KINETIK, Vol. 3, No. 4, November 2018, Pp. 287-294
ISSN : 2503-2259
E-ISSN : 2503-2267

Purchase Recommendation and Product Inventory Management using Content Based Filtering with Sequential Pattern Mining Approach

Aditya Cipta Raharja[*1], Imas Sukaesih Sitanggang[2], Agus Buono[3]
[1,2,3]Institut Pertanian Bogor
adityacr@outlook.com[*1], imas.sitanggang@ipb.ac.id[2], pudesha@gmail.com[3]

Abstract

Today, the product sales at XYZ Bookstore are increase in accordance to the trend in society. In that case, high sales must be supported by good supply and on target. Product sold based on needs of consumers will make possibility to achieve high sales. Using the Sequential Pattern Mining approach, we can specify sales patterns of products in relation to another products. SPADE (Sequential Pattern Discovery using Equivalence classes) is an algorithm that can be used to find sequential patterns in a large database. This algorithm finds frequent sequences of the sales transaction data using database vertical and join process of the sequence. The results of SPADE algorithm is frequent sequences which are used to form the rules. Those can be used as predictors of other items that will be purchased by consumers in the future. The result of this study is a lot of unique sequence appears that can provide the best advice for Merchandiser Officer, for example, there are 1.468 sequences that prove the customer who bought the product in Children's Book category will always bought the same thing in the others day. This research produce some recommendation, one of the recommendation is Children's Book category has a very high chance of being a Best Seller for a long time so that the purchasing officer on XYZ bookstore should ensure that the product's supply of the category is always safe throughout the year. *It means SPADE is successfully used to provide the advice and Merchandiser Officer must ensure the stock of that product is always available to avoid Lost Sales.*

Keywords: *Data Mining, SPADE, Recommendation Model, Sequential Pattern Mining*

1. Introduction

At this time many companies have realized how important to use a strong and good database to face the business challenges. The database must be relied and integrated with all parts in those companies. According to [1] the element with a strong impact in business stability is the cost given by business needs: human needs, technical needs, current expenses, miscellaneous costs, etc. The purpose of reduce costs for sub-assembliesis we can reduce the total cost.

Important data to support product analysis is customer purchase pattern [2]. The customer purchase pattern is very unique because each individual customer who shops in XYZ bookstore will certainly affect the purchase pattern of products in the company. In retail companies that rely on sales revenue, customer purchase patterns are very important to know the relationship between one product with another product for the customer. This is important because the right product will encourage customers to shop in the store again and again [3]. If this pattern has been established and known, it will be very helpful, especially for merchandising section so they can keep the stock of products sold in order to remain available and not empty. It means the company can avoid lost sale opportunity.

XYZ bookstore is one of the leading bookstores in Indonesia. One way for XYZ Bookstore company to compete with other retail companies so they can get more profit is to do the analysis to determine the pattern of product purchase sequence. This is expected to help the inventory of goods in the bookstore to get the target. Sequential PAttern Discovery using Equivalent classes (SPADE) is one of the sequencing pattern mining algorithms that can be used to overcome previous algorithm deficiencies, where a complete database search must be performed multiple times. SPADE uses a vertical id-list to facilitate search in the database. SPADE can search for frequent sequences with multiple database searches only [4]. SPADE produces good performance in terms of computational time compared to previous frequent sequence search

Raharja, A., Sitanggang, I., & Buono, A. (2018). Purchase Recommendation and Product Inventory Management using Content Based Filtering with Sequential Pattern Mining Approach. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 3(4). doi:http://dx.doi.org/10.22219/kinetik.v3i4.663
Receive June 29, 2018; Revise July 19, 2018; Accepted August 03, 2018

algorithms such as Apriori and Generalized Sequential Pattern (GSP). So hopefully the SPADE algorithm will be suitable to applied in thedata sale transaction in case to face big data amount of sale transaction [5].

