Data Mining Application using Association Rule Mining ECLAT Algorithm Based on SPMF

*Jason Reynaldo**, and *David Boy Tonara*

Management Information System, Ciputra University, UC Town, Citraland, Made, Sambikerep, Surabaya 60219, East Java, Indonesia

Abstract. Data mining is an important research domain that currently focused on knowledge discovery database. Where data from the database are mined so that information can be generated and used effectively and efficiently by humans. Mining can be applied to the market analysis. Association Rule Mining (ARM) has become the core of data mining. The search space is exponential in the number of database attributes and with millions of database objects the problem of I/O minimization becomes paramount. To get the information and the data such as, observation of the master data storage systems and interviews were done. Then, ECLAT algorithm is applied to the open-source library SPMF. In this project, this application can perform data mining assisted by open source SPMF with determined writing format of transaction data. It successfully displayed data with 100 % success rate. The application can generate a new easier knowledge which can be used for marketing the product.

Key words: Association rule mining, data mining, ECLAT, mining library, SPMF

1 Introduction

Data mining is an important research domain is currently focused on knowledge discovery in databases. Where data from the database are mined so that information can be generated and used effectively and efficiently by humans. Its objective is prediction and description. One of the aspects of data mining is the Association Rule mining. It consists of two procedures: first, finding the frequent item set in the database using a minimum support and constructing the association rule from the frequent item set with specified confidence. It relates to the association of items wherein for every occurrence of A, there exists an occurrence of B. This mining is more applicable in the market basket analysis. That application is helpful to the customers that buy certain items. That for every item that they bought, what would be the possible item/s coupled with the purchased item [1].

Association Rule Mining (ARM) has become the core data mining tasks and has attracted remarkable attention among researcher data mining. ARM is a data mining without direction or without supervision is a technique that works on a long variable data, and produce results that are clear and easy to understand [2].

* Corresponding author: jreynaldo@student.ciputra.ac.id

A new algorithm for ARM combine features item sets, depending on the format of the database, decomposition techniques, and procedures used search. One of them is ECLAT (Equivalence Class Transformation). The algorithm which was found by Zaki, not only minimize the cost of I/O by simply making a small number of database scans, but also minimize the cost calculations with efficient search schemes. This algorithm is very effective when used for small to medium item set [3].

This study uses ECLAT algorithm which is applied to the open-source library SPMF. SPMF will be used to run the algorithm ECLAT on restaurant xyz sales data as a case study. Therefore, researchers are interested in discussing about ECLAT Algorithm Implementation on sales data with xyz restaurant case studies using a program embed with open-source SPMF.

2 Related

The task of discovering all frequent associations in very large databases is quite challenging. The search space is exponential in the number of database attributes and with millions of database objects the problem of I/O minimization becomes paramount. However, most current approaches are iterative in nature, requiring multiple database scans, which is clearly very expensive. Some of the methods, especially those using some form of sampling, can be sensitive to the data-skew which can adversely affect performance. Furthermore, most approaches use very complicated internal data structures which have poor locality and add additional space and computation over-heads [3].

3 Analysis and design

To get the information and the data, there are some important things such as, observation of how the xyz restaurant run the master data storage systems and for every transaction that occurs at xyz restaurant, as well as conducting interviews to the xyz restaurant manager. Interview was conducted on Friday, December 16th, 2016 at xyz restaurant and below are the results obtained by the author.

3.1 Result interview

- (i) For each sales transaction in the xyz restaurant was already systemized, so every number of transactions, the customer, total payment and others are already recorded in the system, so it is very easy to pick and process the data using the application.
- (ii) For each transaction data processing that are very private for xyz restaurant, system must go through the authorities of the restaurant manager.

3.2 Result interview

In the project, the authors use the attributes of the data that collected by the authors to learn the sale of the food menu in the xyz restaurant. So it is easier for the application to analyze the data and the ease of use for the data mining application, it can be very helpful for manager or anyone on the part of the restaurant to be able to use the application easily and its main purpose is to make it easier for the sales promotion through data mining analysis application which the authors designed and built.

