In such cases, the user has to repeat and refine earlier steps, possibly even restarting the entire process from the beginning. [7] p. 2"

Data mining is an alternative way of querying data. Instead of doing regular queries from regular databases, data mining goes further by extracting more useful information from huge amount of datasets. It is not an easy process. It involves several phases and feedbacks between the phases and sometimes the whole process might be repeated from the beginning in order to give better answers that meet the decision makers expectations [20].

2) Problem Definition

"A data-mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required. [7] p. 2"

The starting point of the data mining process is to understand the business problem. In this step, the project objectives and the business requirements should be defined by working team [10]. Also, the model metrics should be defined in this step to be used to evaluate the model. The team should work together to form the definition of the data mining problem [21].

3) Data Exploration

"Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics are used to explore the data. [7] p. 3"

In this phase domain experts collect, describe and explore data also exchange with the data mining and business experts from the previous phase in order to understand the meaning of the metadata perfectly [10]. Understanding the data, helps to understand the business more which will play an important role in designing the mining model [22].

4) Data Preparation

"Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed. [7] p. 3"

Before building the mining models, domain experts should build the data models and fix any problem related to the data. Any bad data should be removed and any missing data should be brought before moving to the next phase [23]. In this phase the final dataset should be prepared [10].

5) Modeling

"Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required. [7] p. 3"

After completing the data exploration and preparation phases, data mining experts can start the modeling phase by selecting modeling techniques and defining the columns of data needed to build a mining structure and then the mining models [10]. The developed model should meet the expectation [24].

6) Evaluation

"Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions: 1. Does the model achieve the business objective? 2. Have all business issues been considered? At the end of the evaluation phase, the data-mining experts decide how to use the data mining results. [7] p. 3"

In this phase, data mining experts build more than one model and test them and then select the best one among them. Before the deployment phase, the selected model should be evaluated carefully before deploying it into the production environment. If no model perform as expected, the entire process could be repeated from the beginning which is the problem definition [10].

7) Deployment

"Deployment : Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets. The Intelligent Miner™ products assist you to follow this process. You can apply the functions of the Intelligent Miner products independently, iteratively, or in combination. [7] p. 3"

After evaluating the best model, it can be deployed into the production environment. A report about the data mining should be generated in this phase [10]. After the deployment, the data mining tasks can be done. For example, prediction tasks which help the business to make better decisions. The deployed model should add value to the business [25].

VI. RESEARCH OUTCOME

The developed model advances the understanding of the data mining process which helps to identify the main factors (steps) of the process. It also shows how complicated it is. The feedbacks and the iterations are very important to provide accurate results. This understanding helps to implement a real case example of a data warehouse and then apply the data mining techniques. We picked a system with huge amount of data to do the implementation and the observations. The chosen system is a University Housing System which usually has huge amount of data like data related to objects as buildings, apartments, furniture, students and many other things like maintenance requests. First we implemented the operational database for the system. Then we built the data warehouse to apply the data mining techniques.

A. Operational Database

1) Brief

As the data of a university housing has to be maintained in a finite order, it is better to take a help provided by a database, due to which the dealing with the information becomes very simple. Also having a database for the Housing Department helps to make the work more efficient. One of the most important things that the Housing Department should provide in a timely manner is to fix any problem or damage that may happen. So, having the database can speed up the process. Students can make requests by filling an electronic form with some necessary information to store in the database. Then the employee who is responsible to receive the repairing requests from the students will receive the requested information. When the workers are done from the work, the database should be updated by providing all the information which is related to repairing process.

2) Functions of the System

The main functions that are done by this system are :

- Students can make requests in order to get their items repaired.
- The requests given by the students will go to an employee who is responsible to receive the requests
- Student can know the price for the repair work done and the status of his/her repair request.
- Employees can list all the new requests .
- Employees can show the status of the repairing process.
- Employees can update the repairing information.