Product inventory management is inventory management that involves a series of decisions aimed at matching the existing time interval between demand and the supply of products and materials. Other objectives are in order to achieve cost and service level with a specific purpose, observe the characteristics of products, operations, and demand [6]

[7] have previously successfully applied the sequential pattern mining method using the SPADE algorithm to predict the purchase of computer spare parts and returning customer arrivals at the analyzed stores. Based on experimental results, the SPADE algorithm is accurate and applicable to predicte consumer arrivals. Using SPADE coming consumer arrival can be predicted with the accuracy of 75%.

Recommendations are expected to help users in the decision making process, such as what items to buy, what books to read, or what music will be heard, and others [8]. This study uses Content-based filtering because the data who available in this study is data of product purchased along with date of purchase without supported by demographic data from customers who bought the product.

The purpose of this study is to find customer buying pattern with sequential pattern mining approach, and to formulate purchase recommendation on targeted product inventory management. The benefits that can be obtained from this study is the company can know the purchasing pattern for book product that became the main product of the XYZ bookstore. If those patterns are known, then it can be a recommendation material for the purchasing officer (merchandiser) to buy the product in accordance with current market trends. It is expected to increase product sales if the purchasing of goods have been right on target according to the recommendations formed.

2. Research Method

The study is done through several stages: data preprocessing process to produce a ready to process database, sequential sequence determination stage that look for sequential patterns from existing database and sequential pattern analysis step to determine appropriate purchase recommendation according to the purpose of this study. The detail of the stages of this study can be seen in Figure 1.

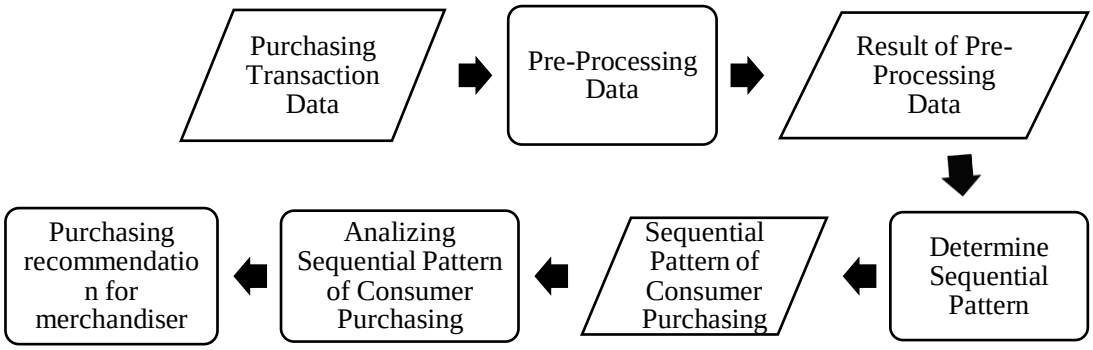

Figure1. Research Steps

The way to solve this sequential problem can be done with several methods such as Generalizes Sequential Pattern (GSP), SPADE, FreeSpan, and PrefixSpan [9]. For example the sequential pattern mining process is as follows: there is a sales transaction table containing the customer ID, date and item. From the transaction table then they are formed a sequence of transactions based on the customer and sorted by date so as to form several sequences. In other words, if a sequence is given, each sequence consists of a series of elements, and each element consists of a number of items. Mining process of the sequential pattern is to search for all subsequent recurrences, the subsequent frequency whose occurrence is greater than the minimum support [10].

3. Results and Discussion

3.1 Data Pre-Processing

The dataset used in this study is obtained from sales data in one branch of XYZ Bookstore located at East Jakarta area and in the period of June 2016 until December 2016. The first stage in data Pre-Processing is data selection. This stage is done in the database management system that used by XYZ Bookstore. Purchase data contain in the database that has many attributes. However, not all attributes are required in this study. There are three attributes used namely Customer ID, Transaction Date, and Product Category. The selection of attributes was done because this study aims to produce sequential patterns of customer purchases within a certain time to create sequential dataset we need only those three attributes. The description of the selected attributes is shown in Table 1.