3.3 Data source

The data that used for this project is the data from the xyz restaurant that located in Surabaya, Indonesia. The data was taken from the xyz restaurant system for 12 months data transactions, it started from January 2016 until December 2016. With the Excel type data, the data obtained about 40 876 rows with eight data attributes columns and had 364 food menu data. The data was reduced to two columns of attributes. The attributes are:

Table 1. Attributes table.

No	Attributes	Explanation
1	Transaction ID	Transaction Id for each transaction.
2	Item Name	Name of goods in accordance with the master data.

3.4 System architecture design

Before getting the output from the analysis results that obtained from the application, there are several stages before preprocessing until post processing, which of these steps will be detailed in several segments below:

Fig. 1. System architecture design.

3.4.1 Collecting transaction receipts

First of all, researcher collect all data that start from interview from restaurant manager to get the transaction data starting from January 2016 until December 2016 with data format of Excel type (.xls) with 40 876 rows with eight columns of data attributes.

3.4.2 Preprocessing (.xls > .txt)

- Data reduction
Data reduction is to reduce the attributes used before entering into the preprocessing application manually by reducing the eight columns attribute into two columns attribute. There are transaction id and item name [4].

- Data transformation

Data Transformation is performed when data is converted into a text file, because SPMF can only read a text file [4].

The purpose of the transformation are: for each name of the menu in the Excel file will be converted into number of codes, the use of the transaction id is used as a separator for each new transaction that has transformed in vertical line and entered for each new transaction id.

- Data integration

Integration of data that occurs in sub preprocessing is to combine the data from January to December 2016 by inputting data per month one by one in the application and put together into a whole data which will be processed by SPMF and produce a result of information [4].

3.4.3 Sales data (.txt)

At the end of the preprocessing, the files format is changed to text files. And the form of numbers as a substitute of the name of the goods in the database used as a code of the food menu.

3.4.4 SPMF process

The output generated from the preprocessing stage above will be processed using open source data mining program that called SPMF. In the program we have chosen the ECLAT algorithm and then run the program to generate the text into the SPMF, so the SPMF can generate all the data based on ECLAT algorithm. The author will explain how ECLAT works on open source data mining SPMF, here is a flowchart drawing to explain how the ECLAT algorithm works.

Fig. 2. Flowchart of ECLAT algorithm [2].

Inputs for this ECLAT algorithm include a database of transactions and a threshold named minimum support (which is filled between 0 % and 100 %).

The transaction database is a set of transactions. For each transaction is a set of items, for example consider the following transaction database. There are five sample transactions (t1, t2, ..., t5) and 5 items (1, 2, 3, 4, 5). For example, the first transaction represents the set items 1, 3, and 4. It is important to note that items are not allowed to appear twice in the same transaction and that items are assumed to be sorted by lexicographical order in transaction.

Table 2. Attributes table.

No	Transaction ID	Items
1	t1	{1, 3, 4}
2	t2	{2, 3, 5}
3	t3	{1, 2, 3, 5}
4	t4	{2, 5}
5	t5	{1, 2, 3, 5}

ECLAT is an algorithm for discovering item sets (group of items) which occurs frequently in a transaction database. A frequent item sets is an item set appearing in at least minimum support transactions from the transaction database, where minimum support is a parameter given by the user.

For example, if ECLAT is run on the previous transaction database with minimum support of 40 % , ECLAT produces the following result:

Table 3. Examples of SPMF output.

No	Transaction ID	Items
1	{1}	3
2	{2}	4
3	{3}	4
4	{5}	4
5	{1,2}	2
6	{1,3}	3
7	{1,5}	2
8	{2,3}	3
9	{2,5}	4
10	{3,5}	3
11	{1,2,3}	2
12	{1,2,5}	2
13	{1,3,5}	2
14	{2,3,5}	3
15	{1,2,3,5}	2

Each frequent item set is annotated with its support. The support of an item set is how many times the item set appears in the transaction database. For example, the item set

{2,3,5} has support of 3 because it appears in transaction t2, t3, and t5. It is a frequent item set because its support is higher or equal to the minimum support parameter [3].

The input file format used by ECLAT is defined as follows. It is a text file. An item is represented by a positive integer. A transaction is a line in the text file. In each line (transaction), items are separated by a single space. It is assumed that all items within a same transaction (line) are sorted according to a total order and that no item can appear twice within the same line [5].

