Modelling Data Mining Dynamic Code Attributes With Scheme Definition Technique

Evasaria M. Sipayung[1], Cut Fiarni[1], Randy Tanudjaja[1]
Department of Information System [1]
Institut Teknologi Harapan Bangsa (ITHB)
Bandung, Indonesia

***Abstract*—Data mining is a technique used in different disciplines to search for significant relationships among variables in large data sets. One of the important steps on data mining is data preparation. On these step, we need to transform complex data with more than one attributes into representative format for data mining algorithm. In this study, we concentrated on the designing a proposed system to fetch attributes from a complex data such as product ID. Then the proposed system will determine the basic price of each product based on hidden relationships among the attributes of data. These researches conclude that the proposed system accuracy of precision rate is 98.7% and recall rate are 70.27%.**

Keywords—data mining; attributes; recall; precision

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information [1]. Now days, most of the companies used information technology as part of its business strategy. Consequently, companies have a vast amounts of data, but unfortunately poor in information and knowledge extracted from that data. Big data is seen as a valuable resource and although the concept of data mining is still new and developing, companies in a variety of industries are relying on it for making strategic decisions [2]. The importance of collecting data to gain knowledge of business transaction is recognized and widely adopt by companies that understand its benefit on business competitive edge. As the result, a powerful system for collecting data and managing it in large databases are in place in all large and mid-range companies.

An immense amount of data transaction of sales of product can be used to gather market intelligence on consumers and communities. By mining data transaction, companies can analyze top product to gain patterns of its characters and a distribution purchasing pattern on each sales areas. By gaining these patterns and knowledge can increase sales and distribution of products.

However, one of the bottlenecks of extracting data into knowledge on data mining processes is the difficulty of data transformation. That involves data representation in an appropriate format for mining algorithm. To overcome the bottleneck on data transformation processes, the construction of a data warehouse, which involves data cleaning and data integration, can be viewed as an important pre-processing step for data mining. Especially with data that contain significant attribute to mining, such as identity number, ID of product etc.

We address this issue in this paper by presenting an integrated framework for knowledge discovery and management, in the context of marketing decisions. Our paper is further organized as follows. First, we will design a scheme function to fetch attributes from ID products as part of data transformation on data mining processes. Second, we extract the relationship of each attributed to gain knowledge of the basic pricing of each product. Then, we test capability of system by analyze the precision and recall rate of basic price of products. We close our discussion by identifying the emerging issues to be addressed in the process of managing the discovered marketing knowledge.

II. BACKGROUND AND RELATED WORK

Data Mining techniques are the result of a long process of research and product development [1]. Companies uses mining tools and techniques to find useful relationships, patterns and anomalies that can help managers make better business decisions. Data mining tools perform analyses that are very valuable for business strategies, scientific research and for companies is to getting to know company's customers better. Managerial insights are no longer the only factor trusted when it comes to decision-making. Data driven decisions can lead to better firm performance [2].

There are several assessment processes for applying Data Mining [3]:

a. Data preparation
b. Clustering analysis for data mining.
c. Results of Expression and Visualization
d. Pattern evaluation

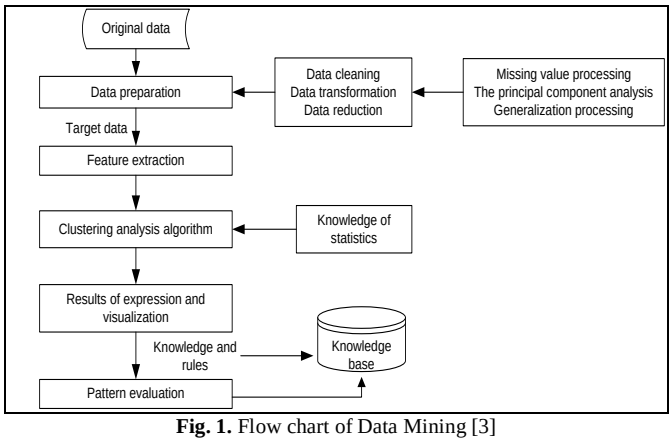

Fig. 1. Flow chart of Data Mining [3]

To manage data complexity, data need to be defined in order to show the relationship between attributes on the data. Dynamic scheme definition is a methodology to define a scheme on existing data after it has been collected and stored, and use the scheme to retrieve data at runtime while processing [4].

