IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

A Tool for KDD: Data Mining

Upinder Kaur[1], Payal Jain[2]

Assistant Professor, Department of Computer Science, G.M.N. College, Ambala Cantt (Haryana) [1]

Lecturer, Department of Computer Science and Application, G.M.N. College, Ambala Cantt (Haryana) [2]

Abstract: With the rapid development of computer and information technology in the last decade, an enormous amount of data has been and will continuously be generated in massive scale. All the large MNC's and organizations rely on database to manage their data and information. These databases are useful for conducting daily business transactions. Also, Data warehousing and data mining are essential elements of decision support, which has increasingly become a focus of database industry. This paper provides an overview of data warehousing and data mining, with an emphasis that how the data are analyzed to derive effective business strategies and discover better ways in carrying out business. This paper also describes various types of data warehousing and data mining methods. Also in this paper we did research on the top trends in data warehousing such as Hadoop, customer experience strategies to improve sales and service etc. We also described some examples and applications of data mining which we see in our day to day life like how data mining is used in Healthcare, data market basket, education system and social media. We described some of the current tools and techniques available at present for data warehousing in terms of the front end and backend tools. We further analyzed problems and issues and identified some of the research areas in the field of data warehousing.

Keywords: Data warehousing, Data Model, analysis, repository, KDD, logistics, Derived model.

1. INTRODUCTION

Now days, almost every enterprise uses a database to store its vital data and information. In today's world many MNC's and major organizations are operated in different countries. Each place of operation may generate large volumes of data. This data is strategically used in building reports to make decisions by decision makers of the company's. This data is available in various forms; it may be in the graphical form, documents, and video or in the form of array. The competition in the marketplace has led business managers and managements to seek a new way to increase their profit and market power, and that by improving their decision making processes. In this sense, the idea of data warehouse and data mining was born. In fact, data warehousing is the process of collecting data from operational functional databases, transforming, and then archiving them into special data repository called data warehouse with the goal of producing accurate and timely management information; whereas, data mining is the process of discovering trends and patterns from data warehouse, useful to carry out data analysis.

2. DATA WAREHOUSING

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker like manager, decision maker, or analysts to make better and faster decisions. Data warehousing technologies have been successfully used in many industries: manufacturing for order shipment and customer support, retail for user profiling and inventory management, financial services for claims analysis, risk analysis, credit card analysis, and fraud detection, transportation for fleet management, telecommunications for call analysis and fraud detection, utilities for power usage analysis, and healthcare for outcome analysis. Architecture of Data warehouse Data Model: In this model we receive data from various sources and in different forms like operational data, flat files, ERP and CRM data. This data is then processed into an ETL system, where basically data is extracted, transformed and loaded. After that data is stored in data warehouse as Meta data, raw data or summary data as required. Once we have the data in the data warehouse then it is used by business intelligence people to generate Reports, Online Analytical Processing (OLAP) and Data Mining purposes. These reports which are generated through Data Warehouse tools have been tremendously use to make business decisions by companies.

2.1. Example of Data Warehousing

Here we have explained a great example of data warehousing that everyone can relate to is what Facebook does. Facebook basically gathers all of your data – your friends, your likes, who you stalk, etc. – and then stores that data into one central repository. Even though Facebook most likely stores your friends, your posts, etc., in separate databases, they do want to take the most relevant and important information and put it into one central aggregated database. Why would they want to do this? For many reasons – they want to make sure that you see the most relevant ads that you're most likely to click on, they want to make sure that the friends that they suggest are the most relevant to you, also keep in mind that this is the data mining phase, in which meaningful data and patterns are extracted from the aggregated data. But, underlying all these motives is the main motive: to make more money – after all, Facebook is a business.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

3. DATA MINING

Now data warehousing is a process that must occur before any data mining can take place. In other words, data warehousing is the process of compiling and organizing data into one common database, and data mining is the process of extracting meaningful data from that database. The data mining process relies on the data compiled in the Data Warehousing phase in order to detect meaningful patterns. In the Facebook example that we gave, the data mining will be done by business users who are not IT engineers, but who will most likely receive assistance from engineers when they are trying to manipulate their data. The data warehousing phase is a strictly engineering phase, where no business users are involved. And this gives us another way of defining the 2 terms: data mining is done by business users with the assistance of engineers, and data warehousing is done exclusively by engineers.