3) Design

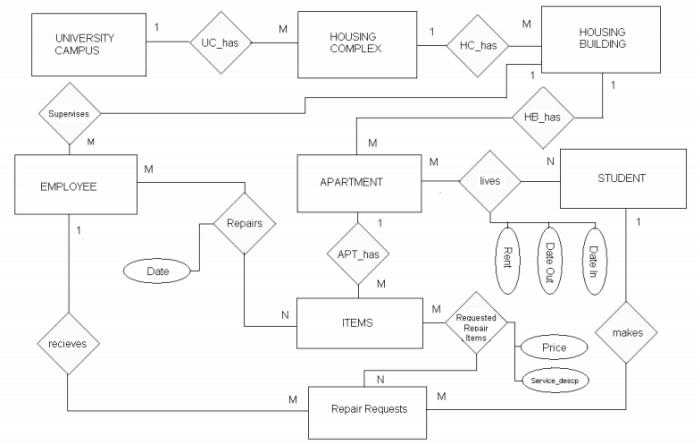

Fig 2: Developed Entity-Relationship Diagram for the proposed Operational Database (ERD)

B. Data Warehouse and Data Mining

1) Brief

The subject of this Data Warehouse is "repair Request Information". It is based on the repair requests made by students live in one of the housing buildings in one of the university campuses.

This data warehouse stores the information provided by students asking to repair damaged items. Information includes:

- Students information.
- Items information.
- Location (campus name, complex name, building #, and apartment #).
- Request information.
- employees Information who are going to take care of the damaged item.

2) Proposed results

This data warehousing with the data mining techniques will help managers to make better decisions. This system gives answers for some important questions that can help managers to make better decisions. For example:

- What items are requested to be repaired together?
- What items are requested to be repaired after each other?
- What items do students make repair requests more?
- What were the most requested items to be repaired last year?

- What was the building name of the largest number of repair requests last month?

3) Design

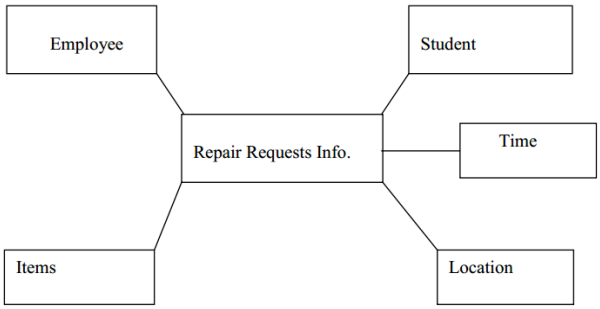

Fig 3 : Developed Star Schema for the proposed Data warehouse

VII. CONCLUSION

Nowadays we have enormous volume of data which lead to the necessity of using data warehousing and data mining. Data warehouse is used as a central store of a subject-oriented, integrated, time-variant and non-volatile collection of data from different sources (operational databases) [1]. For faster performance, data warehousing organizes data in a different architecture – fact table and dimension tables [4]. For that reason modeling the data warehouse is unlike modeling the operational database. A dimensional modeling is used to model the data warehouse (star schema, snowflake schema, or galaxy schema) but the operational database uses entity relationships diagram [3].

Data mining has became an important tool which can extract useful information from the huge amount of data we have nowadays. It also may help to extract information from the Internet which becomes part of our life. It is a complicated process. It involves six phases: (1) Problem definition, (2) Data Preparation, (3) Data Exploration, (4)Modeling, (5)Evaluation, and (6)Deployment [7]. It is an iterative process which includes feedbacks between the phases and sometimes needs to repeat the entire process from the beginning. The iterations are needed in the mining process in order to provide better answers which will be used by the users to make better decisions.

The ability of automation the data mining techniques and the value added of using it, make it attractive to be used in different areas especially science and business areas with huge amount of data [8]. Data mining provides a smart way of analyzing and querying data. It goes further by finding useful relationships in data even hidden relationships or patterns [26]. Web mining is one of the most common example of data mining usage [6]. It helps to extract useful information from the tons of information in 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. in Management 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.