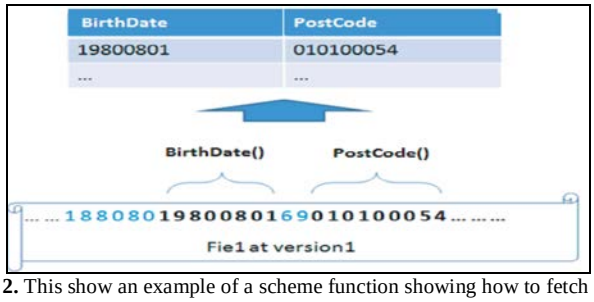

Fig. 2. This show an example of a scheme function showing how to fetch BirthDate and PostCode from identity number [4]

III. MODELLING ATTRIBUTE DATA WITH SCHEME DEFINITION

The following data that used in this research is taken from CV.XYZ, a toy company located in Bandung, Indonesia and has several distribution stores. The toy company produces more than 200 variations of toys, and it's sold in all Indonesian regions, in more than 100 stores.

One of the most crucial part on data mining processes is data transformation, because how to represent data into an appropriate format for mining algorithm, especially on data that contain more than one attributes to mining. With these types of data, hiding data relationship logic in a program is not a good way to manage data complexity. Because big data uses the 'structure later' approach, in most of the cases, we can only know the data scheme after the data has been created [4]. In this research, we will design a scheme function to fetch attributes from ID products as part of data transformation on data mining processes. There are 3 attributes for each toy ID, which are material type, toy type and accessories. There also hidden information which is product price that gain from the 3 attributes of toy ID. Then we will use these modeling on proposed business intelligent system, to gain knowledge on marketing strategies. The system will analyze pattern of top product based on amount of sold toy on data transaction. This information then being used together with the proposed scheme function to know the toy attributes of top product or any other cluster.

A. A Scheme Function to Fetch Attributes of Product ID

As explained of previous part, from ID product or in this case toy ID, a business analyst needs to gain pattern and knowledge on top product characteristic. To gain this knowledge, the proposed system will adopt a scheme function to fetch attributes of toy ID. The algorithm to fetch attribute of ID products for the proposed business intelligent system is designed as follows:

1. Determine for product or toy type. The System will fetch the second and third character from ID toy and matching the character with code of the product library on the system
2. If the product type is a ball, the system will analyze the ball size, by fetching the fourth and fifth character from ID toy. If the toy type is not a ball, system will skip fetching size and just fetch the type of toy.
3. Determine toy's material and thickness. To determine the materials of toys, there are different methods to analyze material code. This happens because the position of code in ID toy are different, it's depend on length characters, is as follows:
    a. If the type of product known as a ball:
        i. If the length characters are 11 or 12, material code position is on eight and nine, with the character length of the material code is 2
        ii. If the length characters are 9 or 10, material code position is at eight of the ID toys
    b. If the type of product known as a figure (toy);
        i. If the length characters is 12, material code position is on the seventh of the ID toys
        ii. If the length characters is 11, material code position is on the eighth of the ID toys, with the character length for the material code are 1 or 2

    After fetching the material code, the system will match the character that's been fetched with the material library that embedded on the system
4. Determine the type of accessories of the toy. Once again, to determine accessories of toys, there are different methods to analyze accessories code. This happens because the position of code on ID toy are different, it's depend on length characters, is as follows:
    a. If the type of product known as a ball:
        i. If the length characters is 12, material code position is at the third last characters of ID toy

ii. If the length characters is 11, material code position is at the very last characters of ID toy

b. If the type of product known as a figure (toy);

i. If the length characters is 12, material code position is iat the fourth last characters of ID toy

ii. If the length characters is 11 or 10, material code position is at the second last characters of ID toy

iii. If the length characters is 9, material code positions is at the second last or the very last characters of ID toy

iv. If the length characters is 8, material code position is at the second last characters of ID toy.

After fetching the accessory code, the system will match the character that's has been fetched with the accessories library that embedded on the system.

