Important and application of data mining

ASSIGN BUSTER

Important and application of Data Mining

Abstract

Today, people in business area gain a lot of profit as it can be increase year by year through consistent approach should be apply accordingly. Thus, performing data mining process can lead to utilize in assist to make decision making process within the organization. This paper elaborate in detail the level of importance and also the application the application of data mining which can be adopt for various fields depends on the objective, mission, goals and purpose of conducting the study within the organization. there are three main areas take as a example which are hotel, library and hotel to observe on how data mining works to these main field.

Keywords: Data Mining, KDD Process, Decision Trees, Ant Colony Clustering Algorithm; Association Rules, Neural Network, Rough Set,

1. 0 Introduction

As we know, organization which conducts business transaction is keeps massive of document or data in a specific database for further retrieval. The data are combine from are a few departments that carried out different task and each of their function parallel with the mission and vision of organization. According (Imberman, 2001) the number of fields in large databases can approach magnitudes of 102 to 103. Therefore, it is necessary to make proper decision making or strategic planning using the existing data where these plays important role in order to ensure any action that are taken place does not given an impact especially bring loss to the organization. Other than that, data became obsolete when it keeps on changing and easily

out dated as the user requirement shifting depends on factors such as trends, money, needs and so forth.

One way to analyze data is using of data mining technique which enable to assist organization by emphasize several steps to produce the valuable output in short period of time compare with the traditional method which may involves more than one methodologies and it derive to longer of time to accomplish the investigation towards a portion of data. Thus, in the business area an action should be done quickly in order to compete with other competitors and to improve performance both in giving service and produce a high quality product. Moreover, process interpretation of the result involves group of people to inject some of the creativity and synthesis which can lead to the solutions on the problem or tasks.

Obviously, data mining a lot assist in various fields with different purposes and depend on the objectives that want to achieve. The rest of this paper is organized as follows. Section 2 tells about definition of data mining. Section 3 determines the importance of data mining. Section 4 explains the application of data mining in various fields. Section 5 draws the conclusions.

2. 0 Definition of Data Mining

There are abroad definitions listed by a few researcher and academician according to their view and opinion based on the study they have done. Moreover, these will help to understand or giving an idea before discusses more in depth towards data mining technique.

Basically, the main purpose use of data mining is to manipulate huge amount of data either existence or store in the databases by determine

suitable variables which is contribute to the quality of prediction that will be use to solve problem. Define by Gargano & Raggad, 1999.

“ Data mining searches for hidden relationships, patterns, correlations, and interdependencies in large databases that traditional information gathering methods (e. g. report creation, pie and bar graph generation, user querying, decision support systems (DSSs), etc.) might overlook”.

Besides that, another author also agreed with opinion toward the data mining definition which is to seek hidden pattern, orientation and also trend. Through (Palace, 1996) added to the previous is:

“ Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases”.

Moreover, data mining also define as process to squeeze of knowledge or information using appropriate framework or model to analyze until produce an output that assist in fulfill the objective of the study. From Imberman, 2001:

“ As knowledge extraction, information discovery, information harvesting, exploratory data analysis, data archeology, data pattern processing, and functional dependency analysis”.

The statement above agreed and adds that the framework or model that adopt definitely to expose the real circumstance. Define by Ma, Chou & Yen, 2000:

“ Data mining is the process of applying artificial intelligence techniques (such as advanced modeling and rule induction) to a large data set in order to determine patterns in the data”.

In the other hand, data mining is taken a few steps during analysis and this step is depending on the methodology that is chosen. Each of the methodology is not much differ from other methodology. Through Forcht & Cochran, 1999:

“ Data mining is an interactive process that involves assembling the data into a format conducive to analysis. Once the data are configured, they must be cleaned by checking for obvious errors or flaws (such as an item that is an extreme outlier) and simply removing them”.

3. 0 Important of Data Mining

As discusses above, it can be seen that data mining will be beneficial a lot of party and multiple range of level in the organization as the model or framework that is apply can reduce time and cost. Then, the results allow the responsible knowledge worker to transform into the strategic value of information effectively by critically analyze the result.