4. APPLICATIONS OF DATA WAREHOUSE

a) Retail Sales - Data is collected at several interesting places in a grocery store. Some of the most useful data is collected at the cash registers as customers purchase products. Modern grocery store scans the bar codes directly into the point of sale (POS) system. Also the data is collected when vendors make deliveries, is another interesting data collection point which gives us the inventory of the store. At the grocery store, management is concerned with logistics of ordering, stocking, and selling products while maximizing profit. Some of the most significant management decisions are on pricing and promotions. Both store management and marketing spend a great deal of time tinkering with pricing and promotions. Some of the big retail companies which are using data are Wal-Mart, Big Bazar, Target etc.

b) Telecommunications - A telecommunications company generates hundreds of millions of call-detail transactions in an year. For promoting proper products and services, the company needs to analyze these detailed transactions. The data warehouse for the company has to store data at the lowest level of detail.

c) Logistics or Transportation – Airlines gives offer and rewards while you fly these days based on the data of frequent flyer. The department is interested in seeing which flights the company's frequent flyers take, which planes they fly, what fare basis they pay, how often they upgrade, how they earn. These requirements can be fulfilled by data warehouse.

4) Education - Utilizing a decision support system is a proactive way to use data to manage, operate, and evaluate educational institute in a better way. Depending on the quality and availability of the underlying data, such a system addresses a wide range of problems by distilling data from any combination of education records maintenance system.

The data mining from data warehouse can be a ready and effective system for the decision makers. That can help in improving the educational system. Also that helps students to find best School or College according to their interests.

5. PROCESS OF DATA MINING

Data mining is used to extract implicit and previously unknown information from data. Data mining is the process which provides a concept to attract attention of users due to high availability of huge amount of data and need to convert such data into useful information. So, many people use the term "knowledge discovery device" or KDD for data mining. Knowledge extraction or discovery is done in seven sequential steps used in data mining:

i) **Data cleaning**: we remove noise data and irrelevant data from collected raw data, at this step.
ii) **Data integration:** At this step, we combine multiple data sources into single data store called target data.
iii) **Data Selection**: Here, data relevant to analysis task are retrieved from data base as pre-processed data.
iv) **Data transformation**: Here, data is consolidating into standard formats appropriate for mining by summarizing and aggregated operations.
v) **Data Mining**: At this step, various smart techniques and tools are applied in order to extract data pattern or rules.
vi) **Pattern evaluation**: At this step, strictly identify tree patterns representing knowledge.
vii) **Knowledge representation**: This is the last stage in which, visualization and knowledge representation techniques are used to help users to understand and interpret the data mining knowledge or result.

The goal of knowledge discovery and data mining process is to find the patterns that are hidden among the huge set of data and interpret useful knowledge and information.

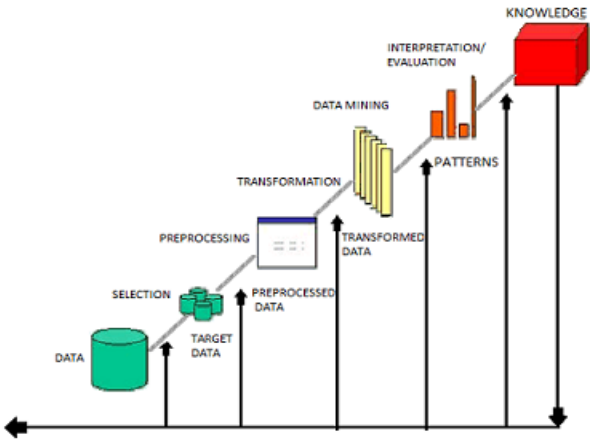

Fig 1. Data Mining Process

In the diagram data mining is the main part of knowledge discovery process.

DOI 10.17148/IJARCCE.2016.51188

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

6. APPLICATIONS OF DATA MINING

1. **Marketing:** Customer profiling, retention, identification of potential customer, market segmentation.
2. **Fraud detection:** Identify credit card fraud and intrusion detection.
3. **Scientific data analysis:** Identify the research decision making data.
4. **Text and web mining:** used to search text or information on web or given raw data.
5. Any other applications that involve large amount of data.

