Complex Data Analysis for Products Bundling

IOP Conf. Series: Materials Science and Engineering **407** (2018) 012102 doi:10.1088/1757-899X/407/1/012102

Complex Data Analysis for Products Bundling

A P Purfini

Department of Computerized Accounting, Universitas Komputer Indonesia,
Jl. Dipatiukur 112-116 Bandung, Indonesia

apriani.puti.purfini@email.unikom.ac.id

Abstract. The aims of this research is meant to analyze product bundling involving complex data composed of consumer reviews and selling transaction numerical data. One of the Association rule implementations is by product bundling analysis. Textual data extraction used sentiment analysis. Bundling analysis involving the corresponding complex data used Association Rule and K-Means Clustering method for data sorting used to form transaction data as input to the Association Rules. Result this research the products to be bundled are products that have qualified characteristics taking into account buyer reference and sales data. It is expected that the proposed method to find the bundle of products can increase the potential of sales.

1. Introduction

Complex data are data with different type and structure from different data sources. Transforming complex data into more structured and able to be explored is a difficult task and requires efficient techniques [1]. One of the challenges for data mining process is how to transform complex data into more structured data and how to find pattern matching from complex data. Product bundling is one of the business marketing strategies used by companies [2]. The emerging problem when companies apply product bundling is how to select bundles to maximize product compatibility in one package [3]. Bundling can also minimize consumer costs, depending on the number of items bundled, the value of those items, and the level of the variations [4, 5].

Previous researches to find the bundle of products focus involved data set with one data type and singular data [6]. Web-based collaborative filtering mechanism for firm's product bundling strategy used one data type [7]. Generative approach to find the bundle of products that best meets the preferences, user requirements and to ensure the satisfaction of the trader's needs such as minimization of dead stock [8]. Machines to find products that meet the requirements and, at the same time, seller needs such as minimization of dead profit and net profit maximization [3]. However, from previous research has not yet explained about finding products to be combined using various data sources.

Therefore, this research is meant to analyze product bundling by involving complex data composed of consumer review and selling transaction numerical data. Bundling analysis involving the corresponding complex data uses Association Rules and K-Means Clustering Method for data sorting used to form transaction data as input to the Association Rules. This method is more efficient than using the analysis using one set data type.

2. Methods

2.1. Sentiment analysis

The data used in this research are textual consumer review. Consumer review contains information on e-commerce post transaction consumer opinion in a form of opinion statements. Consumer review is an unstructured data which require further process in order to make them structured. Sentiment analysis is a method to extract textual data, transform unstructured data into structured data Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text [9]. Sentiment analysis phases composed of feature extraction as a basis of classification process, tokening process with words separation feature. Stop words removing phase contains process to remove unimportant words according to the dictionary in use in order to enhance accuracy. Keywords extractor is a process to classify words obtained from stop words phase into either positive or negative keywords. Keywords classification determination is based on human (user) intuition. The result of this process is a list of keywords for positive and negative category. The aim of this sentiment analysis as the last step of sentiment analysis process is to determine textual consumer review into positive and negative sentiment category.

2.2. Data integration

In this research, the data is composed by two; transaction data and consumer review. Therefore, an integration process to assimilate data is required. Data integration is done after each data has been pre-processed and other data preparation as seen in the figure 1. For consumer review data, after sentiment analysis and sentiment values are obtained for each consumer review.

The next process is decoding sentiment values with the following condition; if the sentiment value is positive then it is decoded as +1, conversely it is decoded as -1. Sentiment decoding process aim is to determine how big the positive or negative the consumer review on the products is. Pre-processing before data integration is meant to do ratio calculation for each transaction. (Figure 1)