The process should be done carefully to avoid the useful variables or algorithm being removes or not be included in the extraction of reliable data. Data mining techniques will help in select a portion of data using appropriate tools to filter outliers and anomalies within the set of data. According to Gargano & Raggad, 1999, there are a few others important of data mining consist of:

· To facilitate the explication of previously hidden information includes the capabilities to discover rules, classify, partition, associate and optimize.

According to (Goebel & Gruenwald, 1999) in order to seek the pattern of data, a few methodologies are use in clarify the vagueness as well as to identifying the relation among one variables and other variables within the databases whereas the outcome will guide in making decision or to forecast the impact when the action were take into consideration. The chosen of methodologies should be determined in a proper way suit with the rules and condition towards the data which is to be analyzed. The methodologies include:

- Statistical Methods: focused mainly on testing of preconceived hypotheses and on fitting models to data.
- Case-Based Reasoning (CBR): technology that tries to solve a given problem by making direct use of past experiences and solutions.
- Neural Networks: formed from large numbers of simulated neurons, connected to each other in a manner similar to brain neurons which enables the network to “ learn”.
- Decision Trees: each non-terminal node represents a test or decision on the considered data item and can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.
- Rule Induction: Rules state a statistical correlation between the occurrences of certain attributes in a data item, or between certain data items in a data set.

- Bayesian Belief Networks: graphical representations of probability distributions derived from co-occurrence counts in the set of data items.
- Genetic algorithms / Evolutionary Programming: formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism.
- Fuzzy Sets: constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems.
- Rough Sets: rough sets are a mathematical concept dealing with uncertainty in data and used as a stand-alone solution or combined with other methods such as rule induction, classification, or clustering methods

· The ability to seamlessly automate and embed some of mundane, repetitive, tedious decision steps not requiring continuous human intervention.

Several steps are taken in processes or analyzes on selected data where the process involves of filtering, transforming, testing, modeling, visualization and documented the result or store accordingly in the databases or data warehouse. Each of the steps functions differently and has responsibility in carries out the process with the purpose to easier and produce the high quality of assumption by automate generate towards specific conditions. For example, data warehouse also keep previous analysis and this allow eliminating the redundant output at certain steps. Through Ma, Chou & Yen,

2000, they stress the characteristics of data mining define how it assist to reach the end process of analyzing. It comprises:

- Data pattern determination: Data-access languages or data-manipulation languages (DMLs) identify the specific data that users want to pull into the program for processing or display. It also enables users to input query specifications. Therefore, users simply select the desired information from the menus, and the system builds the SQL command automatically.
- Formatting capability: It generates raw data formats, tabular, spreadsheet form, multidimensional-display and visualization.
- Content analysis capability: Data mining also has a strong content analysis capability that enables the user to process the specifications written by the end-users.
- Synthesis capability: Data mining allows data synthesis to be timely executed.

· Simultaneously reducing cost and potential error encountered in the decision making process.

Basically, data mining can minimize the error of forecasting by following the steps of selected methodology in well manner to avoid delaying in making decision where this situation will giving big impact for the business area. Therefore, it must be careful in handling the data throughout the steps involves whereby the strategic plan should take into consideration includes of the objectives to done the analysis, the amount of data, the variables, the relationship between variables, test adopted, and so forth. Moreover, if there is need to discuss with the professional towards the study conducted and it

should be included in the planning part. In the context of organization, usually a unit or group of people are given responsible to carries this duty to discover the hidden pattern for another department. Hence, the continuously meeting should be done between the professional and researchers to ensure the end result fulfill their requirement as well as to improve the performance of worker, department and organization.

In term of reducing a cost, compare to the traditional research which take time in acquiring the data from respondents and it depend on the methodologies that are use and the number of sampling. If the questionnaire method, it can be done quickly and less time consuming but if the interviewing method is adopted, it surely take time and researcher have to meets the respondent more than one time, if there is an ambiguity or the answers not meet with the requirement. For certain study, the sampling are involves from the different location which require the researcher to travel in order to gain the genuine opinion from them and this will cost a lot involves of accommodation, food, flight ticket and so forth. For data mining, it uses the existence of data (for example, data of customer transaction, data of student registration, data of patient undergo the operation process and so on) that keep in data warehouse which mostly reduce cost in aspect of acquiring data. Other than that, researcher take first action by search for the study in the data warehouse when the objective being determine at the beginning of study because previous study are store in the data warehouse. If it is found tally, a few step will be skip or easily decided towards the data and it prove that data mining can reducing the cost as well as time. Refer to Gargano & Raggad, 1999, data mining also derive long term benefit which

the cost incurred due to the development, implementation, and maintenance of such systems by a wide margin.