Table 1. The Attributes Used in the Dataset

No	Attributes Name	Information
1.	Customer ID	Unique number of each member in XYZ bookstores
2.	Transaction Date	The date when the customer made a purchase
3.	Product Category	Categorization of products purchased by customers

The dataset used in this study consists of 49,717 transactions by 7,352 member in sales period June 1, 2016 until December 31, 2016. There are 17,015 items contained in the transaction data. The items are divided into 27 categories as can be seen in Table 2.

Table 2. List of Product Categories

No	Code	Category	No	Code	Category
1.	1002	Fiction & Literature	15.	2004	Computing & Internet
2.	2019	Schoolbooks Indonesia Curriculum	16.	2021	Science & Nature
3.	2003	Business & Economics	17.	2002	Art. Architecture & Photography
4.	2018	Religion & Spirituality	18.	2005	Nn (*uncategorized*)
5.	1001	Children`sBooks	19.	2001	Agriculture
6.	2006	Diet & Health	20.	2010	Home & Garden
7.	2017	Reference & Dictionary	21.	2011	Law
8.	2009	Entertainment	22.	2013	Medical
9.	2022	Self-Improvement	23.	2012	Magazines, Tabloid & Journal
10.	2016	Psychology	24.	2008	Engineering
11.	2023	Social Sciences	25.	2015	Philosophy
12.	2099	Others	26.	2024	Sports & Adventure
13.	2007	Education & Teaching	27.	2020	Schoolbooks Singapore Curriculum
14.	2014	Parenting & Family			

In the data transformation stage, the purchase date format is converted to unicode form so that data can be processed into software R. After that, data synchronization was done in advance so that data can be in accordance with the format requested by SPADE package in software R. The percentage of the distribution of items and categories can be seen in Figure 2.

3.2 Result of Pre-Processing Data

The data that have been through the preprocessing stage is ready to be reprocessed in the R application. In the Figure 3, the first column is called the sequence ID, which is the sequence where the event is located. In this study, the column is filled by Customer ID. The second column is called the timestamp event when exactly the event occurs. The column is filled by time (in unicode format) when the purchase transaction occurs. The third column is the number of items in the event. The third column is filled by the number of products purchased by the customer. The next column consists of what items the customer purchases, according to the amount in the third column.

