Functions, Processes, Stages And Application Of Data Mining

Azhar Susanto; Meiryani

Abstract: Data mining is an automatic information discovery process by identifying patterns from large data sets or databases. The process of finding information can be done by a method of grouping data into several groups from a data set that is in data mining called the clustering method. Data mining is defined as a method of finding information that is hidden in a database that is large and difficult to obtain by using only ordinary queries. Unit Implementation and Testing, software design is a series of programs or program units. Then unit testing involves verifying that each program unit meets its specifications (Sommerville, 2003). Programs should be released after they are developed, tested to correct errors found in their quality guarantee testing (Padmini, 2005). There are two testing methods, namely: (1) The white box method is a test that focuses on the internal logic of the software (program source code); (2) The black box method is to provide actual results in accordance with the results needed. In the testing phase, the author conducted a black box method that tests the functionality of the software alone without having to know the internal structure of the program (source code).

Index Terms: Data Mining, Data, Mining, Information Discovery, Databases.

1 INTRODUCTION

Data Mining is an activity that includes collecting, using historical data to find order, patterns and relationships in large data sets. The use of data mining is to specify patterns that must be found in data mining tasks. Data mining presence is motivated by data explosion problems that have been experienced lately where many organizations have collected data for many years (purchase data, sales data, customer data, transaction data, etc.) Data mining capabilities to find valuable business information from a very large database, can be analogous to mining precious metals from their source land, this technology is used to: Prediction of trends and business characteristics, where data mining automates the process of finding predictive information in large databases. The discovery of previously unknown patterns, where data mining sweeps the database, then identifies previously hidden patterns in one sweep. Data mining is useful for making critical decisions, especially in strategy. Data Mining is often translated as Data Excavation, which is actually not right because the word Mining should be translated into Mining and not Excavation. In context, of course there are significant differences between mining activities compared to excavation. Excavation is an activity carried out to move a number of materials from one place to another, as a result of the amount of material being moved which of course will be the same as the amount of material obtained. On the other hand, mining is an activity that is far more than just moving material. In the mining process a person will often only get a small piece of material from a large excavation, but a small piece of material has a much higher value than the material excavated. In addition, the mining process must also be preceded by studies, surveys, preparations and so on. Based on the description above, the Data Mining should be translated into Data Mining and not Data Extraction. In data mining activities, the "mountain" that will be mined is data that has been collected previously.

- *Azhar Susanto; Accounting Department, Faculty of Economics and Business, Padjadjaran University, Bandung, Indonesia*
- *Meiryani; Accounting Department, Faculty of Economics and Communication, Bina Nusantara University, Jakarta, Indonesia 11480 meiryani@binus.edu*

The goal to be achieved from this data mining activity is to obtain a number of information or knowledge that is of high value and can be utilized for the benefit of the community and the organization. In essence, the main purpose of Data Mining is to be able to find repetitive and valuable patterns that are often hidden in data stacks. For example, an activity that can make someone understand after reading a telephone book that the majority of people named Andi live in South Jakarta can be categorized as a data mining process. Whereas, finding where Andi Suhendar lives by searching his name in the telephone book is not a data mining process but can only be categorized as an ordinary query process. Data Mining is an activity and is not an algorithm or program. In the implementation of Data Mining activities, various techniques or algorithms are used which are based on various scientific disciplines such as statistics, artificial intelligence or machine learning. In general, the purpose of data mining can be grouped into 2, namely to be able to understand more about the behavior of the observed data, or often referred to as Descriptions, and to be able to estimate conditions that will occur in the future or called Prediction. With the ability to be able to recognize the existence of patterns that are related to behavior, connectivity, movement of data, it is expected that data mining can help humans understand more about the observed system and then anticipate the possibility of future system movements.

2 LITERATURE REVIEW

2.1 The Concept of Data Mining

Data Mining is a series of processes to explore added value in the form of information that has not been known manually from a database by extracting patterns from data with the aim of manipulating data into more valuable information obtained by extracting and recognizing important patterns or draw from the data contained in the database. Data Mining is a process that uses statistical techniques, mathematics, artificial intelligence, machine learning to extract and identify useful information and related knowledge from various large databases (Turban et al. 2005). There are several other terms that have the same meaning as data mining, namely Knowledge discovery in databases (KDD), knowledge extraction, data analysis/data/pattern analysis, business intelligence and archaeological and data data dredging (Larose, 2005). some