4. 0 The application of Data Mining

Nowadays, data mining is widely use especially to those organization that focuses on consumer orientation. For example, retail, financial, communication, and marketing organizations (Palace, 1996). Besides it, healthcare area also gain benefit by apply the data mining into the daily operations. These various of field shows each of the organization carries different transaction where all of details keep in the databases which enables to perform analysis for multiple purpose likes to increase revenue, gain more customer, improve customer satisfaction and others. Moreover, again through (Palace, 1996) the existence data allow to determine relationships among internal factor consists price, product positioning or staff skills and external factor consists economic indicators, competition and customer demographic.

Hence, there three examples of data mining’s application in different areas which are hotel sector, library scope and also hospital with the goals to reduce or eliminate the weakness by address it using the result that is interpret in well manner to assist in making decision for the best solutions. The examples are as follows:

· A data mining approach to developing the profiles of hotel customers.

A study conduct by Min, Min & Ahmed Emam, 2002 with the objective to target some of the valued customers for special treatment based on their

anticipated future profitability to the hotel. There are a few questions regarding to the customer profiling:

- Which customers are likely to return to the same hotel as repeat guests?
- Which customers are at greatest risk of defecting to other competing hotels?
- Which service attributes are more important to which customers?
- How to segment the customer population into profitable or unprofitable customers?
- Which segment of the customers' best fits the current service capacities of the hotels?

The researchers adopt decision trees for analyzing the data from the abroad method of data mining methodology because the ability to generate appropriate rules using visualization and simplicity. There are three steps having to follows in this process and it includes:

- Data collection: the process of select data that suit with objective from the previous survey. Moreover, remove the unwanted data from databases by filtering out the excel file.
- Data formatting: the process of converted all data in the spreadsheet to Statistical Packages for Social Sciences (SPSS) for the purpose of classification accuracy.
- Rules induction: the process of selection of algorithms to building decision trees which is C5. 0 to generate sets of rules that bring important clues in order for hotel manager to take further action.

As the result, the researcher found that “ if-then” rules as a useful in formulating a customer retention strategy with a predictive ranging from 80. 9 per cent to 93. 7 per cent whereas a predictive accuracy reflect to the rules conditions that affect by times (percentage).

· Using data mining technology to provide a recommendation service in the digital library.

A study conducted by Chen & Chen, 2006 with the purpose to provide recommendation system architecture to promote digital library service in electronic libraries. There are abroad of digital publication format likes audio, video, picture, etc. thus, it lead difficulties in analyzing or defining the keyword and content in order to gain information from the user to improve the service in the digital libraries.

In the methodology section, there are two data mining models selected which consist

o Ant Colony Clustering Algorithm;

This model is capable to find the shortest path or reduce time to find the best output fit with the problem that existence in the organizations. Each of the steps has different function to enable they too see the relation among the variables It takes a few steps which are:

Step 0: parameters and initialize pheromone trails.

Step 1: Each ant constructs its solution

Step 2: Calculate the scores of all solutions

Step 3: Update the pheromone trails.

Step 4: If the best solution has not been changed after some predefined iterations, terminate the algorithm; otherwise go to step 2.

o Association rules to discover the hidden pattern.

This model enables to find co-purchase items and assist in uncovered relationship algorithms in form of association rules. There are two main steps as follows:

Step 1: Find all large item sets

Step 2; use the large items set generated in the first step to generate all the effective association rules.

As the results, these two models encounter more than one solutions and enable to gain a lot of recommendation that can be manipulate into various problem that exists in conducting digital libraries as well as to promote the usage in multiple level of user using the appropriate mechanism and providing suitable services.

· Using KDD process to forecast the duration of surgery.

A study conducted by Combas, Meskens & Vandamme, 2007 with the aim is to identify classes of surgery likely to take different lengths of time according to the patient’s profile as well as to allow the use of the operating theatre to be better scheduled. There are many issues arise in this field that lead to the study. For example, an endoscopy unit use of endoscopy tube (shared resources) during the surgery. However their availability is limited because it

takes 30-45min to clean and sterilize each one. The scheduling of endoscopies (and all other operating theatre procedures) must obviously take into account the availability of these different resources.