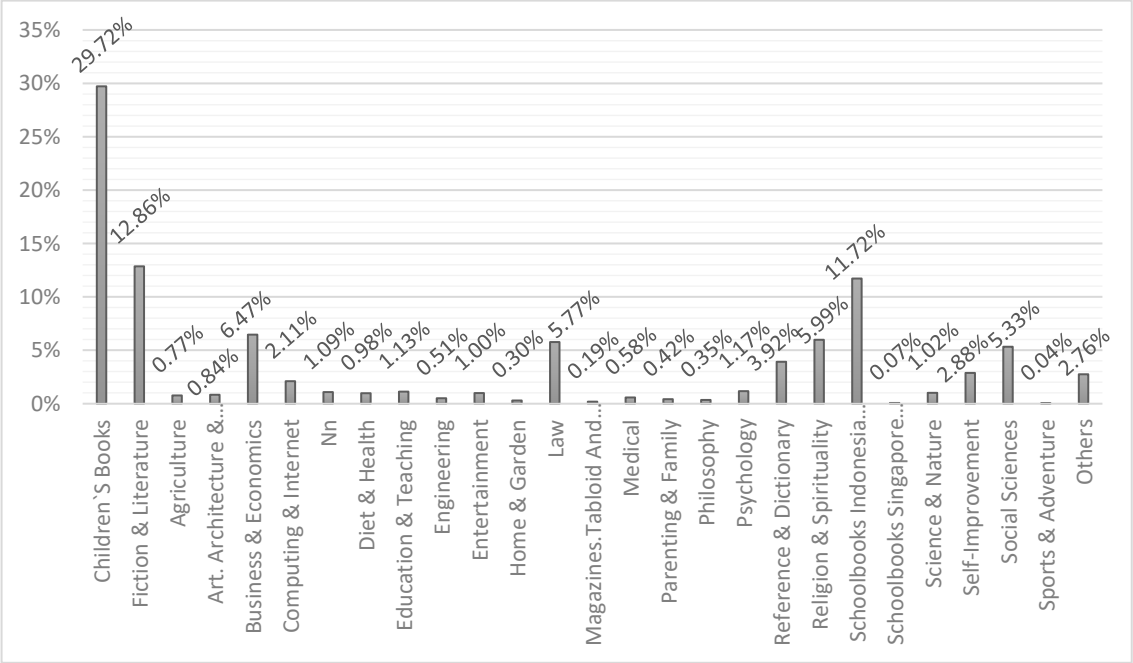

Figure 2. Percentage of Products Number Per Category

	Sequence ID	Event	∑#	Detail Event
1	11193900	1469836800	1	2005
2	11202100	1468800000	6	1001,1002,2001,2004,2009,2019
3	11209400	1468627200	1	1001
4	11234100	1468886400	3	1002,2003,2023
5	11234100	1468972800	2	2011,2023
6	11234100	1469059200	1	2023
7	11234100	1469145600	2	2007,2019
8	11234100	1469318400	1	2019
9	11234100	1469491200	3	2004,2011,2022
10	11234100	1469664000	3	1002,2004,2023

Figure 3. Example of Data Pre-Processing Result

3.3 Determine Sequential Pattern

This stage find the sequential pattern of the purchase transaction by determiningthe interrelationship between items / events on the sequential pattern. Before starting the process of determining the sequential pattern, we must first define the minimum support. The minimum support is choosen by trial and error, we try minimum value of support from 1% to value of minimum support which still produce sequential pattern. The minimum support values with the sequence resulted can be seen in Table 3.

The results obtained after the process is the ideal of minimum support that we can used is 0.001. These figures are used on the basis of analysis by comparing the minimum support that is perceived to provide large number ofsequence.

Table 3 shows the difference in the number of customer purchase patterns in XYZ Bookstore. The difference in the number of sequential patterns is influenced by the value of minimum support usedNumber of data processed is 49,717 so when the minimum support is 0.001, it means that the number of support for each sequence is more than 50 sequences. Therefore, the number of customer purchase patterns with a minimum support of 0.001 is 203 sequences. Each sequence has a support value of more than 50.

The purchasing recommendations used in this study are based on content-based filtering, that will searches which items are most often purchased by customers. Purchase recommendations are made after the sequential purchase patterns are generated. The results show that the minimum support results the most sequence occurrence is 0.001 or203 sequence. The frequent sequences generated are k-frequent sequence with k = 1, 2, 3, 4, 5, 6, 7.

Table 3. The Number of Sequential Purchase Patterns

No	*Minimum Support*	Number of *Frequent* Sequence	*Sequence Level* - (*Frequent*)						
			1	2	3	4	5	6	7
1.	1 sto 0.2	0	0	0	0	0	0	0	0
2.	0.1	2	2	0	0	0	0	0	0
3.	0.08 to 0.09	3	3	0	0	0	0	0	0
4.	0.05 to 0.07	5	5	0	0	0	0	0	0
5.	0.04	6	6	0	0	0	0	0	0
6.	0.03	8	8	0	0	0	0	0	0
7.	0.02	12	12	0	0	0	0	0	0
8.	0.01	21	18	2	1	0	0	0	0
9.	0.009	25	21	3	1	0	0	0	0
10.	0.008	31	27	3	1	0	0	0	0
11.	0.007	33	29	3	1	0	0	0	0
12.	0.006	35	31	3	1	0	0	0	0
13.	0.005	47	38	7	1	1	0	0	0
14.	0.004	59	45	12	1	1	0	0	0
15.	0.003	69	49	18	1	1	0	0	0
16.	0.002	100	72	25	1	1	1	0	0
17.	**0.001**	**203**	**123**	**63**	**11**	**3**	**1**	**1**	**1**

3.3.1 1-frequent sequence

An interesting pattern that can be obtainedfrom 1 frequent sequences is to find the highest support value of the value that appears. From 1 sequences that appear on this frequent sequences, the four highest sequences can be seen in Table 4.