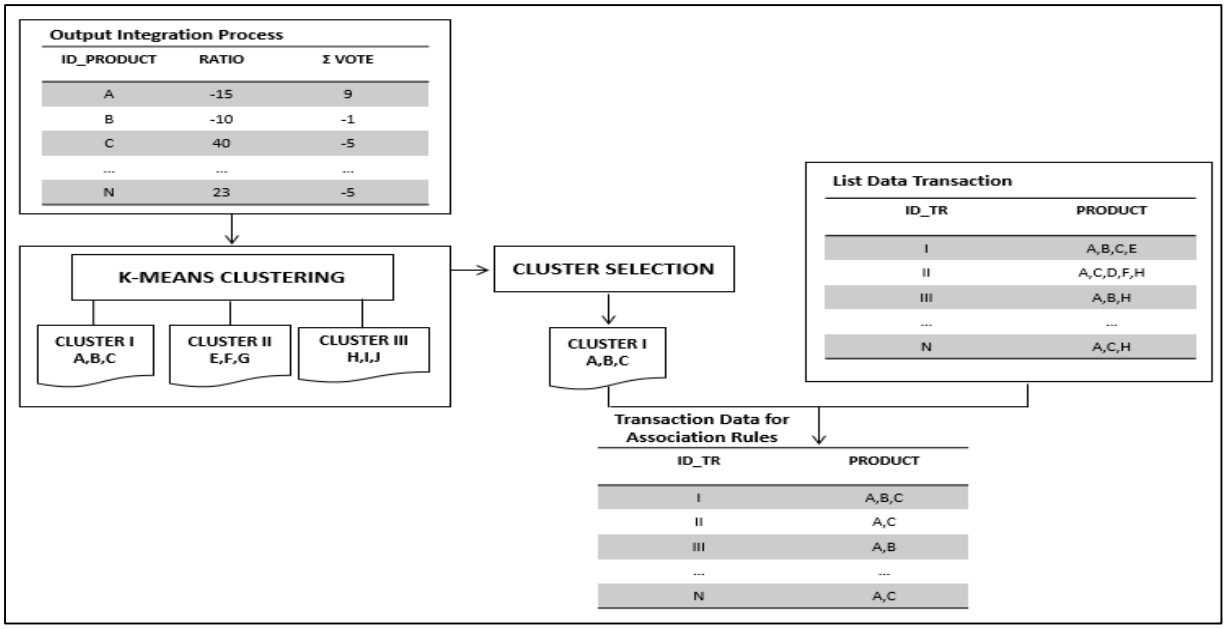

Figure 1. Data integration.

2.3. Create data transaction

Prior to the Association Rule, sorting the data is required. For data sorting, we propose clustering method using K-Means. The k-means clustering algorithm is the most commonly used [10] because of its simplicity [11].

The integration data result from process between selling transaction and sentiment data are later clustered. Clustering process uses K-Means algorithm in order to find list of products within the best quality products. Best quality products are those having high selling and good perspective to consumers. Clustering produces list of products within selected clusters i.e. clusters satisfying the best quality products criteria. Creating transaction data requires selling transaction data having ID_Transaction, and Data_Product attributes. Data_Product sorting process, data_product to be included in the list of data_product from clustering process. Creating transaction data is mean to produce Data_Transaction satisfies the best quality products which furthermore acts as item set for association rule.

3. Results and discussion

This part explains the implementation from data analysis concept for product bundling. We propose to use numerical transaction data and consumer review. The implementation covers experiment data. The data used in this research originated from e-commerce i.e. selling transaction data and consumer review. In this research, we propose real data, consumer review transaction data in xlsx format. Selling transaction data are taken from e-commerce transaction data within a year from 3rd of January 2014 until 31st of December 2014. There are 95,220 transactions within a year period for analysis. The obtained transaction data are grouped by 4 categories; fashion, electronics, hobby & children care. The attributes in this research are order_id, order_date, product_id, product_name, category_product, and qty.

The next one is consumer review. The consumer review data are obtained within the same period of transaction data. Apart from recorded transaction data for each transaction, to obtain consumer review data requires direct sampling from e-commerce web, consumer review data for each particular data transaction of certain product and later stored in xlsx format. The consumer review attributes are product_name and review_content. There are 3,910 consumer reviews, because not all products within the transaction data have reviews.

Pre-processing is meant to produce transaction data by ratio value calculation. The ratio value is used to observe the sale of each product. Transaction Data Pre-processing covers processes are Recording transaction data in xlsx format. The data transaction attributes are product_id, order_date, product_name, price and total product. Creating new attribute qty, This_Month and Next_Month. Derive qty acts as a new attribute creator which is qty as a result of calculation product (total product - price). This_Month derive node acts as This_Month creator by taking monthly from order_date attribute for each transaction. Next_Month derive node acts as creator for the Next_Month attribute. Aggregate Node calculates the summarize of qty for each product within one month period. The next process is to create Next_Month by changing This_Month attribute and Qty_Sum_Next by changing Qty_Sum. The next node calculates ratio value by using formula.

The sentiment analysis is used to extract consumer review data into a structured form. Stop words selection feature and keyword extraction by user intuition. The sentiment analysis process composed of Inserting consumer review, with a format. Attributes used in this process are Product_Name with Review_Content. Text Mining Node is used for tokenizing and documenting classification process. Tokening process is used to parse input string based on word composition. Stop words and keyword extraction process done by user intuition could not be used for data mining in Indonesian language. This node is also useful to classify consumer review text based on previous keywords extraction. Output Node table is meant to display consumer review text result.