7. DATA MINING TECHNIQUES

There are various major data mining techniques that have been developed and used in data mining projects recently including **association, rule classification, clustering, prediction and evaluation pattern etc.,** are used for knowledge discovery from database.

i). Association: It is one of the most popular data mining techniques. In this technique we mine frequent patterns lead to discovery of interesting association and correlations within data.

Example:
Association technique is used in marketing analysis to identify items which are frequently purchased within the same transactions.
An example of such a rule, mined from the All Electronics transactional database, is

buys(X; "computer"))buys(X; "software") [support = 1%; confidence = 50%]

where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all of the transactions under analysis showed that computer and software were purchased together as single-dimensional association rules. Dropping the predicate notation, the above rule can be written simply as **"computer) software [1%, 50%]".**

ii). Classification: It is the process of finding a model or function that describes & distinguish data classes or concepts for the purpose of being able to use the model to predict the class of object whose class label is unknown.
In classification, we make software that can learn how to classify the data items into group. Derived model can be presented as classification or rules. So, Classification techniques:

1. Regression
2. Distance
3. Decision
4. Rules
5. Neural networks

iii). Clustering: This involved seeking to identify a finite set of categories and grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters. This technique has been applied in many fields, for example:

• **Healthcare:** clustering categories and attributes used in analyzing the similarities between community health centers [1].

• **Retailing:** clustering the segmentation for possible product line and brand extension to identify market to customer clusters [2];

• **Financial/Banking:** identifying groups of corporate bond clusters according to the industry and a specific segment within an industry; then tuning cluster data for each industry as a template for predicting rating changes [3].

• **Construction Industry:** clustering textual data to discover groups of similar access patterns [4].

• **Collaboration and Teamwork:** identifying groups of workers with similar task-related information needs based on the similarities of workers' knowledge flow [5]. Common tools used for clustering include k-means, principal component analysis, the Kolmogorov-Smirnov test and the quantile range test and polar ordination.

iv). Prediction: The classification predicts categorical (discrete, unordered) labels, prediction models continuous-valued functions. That is, prediction is used to predict missing or unavailable numerical data values rather than class labels. But, the term prediction may refer to both numeric prediction and class label prediction.

Example: Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Prediction also encompasses the identification of distribution trends based on the available data.

Applications of prediction:
1. Credit approval
2. Target marketing
3. Medical diagnosis
4. Treatment effectiveness analysis

v). Evaluation Pattern:
Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.

Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of time related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Example: Evolution analysis. Suppose that you have the major stock market (time-series) data of the last several years available from the New York Stock Exchange and you would like to invest in shares of high-tech industrial companies.

A data mining study of stock exchange data may identify stock evolution regularities for overall stocks and for the stocks of particular companies.

Such regularities may help predict future trends in stock market prices, contributing to your decision making regarding stock investments.

8. USE IN KDD

This section described knowledge types in 8 organization domains for data mining collaboration process in the knowledge creation.

• **Health-care System domain,** the dataset composed of three databases: the health-care providers' database; the out-patient health-care statistics database; and the medical status database [1].Another data source was from hospital inpatient medical records [6].

• **Construction Industry domain,** A sample data set was in the form of Post Project Reviews (PPRs) as defining good or bad information [4]. Multiple Key Term Phrasal Knowledge sequences (MKTPKS) formation was generated through applications of text mining and was used an essential part of the text analysis in the text documents classification [4].

• **Retailing domain:** Customer data and the products purchased have been collected and stored in databases to mine whether the customers' purchase habits and behavior affect the product line and brand extensions or not [2].

• **Financial domain:** There were two datasets posed in financial domain: (i) to identify bond ratings, knowledge sets contained strings of data, models, parameters and reports for each analytical study; and (ii) to predict rating changes of bonds, cluster data of bond features as well as the model parameters were stored, classified, and applied to rating predictions [3].

•**Small and Middle Businesses (SMBs) domain:** Knowledge types in small and middle businesses in case of Food Company were related to the corporate conditions or goals of the problem among all departments to develop a decision system platform and then formed the knowledge tree to find relations by human-computer interaction method and optimize the process of decision making[7].