3.4.5 SPMF result

The output of the SPMF program is a file that contain the total number of transactions per set item that has been analyzed in accordance with the support or parameters.

3.4.6 Post processing (.txt > .pdf)

This stage is very important in this application because, at this stage, the data that initially on the preprocessing number of code which is an item of the transaction will be changed and put together into a sentence and paragraph which allows users to read the results of analysis that has been in data mining. In this post processing the author uses iTextsharp in visual studio plug-in which is used to assist in generating a new files with portable document format (.pdf) file format. Which is very easy and flexible for users in reading the results.

3.4.7. Knowledge (.pdf)

At this stage, knowledge is generated from the stages before, knowledge is generated in accordance with the result of SPMF. The purpose of the SPMF data changed to a sentence in the PDF is to make it easier for the novice user, because not everyone can read from the results that is generated by SPMF because SPMF does not provide the data extract directly into an easily understood result. Knowledge that can be used by ordinary users to facilitate in learning what food menu most often appear in every transactions in the restaurant and can be used as a reference promotion menu according to the number of transactions that occur as in the example.

SPMF Result

The result of the SPMF analysis with Eclat algorithm declare that :

Total number of sales items 2 :

Item "Teh manis (ES) " and Item "Nasi putih" has the highest number of transaction data 1038 from 40864 (2,54 %) transactions. Item "Teh manis (ES) " with the number 1232 from 40864 (3,01 %) and Item "Nasi putih" with the number 4168 from 40864 (10,20 %).

Item "Sayur asem" and Item "Nasi putih" has the highest number of transaction data 916 from 40864 (2,24 %) transactions. Item "Sayur asem" with the number 1447 from 40864 (3,54 %) and Item "Nasi putih" with the number 4168 from 40864 (10,20 %).

Item "Air mineral 600 ml" and Item "Nasi putih" has the highest number of transaction data 813 from 40864 (1,99 %) transactions. Item "Air mineral 600 ml" with the number 951 from 40864 (2,33 %) and Item "Nasi putih" with the number 4168 from 40864 (10,20 %).

Fig. 3. Generated knowledge.

3.5 Database design

Here is a database design that will be used to perform some data storage on the application, the author uses MySQL as a database.

- 1. For temporary transactions storage

Transaksi
no_transaksi
nama_barang

Fig. 4. Transaction database.

- no_transaksi: Varchar (50)
- nama_barang: Varchar(45)

- 2. Master database for items that converted into a number of code

Master Barang
no_barang
nama_barang

Fig. 5. Master item database.

- no_barang: Integer
- nama_barang: Varchar(100)

- 3. For temporary files result from SPMF

tempspm
jumlah
menu2

Fig. 6. Temporary database for SPMF result.

- jumlah: Integer
- menu2: Varchar (45)

4 Results and discussions

After the design and implementation have been done in the previous chapter, then will be proceed some test about data mining algorithm and application, which will be explained in this chapter.

4.1 Data accuracy test in preprocessing

4.1.1 Writing test for transactional data

Before transactional data test is done in preprocessing data process, each food menu in transactional data of restaurant xyz is stored one by one in the database and then transformed into numeric code which describe the names of the food and if there is new name in the transaction, a new code in the database will be generated. The stored data thus will be useful to help SPMF working with ECLAT algorithm. The sum of food menu in the previous 12 months that successfully get in the database are 364 menus (which means the code for each menu is filled until 364), and the data which failed to get in is zero, thus the food menu percentages that is already transformed is 100 % from all the transactional data in the last 12 months.

4.1.2 Preprocessing performance test

In this sub-chapter will be shown the results of performance test when the preprocessing is done. Here is the PC (personal computer)/laptop specification that was used in the performance test:

PC/laptop type	: Lenovo G400s
Display size	: 14 inch TFT LCD
Resolution	: 1366 × 768 px
Aspect ratio	: 16:9
Graphics Card	: Nvidia GeForce GT 720M
Graphics Memory	: 2 GB
Processor	: Intel(R) Core(TM) i5-3230M
Processor Speed	: 2.60 GHz (4 CPUs)
Memory	: 4 GB RAM
Storage	: 500 GB HDD

In this sub chapter will be shown the result of previous preprocessing with writing format and file in PDF, with the sentence structure and passages which make the user feel easier to read the application results and analysis, which designed for data mining.