Table 4. 1 Frequent Sequence with High Value of Support

Sequence	*Support (%)*	*Absolute Support*
<{CHILDREN`S BOOKS}>	**0.16392**	**8150**
<{FICTION & LITERATURE}>	0.11617	5776
<{SCHOOLBOOKS INDONESIA CURRICULUM}>	0.08407	4180
<{CHILDREN`S BOOKS,FICTION & LITERATURE}>	0.05577	2773

From Table 4 data, it can be concluded that the Children's Book category is the most frequent category for the sales transactions in the period of June 2016 to December 2016.

3.3.2 2-frequent sequence

In this 2-frequent sequence, it means the customer is repeat to comes 2 times. The result is 63 sequences are appearing, and four sequences with the highest support values can be seen in Table 5.

In the Table 5, we can be seen that the products with the most sold categories remain in the Children's Book and Fiction & Literature categories. In addition, the interesting patterns that appear on these 2-frequent sequence are the sequence of <{SCHOOLBOOKS INDONESIA CURRICULUM}, {SCHOOLBOOKS INDONESIA CURRICULUM}> with support 0.00911 or 453 transactions.

3.3.3 3-frequent sequence

On the 3-frequent sequence, it means the customer is repeat to come 3 times. The result is sequences appearing as many as 11 sequences, and 5 sequences with the highest support values can be seen in Table 6.

Results on these 3-frequent sequence are still largely the same as those found on 2-frequent sequence. The dominating categories are still in the Children's Book, Fiction & Literature, and Schoolbooks Indonesia Curriculum categories. But there are added sequences like<{BUSINESS & ECONOMICS}, {BUSINESS & ECONOMICS}, {BUSINESS & ECONOMICS}> with a relatively small support value of 0.00150 and the number of events is 74.

Table 5. 2-Frequent Sequence with High Value of Support

Sequence	*Support*(%)	*Absolute Support*
<{CHILDREN`S BOOKS},{CHILDREN`S BOOKS}>	0.02952	1468
<{FICTION & LITERATURE},{FICTION & LITERATURE}>	0.01673	832
<{SCHOOLBOOKS INDONESIA CURRICULUM},{SCHOOLBOOKS INDONESIA CURRICULUM}>	**0.00911**	**453**
<{CHILDREN`S BOOKS,FICTION & LITERATURE},{CHILDREN`S BOOKS}>	0.00599	298
<{FICTION & LITERATURE},{CHILDREN`S BOOKS,FICTION & LITERATURE}>	0.00544	271

Table 6. 3-Frequent Sequence with High Value of Support

Sequence	*Support* (%)	*Absolute Support*
<{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS}>	0.01061	528
<{FICTION & LITERATURE},{FICTION & LITERATURE},{FICTION & LITERATURE}>	0.00531	264
<{SCHOOLBOOKS INDONESIA CURRICULUM},{SCHOOLBOOKS INDONESIA CURRICULUM},{SCHOOLBOOKS INDONESIA CURRICULUM}>	0.00163	81
<{FICTION & LITERATURE},{FICTION & LITERATURE},{CHILDREN`S BOOKS,FICTION & LITERATURE}>	0.00163	81
<{BUSINESS & ECONOMICS},{BUSINESS & ECONOMICS},{BUSINESS & ECONOMICS}>	**0.00150**	**74**

3.3.4 4-frequent sequence untill 7-frequent sequence

There are 4 untill 7 sequences that appear in this frequent range can be seen in Table 7. It means the sama customer is repeat to come in 4 untill 7 times. The results shown in Table 7 are the formation of <{CHILDREN`S BOOKS}, {CHILDREN`S BOOKS}, {CHILDREN`S BOOKS}, {CHILDREN`S BOOKS}, {CHILDREN`S BOOKS}, {CHILDREN` S BOOKS}, {CHILDREN`S BOOKS}> with support value 0.001224.