data mining definitions from several sources (Larose, 2005): Data mining is the process of finding something meaningful from a new correlation, existing patterns and trends by sorting through large data stored in the repository, using pattern recognition technology and mathematical and statistical techniques. Data mining is the analysis of observational databases to find unexpected relationships and to summarize data with new methods or methods that are understandable and useful to data owners. Data mining is an interdisciplinary field of science that brings together learning techniques from machines (machine learning), pattern recognition, statistics, databases, and visualization to overcome the problem of extracting information from large databases. Data mining is defined as a process of extracting useful and potential information from a set of data contained implicitly in a database. Data mining is also known by other names such as: Knowledge discovery (mining) in databases (KDD), extraction of knowledge (knowledge extraction) Analysis of data / patterns and business intelligence (business intelligence) and is an important tool for manipulating data for presenting information as needed user with the aim to assist in analyzing the collection of behavioral observations, in general the definition of data-mining can be interpreted as follows The process of finding interesting patterns from stored data in large numbers. Extraction of useful or interesting information (non-trivial, implicit, as far as the potential usefulness is known), pattern or knowledge of the data stored in large amounts. Exploration of analysis automatically or semi-automatically on large amounts of data to find meaningful patterns and rules. Data mining really needs to be done especially in managing very large data to facilitate recording activities of a transaction and for the process of data warehousing in order to provide accurate information for its users.

2.2 Data Mining Stages

The main reason why data mining is very interesting to the information industry in recent years is due to the availability of large amounts of data and the increasing need to convert the data into useful information and knowledge because it is in accordance with the focus of this field of extracting or mining knowledge from data of large size/amount, this information will be very useful for development. Following the steps: Data cleaning (to eliminate inconsistent data noise) Data integration (where broken data sources can be put together) Data selection (where data relevant to the analysis task is returned to the database). Data transformation (where data changes or unites into the right form to mine with a summary of performance or aggression operations). Knowledge Discovery (an essential process in which intelligent methods are used to extract data patterns) Pattern evolution (to identify truly interesting patterns that represent knowledge based on some interesting actions) Knowledge presentation (where an overview of visualization techniques and knowledge is used to provide knowledge that has been mined to the user).

2.3 Data Mining Functions

Data mining has important functions to help get useful information and increase knowledge for users. Basically, data mining has four basic functions, namely: Prediction function. The process of finding patterns from data using several variables to predict other variables of unknown type or value. Function Description (description). The process of finding an important characteristic of data in a database. Classification function. Classification is a process to find a model or function to describe the class or concept of a data. The process used to describe important data and can predict data trends in the future. Association functions (association). This process is used to find a relationship that is contained in the attribute value of a data set.

2.4 Data Mining Process

The processes commonly carried out by data mining include: description, prediction, estimation, classification, clustering and association. In detail the data mining process is explained as follows (Larose, 2005):

a. Description

Description aims to identify patterns that appear repeatedly on a data and change the pattern into rules and criteria that can be easily understood by experts in the application domain. The rules produced must be easy to understand in order to effectively increase the level of knowledge in the system. Descriptive tasks are data mining tasks that are often needed in postprocessing techniques to validate and explain the results of the data mining process. Postprocessing is a process used to ensure only valid and useful results that can be used by interested parties.

b. Prediction

Predictions are similar to classifications, but data are classified based on behavior or values predicted in the future. Examples of predictive tasks, for example, are to predict a reduction in the number of customers in the near future and stock price predictions in the next three months.

c. Estimation

Estimates are almost the same as predictions, except the target variable is estimated more in the numerical direction than in the direction of the category. The model is built using a complete record that provides the value of the target variable as a predictive value. Furthermore, in the next review the estimated value of the target variable is based on the value of the predictive variable. For example, an estimate of systolic blood pressure in hospital patients is based on patient age, gender, weight, and blood sodium level. The relationship between systolic blood pressure and predictive variable values in the learning process will produce an estimation model.

d. Classification

Classification is the process of finding a model or function that describes and distinguishes data into classes. Classification involves the process of examining the characteristics of an object and inserting an object into one of the classes that has been previously defined.

e. Clustering

Clustering is the grouping of data without being based on a particular data class into the same object class. A cluster is a collection of records that have similarities with each other and have an incompatibility with records in other clusters. The aim is to produce groupings of objects that are similar to each other in groups. The greater the similarity of objects in a cluster and the greater the difference in each cluster, the better the quality of cluster analysis.

f. Association
The task of associations in data mining is to find attributes that appear at a time. In the business world it is more commonly called shopping basket analysis (market basket analisys). The task of the association seeks to uncover rules for measuring the relationship between two or more attributes.

