International Journal of Civil Engineering and Technology (IJCIET)
Volume 9, Issue 7, July 2018, pp. 1660–1669, Article ID: IJCIET_09_07_178
Available online at http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=9&IType=7
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

A NOVELTY OF DATA MINING FOR PROMOTING EDUCATION BASED ON FP-GROWTH ALGORITHM

Ali Ikhwan*
Faculty of Information System, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
Doctoral Student, School of Computer and Communication Engineering,
Universiti Malaysia Perlis, Pauh, Perlis

Milfa Yetri
STMIK Triguna Dharma

Yohanni Syahra
STMIK Triguna Dharma

Jufri halim
STMIK Triguna Dharma

Andysah Putera Utama Siahaan
Faculty of Science and Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia
Doctoral Student, School of Computer and Communication Engineering,
Universiti Malaysia Perlis, Pauh, Perlis

Solly Aryza
Faculty of Science and Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia
Doctoral Student, School of Computer and Communication Engineering,
Universiti Malaysia Perlis, Pauh, Perlis

Yasmin Mohd Yacob
School of Computer and Communication Engineering,
Universiti Malaysia Perlis, Pauh, Perlis

ABSTRACT

The development of education at this time the increasing number of campuses are growing. Therefore every university wants to gain a lot of students in promoting the university. Many ways can be done for the determination of promotional strategies one of them by using techniques that exist in the data mining. The method used in this research by using the FP-Growth Algorithm. The FP-Growth algorithm is one of the alternative algorithms that can be used to select the most common data stack

(Frequent Item Set) in a data set. The FP-Growth algorithm is a development of the Apriori algorithm. As for some events in FP-Growth does not generate candidate because FP-Growth has the concept of Tree build in doing Itemset search. This research is done by studying some research which is often considered by great jokes especially marketing part in determining promotion what become its target. Variables used are Last Education, Home Address, Department, Choice Prodi. The Research Result is a software system for implementing The FP-Growth algorithm that uses the FP-Tree Development concept in finding Frequent Itemset.

Key words: Data Mining, Association Rules, Frequent Items Set, FP-Growth.

Cite this Article: Ali Ikhwan, Milfa Yetri, Yohanni Syahra, Jufri Halim, Andysah Putera Utama Siahaan, Solly Aryza, Yasmin Mohd Yacob, A Novelty of Data Mining for Promoting Education Based on FP-Growth Algorithm. *International Journal of Civil Engineering and Technology*, 9(7), 2018, pp. 1660-1669.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=9&IType=7

1. INTRODUCTION

Competition in the business world is rigorous [1]–[6], especially in promoting universities. Each campus is trying to provide the best facilities. Therefore every campus is trying to find the right strategy to promote every campus. According to developer experts in finding a strategy that can determine marketing strategy in promoting universities by maximizing service to the community either manually or computerized. One of the techniques used in the implementation of Data Mining is in the field of promotion. When promotional goals are not well defined and precise, in the sense of not seeking potential promotional targets, it will only spend a lot of time and cost that should be minimized through the selection of good promotional targets.

One way that can be applied is to apply the use of Data Mining [7]–[10] Because in the data mining itself there are ways and techniques in the fulfillment of needs, one of which is the need for extensive information, and from the information, we can we use as a decision on our choice or determine quality in determining a decision. The collection of data or information has many potentials to be a conclusion in deciding by doing analysis and dig a un-formation contained in a data. So it is possible to create a strategy in support of education promotion. From the broad definition of Data Mining, there are many types of analysis techniques that can be classified in Data Mining. Data Mining Technique to be used in this research is Association Rule. The notion of an association rule is to find associative rules between a combination of items.

As for this research will be discussed how to implement one of an algorithm in data mining, that is a common algorithm Pattern-Growth (FP-Growth). This algorithm is part of the association technique in data mining. The FP-Growth itself is one of the alternative algorithms that can be used to determine the most frequent set of data in a set of data. The FP-Growth algorithm is a data structure used in a tree called FP-Tree. Using FP-Tree, FP-Growth algorithm can directly extract frequent itemset from FP-Tree [11]–[13].