3.4 Analizing Sequential Pattern of Consumer Purchasing

After we know the result of the sequential pattern mining then we can analizing the result based on content based filtering with sequential pattern mining approach. The following recommendations are :

1. From the data Table 4 it can be concluded that the Children's Book category become the most frequent category of sales transactions in the period June 2016 to December 2016. So the availability of products in this category is mandatory and shouldbe maintained because if the stock of books in this category is empty then it can be cause complaints from customers because customers do not get the desired goods. Other interesting sequences are <CHILDREN`S BOOKS, FICTION & LITERATURE}> with support value of 0.05577 which means as many as 2,773 transactions contained Children's Book and Fiction & Literature products. In other words, many customers who purchase products with the Children's Book category also purchase products with the category Fiction & Literature at the same time. The benefit is this can be a good opportunity for XYZ bookstores to create promotional programs in that category, for example if the customer purchases the Children's Book category book then they can also get a special price for products with the category Fiction & Literature.

2. In Table 5 it can be seen that the products with the most sold categories remain in the Children's Book and Fiction & Literature categories. In other words, if the customer buys the product today, the same customer will come back the next day to buy the product of that same category. In addition, the interesting patterns that appear on 2-frequent item set are <{SCHOOLBOOKS INDONESIA CURRICULUM}, {SCHOOLBOOKS INDONESIA CURRICULUM}> with the support value of 0.00911 or 453 events. This transaction occurs in May to June 2016 which is the beginning of school year. The benefit is this can be a reminder to the purchasing officer or merchandiser to anticipate the start of the new school year by filling out the products in XYZ bookstore with school book products.
3. The results that can be seen in Table 7 confirm that the Children's Book category is the most demanded category and the sales trend has never declined during this study period. Even with the sequence <{CHILDREN`S BOOKS} sequences, {CHILDREN`S BOOKS}, {CHILDREN`S BOOKS}, {CHILDREN`S BOOKS}, {CHILDREN`S BOOKS}, {CHILDREN`S BOOKS}, {CHILDREN` S BOOKS}> with support value of 0.001224 proves that there are 61 instances where customers buy products with this category at 7 different days during this study period. This sequence indicates that the product of the category CHILDREN` S BOOKS really become the flagship product of XYZ bookstore.

3.5 Purchasing Recommendation for Merchandiser

The recommendation related to the application development used by merchandising employees is by adding the sequential sequence pattern data into the currently used application module. The design of recommendation application can be seen in Figure 4.

Table 7. 4-Frequent Sequence until 7- Frequent Sequence with High Value of Support

Sequence	*Support* (%)	*Absolute Support*
<{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS}>	0.005169	257
<{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS}>	0.002857	142
<{FICTION & LITERATURE},{FICTION & LITERATURE},{FICTION & LITERATURE},{FICTION & LITERATURE}>	0.001905	95
<{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS}>	0.001768	88
<{FICTION & LITERATURE},{FICTION & LITERATURE},{FICTION & LITERATURE},{FICTION & LITERATURE},{FICTION & LITERATURE}>	0.001224	61
<{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS},{CHILDREN`S BOOKS}>	**0.001224**	**61**