After all the code being fetch and analyze, the system now can determine type, material and accessories of the product. From this attribute, system will analyze the relationship of each attributes in order to gain hidden information, which is the basic price of the product.

.

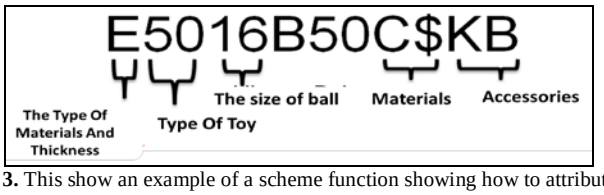

Fig. 3. This show an example of a scheme function showing how to attributes of product from toy ID.

B. *Determine Basic Price from Product ID*

Based on the analysis conducted on the product group, found that in the group type of toy products, the price is set based on the product code that has been determined by the management company based on characteristic of each toy. These characteristics are becoming attributes of ID product. Hence, to determine the basic price of each toy, system need to fetch each attribute combine it's with pricing rules from the management company. The logic of toy ID analysis to determine the basic price of each product is described in the following flowchart.

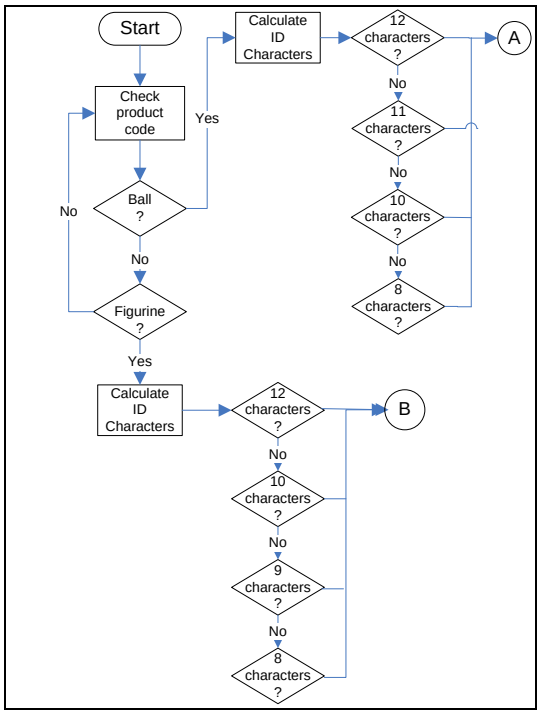

Fig. 4. This show flowchart diagram of determining product or toy type

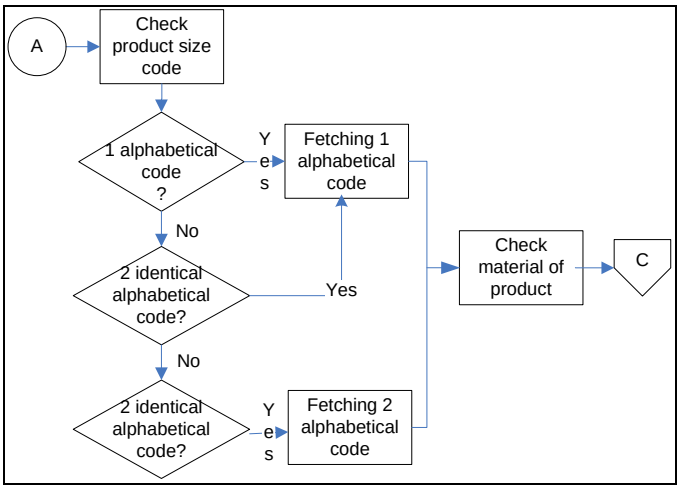

Fig. 4. This show flowchart diagram of determining toy's material and thickness

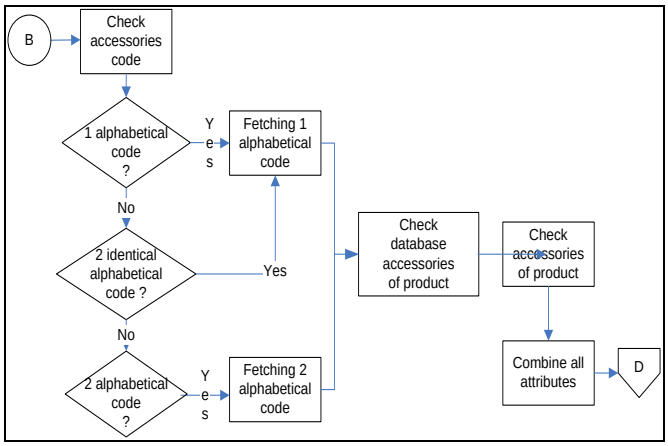

Fig. 4. This show flowchart diagram of determining type of accessories of the Toy