There are the three biggest item sets that shown in final result. The final result of the application is a form of sentences and the order of biggest item set is on the top of the list, and is getting smaller in the third item set. The order in Figure 7 is matched with the order of code in item master database.

Table 4. Performance test result.

<i>Month</i>	<i>Total Rows</i>	<i>Time Passed</i>
January	3716	1 min 59 s
February	3436	1 min 51 s
March	2918	1 min 33 s
April	3419	1 min 40 s
May	3122	1 min 36 s
June	3003	1 min 31 s
July	4892	2 min 32 s
August	3263	1 min 40 s
September	3293	1 min 46 s
October	3113	1 min 37 s
November	3315	1 min 44 s
December	3656	1 min 53 s

SPMF Result

The result of the SPMF analysis with Eclat algorithm declare that :

Total number of sales items 2 :

Item "Teh manis (ES) " and Item "Nasi putih" has the highest number of transaction data 1038 from 40864 (2,54 %) transactions. Item "Teh manis (ES) " with the number 1232 from 40864 (3,01 %) and Item "Nasi putih" with the number 4168 from 40864 (10,20 %).

Item "Sayur asem" and Item "Nasi putih" has the highest number of transaction data 916 from 40864 (2,24 %) transactions. Item "Sayur asem" with the number 1447 from 40864 (3,54 %) and Item "Nasi putih" with the number 4168 from 40864 (10,20 %).

Item "Air mineral 600 ml" and Item "Nasi putih" has the highest number of transaction data 813 from 40864 (1,99 %) transactions. Item "Air mineral 600 ml" with the number 951 from 40864 (2,33 %) and Item "Nasi putih" with the number 4168 from 40864 (10,20 %).

Fig. 7. Final result of the application.

5 Conclusion

In this chapter will be described and explained about the conclusions of the results from the project. This application can successfully changed all data from the sales transactions. It can perform ECLAT data mining assisted by open source SPMF with determined writing

format of transaction data. It successfully displayed data with 100 % success rate according to the highest amount of sales in transaction. The application can generate a new easier knowledge in PDF file format which can be used for marketing the product.

Reference

1. S. A. Abaya. *IJSER*, **3**,7 (2012).
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.302.2946>
2. M.J. Zaki. *IEEE Transactions on Knowledge and Data Engineering*, **12**,3:372-390 (2010). <https://dl.acm.org/citation.cfm?id=628066>
3. Philippe Fournier. *Philippe-fournier-viger* [Online] from <http://www.philippe-fournier-viger.com/spmf/index.php?link=documentation.php#e1> (2017) [Accessed on 16 March 2017].
4. J.J. Davis, A.J. Clark. *Computers & Security*, **30**, 6-7:353-375 (2011).
<https://www.sciencedirect.com/science/article/pii/S0167404811000691>
5. S. Paul. *IJCSIT*, **2**,2:88–101 (2010). <http://airccse.org/journal/jcsit/0202csit8.pdf>.
6. E.J. Gardiner, V.J. Gillet. *J. Chem. Inf. Model.*, **55**, 9 :1781–1803 (2015).
<http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00198>
7. Z. Ma, J. Yang, T. Zhang, F. Liu. *IJDTA*, **9**, 5:251–266 (2016).
http://www.sersc.org/journals/IJDTA/vol9_no5/26.pdf.
8. S. Paul. *IJCSIT*, **2**,2 :88–101 (2010).
https://www.researchgate.net/publication/43121793_An_Optimized_Distributed_Association_Rule_Mining_Algorithm_in_Parallel_and_Distributed_Data_Mining_with_XML_Data_for_Improved_Response_Time
9. S.P. Ye, J.L. Ding. *Applied Mechanics and Materials*, **488-489** :961–965 (2014).
<https://www.scientific.net/AMM.488-489.961>
10. I. Bruha, A. Famili. *ACM SIGKDD Explorations Newsletter-Special issue on “Scalable data mining Algorithms”*, **2**, 2:110–114 (2000). <https://dl.acm.org/citation.cfm?id=381059>