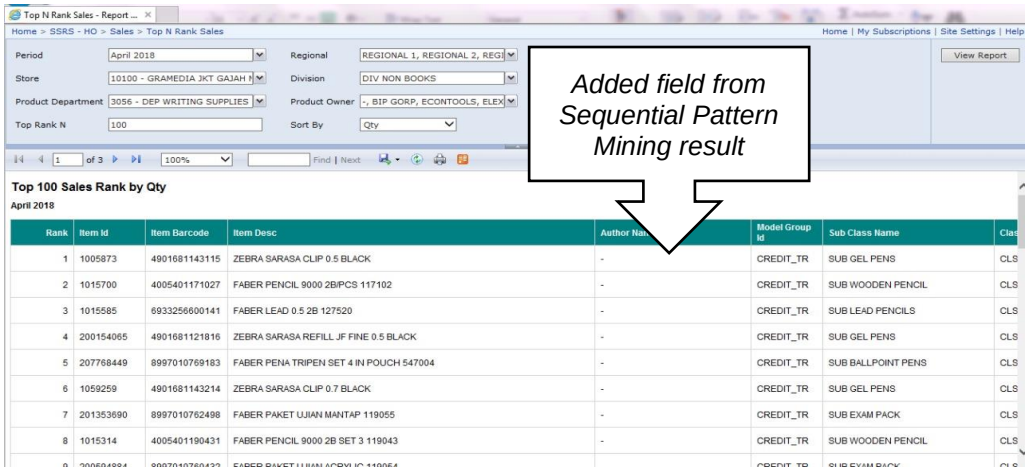

Figure 4. Recommendation application from Sequential Pattern Mining result

4. Conclusion

The SPADE algorithm successfully finds sequence patterns from the XYZ Book Store sales transaction data. Minimum support used to find the sequential pattern is 0.001%. Pattern analysis shows that there are 1,468 transactions from customers who buy products with the Children's Book category will buy products of the same category in another day. This study found that there are 7-frequent sequence which is supported by 61 transactions. This pattern means that there is a tendency of customers to buy the product of the same category for 7 different times consecutively. It also proves that the Children's Book category has a very high chance of being a Best Seller for a long time so that the purchasing officer on XYZ bookstore should ensure that the product's supply of the category is always safe throughout the year. If the category is promoted as a discount it will greatly boost and sales and if the product are given a bundling program with products with less-selling categories will be able to boost sales of the less-selling books. In the further study we can combined the sales transaction data with customer demographic data so that the results obtained illustrate the characteristics of customers with the product they buy.

References

[1] Sorin. *"MySQL Databases as Part of the Online Business, using a Platform Based on Linux," Database Systems Journa,* Vol. II, No. 3, 2011.
[2] Pachauri M, *"Consumer Buying Behaviour – A Literature Review,"* The Marketing Review Journal.e-ISSN : 1469-347X(p). DOI : 10.1362/1469347012569896, 2001.
[3] Levi E, Weitz S, "*Retailing Management. 7th ed*," New York: McGraw Hill, 2009.
[4] Zaki*,"SPADE: An Efficient Algorithm for Mining Frequent Sequences,"* Machine Learning Journal, 42(1/2):31–60, Jan/Feb, 2001.
[5] Kumar M, *"Sequential Pattern Mining With Multiple Minimum Support by MS-SPADE,"* International Journal of Database Management Systems (IJDMS), Vol.4, No.4, August 2012.
[6] Wanke, *"A Conceptual Framework for Inventory Management: Focusing on Low-Consumption Items,"* Production and Inventory Management Journal, Vol. 49, No. 1, 2014.
[7] Juliastio R and Gunawan*. "Sequential Pettern Mining dengan SPADE untuk Prediksi Pembelian Spare Part dan Aksesoris Komputer Pada Kedatangan Kembali Konsumen". Seminar Nasional "Inovasi dalam Desain dan Teknologi,"* IDeaTech 2015. ISSN: 2089-1121.
[8] Ricci F, *"Recommender System Handbook,"* Springer – London, 2011
[9] Pei J, Han J, Mortavazi-Asl B, Wang J, Pinto H, Chen Q, Dayal U, and Hsu M, *"Mining Sequential Patterns by Pattern-Growth:The PrefixSpan Approach,"* IEEE Transactionon Knowledge and Data Engineering, Vol. 16, No. 10, October 2004
[10] Agrawal R and Srikant R, *"Mining Sequential Patterns,"* In 11th International Conference on Data Engineering, Taiwan, 1996.