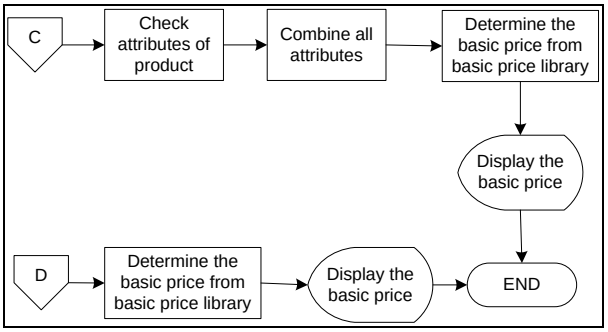

Fig. 4. This show flowchart diagram of determining basic price form toy ID

The system will split the character of toy ID based on the type of existing products. The following flow logic decoding to obtain the price that corresponds to the attributes of the product:

1. Determine the toy character based on the type of product, whether the product type ball or figure (toy).
2. If the figure type toy, set the price based on the product code that has been stored in the database.
3. If the ball-type toy, then specify the existing ball materials, as well as the attributes attached to the product

IV. EXPERIMENT RESULTS

To analyze the system capability to extract attributes form ID toy, we used expressions by a process of generalization from the labeled examples and so the set of attribute occurrences that they pull out will be a superset of the labeled examples. The amounts of testing data that used are 184 items. The results of the proposed system are:

$$\textit{Recall}\ [5, 7] = \frac{ce}{ce + te} \times 100\% \quad (1)$$

$$= \frac{130}{185} \text{ x } 100\% = \mathbf{70.27\%}$$

$$\textit{Precision}\ [6, 7] = \frac{ce}{ce + fe} \times 100\% \quad (2)$$

$$= \frac{182}{185} \text{ x } 100\% = \mathbf{98.37\%}$$

Where *ce* is the number of entities extracted correctly,
te is the number of true entities not extracted, and
fe is the number of false entities extracted

V. CONCLUSION

This study utilizes data pre-processing on complex data as part of data mining on business intelligent. The steps to define a scheme on existing data and fetching attributes of product and how to found the basic price of product as a hidden relationship between attributes were carried out and explained in detail. The level of accuracy of proposed system to analyze product attributes such as type, material and accessories with sensitivity test and specificity, give results precision rate is 98.37% and recall rate is 70.27%. The use of the scheme function technique to fetch attributes of Product ID may provide us with more varied and significant findings, and may lead to the increase in the quality of knowledge from data mining processes.

REFERENCES

[1] T.Imielinski and H. Mannila. Communications of ACM. A database perspective on knowledge discovery, 39:58-64, 1996.

[2] Gancheva, V. Market Basket Analysis of Beauty Products. Thesis on Erasmus University Rotterdam, 2013.

[3] Yanrong Guo, Baoguo Wu, Yang Liu. Multidimensional Data Mining using a K-mean Algorithm based on the Forest Management Inventory of Fujian Province, China. TELKOMNIKA. 11(12):7290-7294, 2013.

[4] Zhu, J.: Data Modeling for Big Data, CA, Beijing (2012).

[5] B. Adelberg. NoDoSe: A tool for semi-automatically extracting structured and semi-structured data from text documents. In ACM International Conference on Management of Data (SIGMOD), 1998.

[6] D. Angluin. On the complexity of minimum inference of regular sets. Information and Control, 39(3):337–350, 1978.

[7] Hoda Waguih. A Data Mining Approach for the Detection of Denial of Service Attack. IAES International Journal of Artificial Intelligence, 2(2):99-106, 2013.