To solve food supply chain networks problems, Li et al. (2010) developed EW&PC prototype which composed of major components of: (i) knowledge base, (ii) task classifier and template approaches, (iii) DM methods library with expert system for method selection, (iv) explorer and predictor, and (v) user interface [8]. This system built decision support models and helped managers to accomplish decision-making.

• **Research Assets domain:** In Cantu & Cellbos (2010) focused on managing knowledge assets by applied acknowledge and information network (KIN) approach.

This platform contained three components types of research products, human resources or intellectual capital, and research programs. The various types of research assets were handled on domain ontologies and databases [9].

•**Business domain:** There were two types of knowledge attributes conducted: condition attributes and decision attribute [10]. Condition attributes included four independent attributes of the KM purpose, the explicit-oriented degree, the tacit-oriented degree, and the success factor. Decision attribute included one dependent attribute of the KM performance [10].

• **Collaboration and Teamwork domain:** A dataset used from a research laboratory in a research institute. It contained 14 knowledge workers, 424 research documents, and a workers' log as that recorded the time of document accessed and the documents of workers' needed [5]. For the workers' log, it was generated to 2 levels of codified-level knowledge flow and topic-level knowledge flow [5].

The two types of knowledge flow were determined to describe a worker's needs. To collect the knowledge flow, documents in the data set were categorized into eight clusters by data mining clustering approach [5].

9. CONCLUSION

In this paper we briefly discussed what Data Warehousing and Data Mining means. Also we discussed some great applications of data warehouse and data mining in today's world. We talked about the benefits by the help of some examples of Data warehousing application, how data plays a big role in large enterprises to make big decisions by managers or decision makers.

Also we researched some top trends with Data warehousing and data mining technology like Data-fication, Consolidation, and hadoop helps businesses to achieve their goals. With the help of data mining how business intelligence is effectively generating business reports which helps business decision makers to innovate something and increase the profit of the company.

Beside this we talked about some issues and challenges and how we will work on it to make our Data warehouse stronger and better for future world.

REFERENCES

[1] Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M. &Kobler, A. (2007).Data mining and visualization for decision support and modeling of public health-care resources. Journal of Biomedical Informatics, 40, 438-447. doi:10.1016/j.jbi.2006.10.003

[2] Liao, S.H., Chen, C.M., Wu, C.H. (2008). Mining customer knowledge for product line and brand extension in retailing. Expert Systems with Applications, 34(3), 1763-1776.doi:10.1016/j.eswa.2007.01.036

[3] Cheng, H., Lu, Y. & Sheu, C. (2009). An ontology-based business intelligence application in a financial knowledge management system. Expert Systems with Applications, 36, 3614–3622. Doi:10.1016/j.eswa.2008.02.047

[4] Ur-Rahman, N. & Harding, J.A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Systems with Applications, 39, 4729- 4739. doi:10.1016/j.eswa.2011.09.124

[5] Liu, D.R. & Lai, C.H. (2011). Mining group-based knowledge flows for sharing task knowledge. Decision Support Systems,50(2), 370-386. doi:10.1016/j.dss.2010.09.004

[6] Hwang, H.G., Chang, I.C., Chen, F.J. & Wu, S.Y. (2008). Investigation of the application of KMS for diseases classifications: A study in a Taiwanese hospital. Expert Systems with Applications, 34(1), 725-733. doi:10.1016/j.eswa.2006.10.018

[7] Li, X., Zhu, Z. & Pan, X. (2010). Knowledge cultivating for intelligent decision making in small & middle businesses. Procedia ComputerScience,1(1),2479-2488. doi:10.1016/j.procs.2010.04.280

[8] Li, Y., Kramer, M.R., Beulens, A.J.M., Van Der Vorst, J.G.A.J. (2010). A framework for early warning and proactive control systems in food supply chain networks. Computers in Industry, 61, 852–862. Doi:101.016/j.compind.2010.07.010

[9] Cantú, F.J. & Ceballos, H.G. (2010). A multi agent knowledge and information network approach for managing research assets. Expert Systems with Applications, 37(7), 5272-5284. doi:10.1016/j.eswa.2010.01.012

[10] Wu, W., Lee, Y.T., Tseng, M.L. & Chiang, Y.H. (2010). Data mining for exploring hidden patterns between KM and its performance. Knowledge-Based Systems, 23, 397-401. doi:10.1016/j.knosys.2010.01.014