The researchers adopt Knowledge Discovery in Databases (KDD) process to analyze this massive data from the databases. The step as follows:

- Step 1: data preparation which the selected data must be fulfill of requirement includes secondary diagnoses, “ Previous active history” and system affected.
- Step 2: data cleaning where filter data by concerning surgical procedures that had been performed at least 40 times (at least 20 times for combinations involving both surgery and specific surgeons).
- Step 3: data mining which to decide appropriate method to test on the portion of data which it involves rough set and neural network.
- Step 4: validation by comparison consist process of interpretation by comparing the result from two methods that perform data analysis in order to observe the rate of good classification.
- Then, researcher added up another three steps in order to fit with the objective that is proposed and to produce the best outcomes to forecast the durations of surgery. It consists of:
- o Step 5: Measuring the impact of predicting the duration of surgery on planning which in this step the duration of surgery supplied by the prediction models (empirical laws, rule-based laws, etc.) based on information stored in the database is used to feed a series of algorithms and heuristics for planning purposes

- o Step 6: Simulation involves the present time will allow to simulate the activity of the different theatre suites in terms of the operating sequence determined by planning methods on the two scenarios which are operating data and patient’s profile
- o Step 7: validation & selection of the best model where the results supplied by the simulation model should enable to assess the quality of scheduling on the basis of a series of performance indicators likes the length of time for which the operating theatres are not in use, the number of potential additional hours, and errors in predicting the duration of surgery.
- As the results, researchers are not particularly satisfactory. The main problem seems to be the choice of variable grouping, which might possibly have an effect on prediction quality.

5. 0 Conclusion

As a conclusion, data mining can be consider as an effective and efficient way to discover or to transform the invisible to visible data that retrieve from databases which have capabilities to store huge amount of data by using the right tools in assist or enable to analyze, synthesis and manipulate the content of data for various purposes and often depend on the main businesses that carries out to define the target.

From the discussion above, it can be seen that there are a lot of advantages when perform data mining especially in the business area which allow the organization to predict the trends, customer requirement, the relationship and so forth as early preparation can be identify in order to seek another or a few others way to ensure that organization can still operate their daily

operation after determine that organization not agree towards the result have been gain.

In order to produce the end result that satisfying the organization and minimize the error as it successfully implement the information in order to perform business transaction. The key variables should be assign in well manner meet or suitable with the objective that propose in conducting the study because it have to repeat the procedures when found the errors as the decision making process could not been done according to the timeline.

6. 0 References

Chen, Chia-Chen & Chen, An-Pin. (2006). Using data mining technology to provide a recommendation service in the digital library. The Electronic Library. 25(6): 711-734.

Combas, C., Meskens, N & Vandamme, J. P. (2007). Using a KDD process to forecast the duration of surgery. International Journal of Production Economics. 112: 279-293.

Forcht., Karen A. & Cochran, Kevin. (1999). Using data mining and datawarehousing techniques. Industrial Management & Data Systems. 99(5), 189-196.

Gargano., Michael L. & Raggad, Bel G. (1999). Data mining – a powerful information creating tool. OCLC Systems & Services. 15(2), 81-90.

Goebel, Michael & Gruenwald, Le. (1999). A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explorations Newsletter. 1: 20 – 33.

Imberman, Susan P. (2001) Effective Use of the KDD Process and Data Mining for Computer Performance Professionals. in International Computer Measurement Group Conference. Anaheim: USA, 611-620.

Ma, Catherine, Chou, David C. &. Yen, David C. (2000). Data warehousing, technology assessment and management. Industrial Management & Data Systems. 100(3), 125-135.

Min, Hokey., Min, Hyesung & Ahmed Emam. (2002). A data mining approach to developing the profiles of hotel customers. International Journal of Contemporary Hospitality Management. 14(6): 274-285.

Palace, Bill. (1996, Spring). Data Mining: What is Data Mining? retrieved March 2, 2010, from: http://www. anderson. ucla. edu/faculty/jason. frand/teacher/technologies/palace/datamining. htm