3 RESULT AND DISCUSSION

As a series of processes, data mining can be divided into several stages illustrated in Figure. These stages are interactive, users are directly involved or through the knowledge base (Han, 2006). There are 6 stages of data mining, namely:

1. Data cleaning
Data cleaning is a process of removing noise and inconsistent data or irrelevant data. In general, the data obtained, both from a company's database and experimental results, has imperfect contents such as lost data, invalid data or also just typos. In addition, there are also data attributes that are not relevant to the data mining hypothesis that they have. The irrelevant data is also better discarded. Data cleaning will also affect the performance of data mining techniques because the data handled will decrease in number and complexity.

2. Data integration (data integration)
Data integration is the merging of data from various databases into one new database. Not infrequently the data needed for data mining not only comes from one database but also comes from several databases or text files. Data integration is performed on the aribut attribute which identifies unique entities such as name attributes, product types, customer numbers and others. Data integration needs to be done carefully because errors in data integration can produce deviant results and even misleading the action later. For example, if data integration based on product type turns out to combine products from different categories, then there will be a correlation between products that actually do not exist.

3. Data Selection (Data Selection)
The data that is in the database is often not all used, therefore only the data that is suitable for analysis will be taken from the database. For example, a case that examines the tendency factor for people to buy in the case of market basket analysis, there is no need to take the customer's name, just with the customer's id.

4. Data Transformation (Data Transformation)
Data is converted or combined into a format suitable for processing in data mining. Some data mining methods require special data formats before they can be applied. For example, some standard methods such as association analysis and clustering can only accept categorical data input. Therefore the data in the form of a continuous numeric number needs to be divided into several intervals. This process is often called data transformation.

5. Mining process,
It is a major process when methods are applied to find valuable and hidden knowledge from data.

6. Evaluate patterns (pattern evaluation),
To identify interesting patterns into knowledge based found. In this stage the results of data mining techniques in the form of distinctive patterns as well as predictive models are evaluated to assess whether the existing hypotheses are indeed achieved. If it turns out that the results obtained are not in accordance with the hypothesis there are several alternatives that can be taken such as making feedback to improve the data mining process, trying other data mining methods that are more appropriate, or accepting these results as an unexpected result that might be useful.

7. Knowledge presentation,
Is a visualization and presentation of knowledge about the methods used to obtain knowledge obtained by users. The last stage of the data mining process is how to formulate decisions or actions from the results of the analysis obtained. There are times when this must involve people who don't understand data mining. Therefore the presentation of the results of data mining in the form of knowledge that can be understood by everyone is one step needed in the data mining process. In this presentation, visualization can also help communicate the results of data mining (Han, 2006) The stages carried out in the data mining process starts from the selection of data from the source data to the target data, the preprocessing stage to improve data quality, transformation, data mining and the stages of interpretation and evaluation that produce output in the form of new knowledge that is expected to contribute better. The details are explained as follows (Fayyad, 1996) :

1. Data selection
Selection of data from a set of operational data needs to be carried out before the information excavation stage in KDD starts. The selection data used for the data mining process is stored in a file, separate from the operational database.

2. Pre-processing / cleaning
Before the data mining process can be implemented, it is necessary to do a cleaning process on the data that is the focus of KDD. The cleaning process includes, among other things, removing duplicate data, checking inconsistent data, and correcting errors in data.

3. Transformation
Coding is a transformation process on selected data, so that the data is suitable for the data mining process. The coding process in KDD is a creative process and is very dependent on the type or pattern of information to be searched in the database.

4. Data mining
Data mining is the process of finding patterns or interesting information in selected data using a particular technique or method. Techniques, methods, or algorithms in data mining vary greatly. The choice of the right method or algorithm depends on the overall purpose and process of the KDD.