Data integration process assimilates transaction data with consumer review. Before integrating data, each data has been pre-processed. The integration process can be explained are pre-processed transaction data as explained produces transaction data having ratio attribute. Pre-processed consumer review from

sentiment analysis produces structured textual data. Merging process between transaction data and sentiment analysis data produces data.

Prior to data mining using association rule, transaction data creation is required. It is meant to produce transaction data satisfying the best quality product criteria. Data creation process uses K-Means algorithm. Transaction data clustering process are explained: Integration data has set result as input to clustering process. Binning node is used with the purpose to reduce variation in data and to increase the performance of data mining algorithm. Binning process in this research implements 4 bins. Binning in this research is used for ratio attribute and Σvote. Clustering process uses K-Means algorithm to produce data within the best quality products category. The next process after obtaining product list based on clustering is by creating data matrix.

Association Rules is a data mining method to collect group of items frequently present simultaneously. The created data matrix from previous process later act as input for association rules. The association rules are explained each category matrix from previous process acts as input for Association Rule. Aggregate Node is meant to do summarizing for each product. Apriority Node Modelling is used for association rule, and the result will be used by companies for product bundling decision.

The clustering process is done to group data based on sentiment analysis result and product sale. Better clustering process will group similar data in one group, and different data will be grouped into several different groups. The evaluation from clustering process is done by using silhouette coefficient as a cohesion measurement and separation from created cluster. The approximated silhouette value is 1 and this value indicates good clustering:

- Cluster 1: 73,8 % product in this cluster; consists of the data with the level of sales rise and the positive sentiment; its means the product be categorized of superior products.
- Cluster 2: 26.1% product in this cluster; consists of the data with the level of sales rise and the negative sentiment.
- Cluster 3: 0.1% product in this cluster; consists of the data with the level of sales drop and the positive sentiment.

The number of produced rules give opportunities to see patterns frequently present in the database. So they are used to provide probabilities as a basis of decision-making. Not all discovered rules in this research are interpreted. The interpreted rules are those having high confidence value (objective reason) and having high relevance with the requirement (subjective reason). The rule results satisfy minimum support are processed to obtain association rule. The result of association rule can be used to determine association pattern among items.

The results of the research different from other researches, other researches to find the bundle of products focus involved data set with one data type and singular data[3,6,7,8], while my research uses data that varies with different data types to find the bundle of products. The products to be bundled are products that have qualified characteristics taking into account buyer reference and sales data.

4. Conclusion

This research is intended to apply association rules for complex data. Pre-processing of complex data can be done by text mining to define sentiment and ratio calculation to define sales tendency. To obtain high quality bundling product, association rule and clustering. The clustering algorithm to group products as input for association rule with positive sentiment and tendency of sale to rise.

Acknowledgement

This research was supported by Universitas Komputer Indonesia, Indonesia.

References

[1] Boussaid O, Bentayeb F and Darmont J 2003 “A MultiAgent,” *10th ISPE International Conference on Concurrent* (Portugal).

[2] Stefan Stremerch G J 2002 "Strategic Bundling of products and Prices: A New Synthesis for Marketing," *Journal of marketing* **66** pp 55-72.

[3] Cosimo Birtolo D D C L R V 2013 "Searching optimal product bundles by means of GA-based Engine and Market Basket Analysis," *FSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS) 2013 Joint* **12**(23) pp 448-453.

[4] Arora R 2008 "Price bundling and framing strategies for complementary products," *Journal of Product & Brand Management* **17**(7) pp 475-484.

[5] Estelami H 1999 "Consumer savings in complementary product bundles," *Journal of Marketing Theory and Practice* **3** (7) pp 74-85.

[6] Sakurai S 2004 "Analysis of Daily Business Reports based on Sequential Text Mining Method," *Proc. of the 2004 IEEE Intl. Conf. on Systems, Man and Cybernetics* **4** pp 3279-3284.

[7] Guo-rong L a X-z Z 2006 "Collaborative filtering based recommendation system for product bundling," *Management Science and Engineering* **8** pp 251-254.

[8] Cosimo Birtolo D D C M A R A 2011 "A generative approach to product bundling in the e-Commerce domain," *Nature and Biologically Inspired Computing (NaBIC) 2011 Third World Congress on* **12** pp 169-175.

[9] Liu B 2010 *Sentiment Analysis and Subjectivity* (Chicago: N. Indurkhya and F. J. Damerau).

[10] Jain R C D a A K 1998 *Algorithms for Clustering Data* (Prentice Hall).

[11] Ng A 2012 "Clustering with the K-Means Algorithm," *Machine Learning.*