5. Interpretation / evaluation
The pattern of information generated from the data mining process needs to be displayed in a form that is easily understood by interested parties. This stage is part of the KDD process called interpretation. This stage includes examining whether the pattern or information found is contrary to the

facts or hypotheses that existed before. the process carried out by data mining? according to larose there are several processes carried out by data mining, namely description (identifying hidden hidden patterns and changing patterns into rules that can be understood by experts), predictions (classifying based on expected behavior to come), estimation (such as predictions except for estimation variables are more numerical), classification (the process of finding a function model and describing data into classes), clustering (grouping data based on a particular class on the object), association (finding attributes that appear in time). How is data mining implemented? quite a lot, especially in the fields of business, sports, computational science, health, and education. can explain the application of data mining in education? Of course. many applications of data mining in the field of education such as predicting student achievement based on lecturers, motivation, discipline, economics and learning outcomes. for data mining using Cross Industry Standard Process for Data Mining (CRISP-DM) method which consists of 6 stages:

- understanding of business (digging raw data with the help of the C4.5 algorithm)
- understanding of data (data in the form of questionnaires filled in the form of motivation, discipline, economy, and past learning outcomes and data will be processed
- data preparation
- modeling (training data processing stage)
- evaluation (testing of the C4.5 algorithm)
- spread

4 Conclusion

With extensive data mining definitions, there are many types of analytical methods that can be classified in data mining.

Association rules

Association rules (association rules) are related to the study of "what is together". An example can be a study of transactions in a supermarket, for example someone who buys baby milk also buys bath soap. In this case it means baby milk along with bath soap. Because it originally came from the study of customer transaction databases to determine the habits of a product purchased with what products, then the association rules are also often called market basket analysis. The association rules want to provide this information in the form of "if-then" or "if-then" relationships. This rule is calculated from probabilistic data (Santoso, 2007). Association analysis is also known as one of the data mining methods that forms the basis of various other data mining methods. Particularly one of the stages of association analysis called high frequency pattern analysis attracts the attention of many researchers to produce efficient algorithms. The importance of an associative rule can be known by two parameters, support (support value), which is the percentage of combination items. in the database and confidence (certainty value), namely the strong relationship between items in associative rules. Association analysis is defined as a process for finding all associative rules that meet the minimum requirements for support (minimum support) and minimum requirements for confidence (minimum confidence) (Pramudiono, 2007). There are several algorithms that have been developed regarding association rules, but there is one classic algorithm that is often used, namely the priori algorithm. The basic idea of this algorithm is to develop frequent itemset. By using one item and recursively developing frequent itemset with two items, three items and so on to frequent itemset with all sizes. To develop frequent sets with two items, can use frequent set items. The reason is that if the set of one item does not exceed the minimum support, then any larger itemet size will not exceed that minimum support. In general, developing sets with fc-items uses frequent sets with k - 1 items developed in the previous step. Each step requires a single check of the entire database. In associations there are terms antecedent and consequent, antecedent to represent the "if" and consequent to represent the "then" part. In this analysis, antecedent and consequent are a group of items that do not have a relationship together (Santoso, 2007). From the large number of rules that might be developed, it is necessary to have rules that are quite strong, the level of dependence between items in antecedent and consequent. To measure the strength of the rules of this association, measures of support and confidence were used. Support is the ratio between the number of transactions that contain antecedent and consequent transactions. Confidence is the ratio between the number of transactions that include all items in antecedent and consequent with the number of transactions that include all items in antecedent. The first step in a priori algorithm is, the support of each item is calculated by scanning the database. After support from each item is obtained, items that have support greater than minimum support are selected as high frequency patterns with length 1 or often abbreviated as 1-itemset. The abbreviation k-itemset means a set consisting of k items. The second iteration produces 2 items which each set has two items. First, a 2-itemset candidate is made from a combination of all 1-itemet. Then for each. This 2-itemset candidate counts its support by scanning the database. Support means the number of transactions in the database containing both items in the 2-itemset candidate. After the support of all 2-itemet candidates is obtained, the 2-itemset candidate that meets the minimum support requirements can be determined as a 2-itemset which is also a high frequency pattern with length 2 (Pramudiono, 2007).

Acknowledgment

The authors wish to thank to Padjadjaran University, Bandung Indonesia and Binus University, Jakarta, Indonesia.

References

[1] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann., 1993.
[2] MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.
[3] Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag.
[4] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.
[5] McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York.
[6] Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.
[7] Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.
[8] Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. TPAMI. 18(6).

[9] Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.
[10] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, 1984.
[11] Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.
[12] Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. SODA, 1998.
[13] Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.
[14] Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996.
[15] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
[16] Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
[17] Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992.
[18] Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.
[19] Turban, E, 2005, Decision Support Systems and Intelligent Systems Indonesian Edition Volume 1. Andi: Yogyakarta.
[20] Larose, Daniel T. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Willey & Sons, Inc.
[21] ayyad, Usama. 1996. Advances in Knowledge Discovery and Data Mining. MIT Press.
[22] https://beyonder.asia/pengertian-fungsi-proses-dan-tahapan-data-mining/