2. METHODOLOGY

2.1. Data Mining

Data Mining is the process of obtaining information by searching for patterns and hidden relationships in the heap of data. (Fadlina, 2014) Data Mining or often referred to as

knowledge discovery in database (KDD) is an activity that includes the collection, use of historical data to find regularities, patterns or relationships in large data. Data Mining Output can be used to help decision making in the future. The development of KDD causes the use of pattern recognition is reduced because it has become part of Data Mining [14].

2.2. Stage Associate

The Associate analysis is also known as one of the Data Mining techniques that became the basis of various other Data Mining techniques. Particularly one of the stages of an association analysis called high frequent pattern mining captures the attention of many researchers to produce efficient algorithms. The importance of an associative rule can be identified by two parameters, support (the value of support) is the percentage of the combination of items in the database and confidence [15]–[18].

2.2.1. Data Preparation

In many scientific fields, especially computer science, quality data is required through the preparation process of raw data. In practice, it was found that cleaning and preparation data required a total of 80% of effort to engineer data, thus making data preparation a crucial process. The importance of this process can be seen from three aspects, such as:

1. Real world data is dirty data. Real world data can contain data that is not, there is noise, not consistent, due to:
   a. Incomplete, ie lack of attribute value or only Contains aggregate data (example: address = "")
   b. Noise, which still contains errors and outliers.
   c. Inconsistent, ie data containing discrepancies in the code and the name or the shortness of the data is inconsistent.
2. High-performance mining systems require quality data. Data preparation or preparation produces fewer datasets than the original dataset, this can improve the efficiency of Data Mining. This step contains:
   a. Select relevant data
   b. Reduce data
3. Quality data produce a quality pattern. With data preparation, then the resulting data is quality data, which leads to a quality pattern as well, by:
   a. Returns incomplete data
   b. Correct an error, or eliminate outliers
   c. Fixed conflicting data

2.2.2. Association Rules

Association rule is one method that aims to find patterns that often appear among many transactions, where each transaction consists of several items so that this method will support the recommendation system through the discovery of patterns between items in transactions that occur [16], [19]–[21].

The basic methodology of association analysis is divided into two stages:

1. Analyse the high-frequency pattern
   This stage looks for a combination of items that meet the minimum requirements of the support value in the database. The value of an item's support is obtained by the following formula:
   a. $$Support\ (A) = \frac{\text{Number of transactions containing A}}{\text{Total transactions}} \quad (1)$$
   b. $$Support\ (A\ C\ B) = \frac{\text{Number of transactions containing A and B}}{\text{Total transactions}} \quad (2)$$

2. Establish associative rules
   After all high frequency patterns are found, then the associative rule that meets the minimum requirement for confidence by calculating the confidence of the associative rules A _B The confidence value of rule A _B.

2.2.3. Frequent Itemset

The first step in the association rule is to generate all possible item sets with possible itemsset that appear with the m-item is 2m. Because of the magnitude of computing to calculate the frequent itemset, which compares each candidate item set with each transaction, then there are several approaches to reduce the computation, one with an a priori algorithm.

2.2.4. Apriori

Apriori algorithm is used to find frequent items set that meets minsup and then get rule that meets minconf from frequent itemset earlier. This algorithm controls the development of candidate itemsets from frequent items set results with support-based pruning to eliminate unattractive itemsets by setting the minsup. The principle of this a priori is that when the itemset is classified as a frequent itemset, which has more support than previously set, then all subsets are also classed as frequent itemset, and vice versa.

2.3. Rule Generation

After getting a frequent itemset using an a priori algorithm, the next step is to get a rule that meets confidence. Since the rule generated comes from the frequent itemset, in other words, in calculating the rule using confidence, there is no need to calculate its support because all of the candidate rules that have been met meet the specified minsup. This calculation also does not need to loop scanning on the database to calculate confidence, simply by taking the itemset from the support.

2.4. FP-GROWTH Algorithm

(Sensuse, 2012) FP-growth is an alternative algorithm that can be used to determine the most frequent itemset in a data set. FP-growth uses a different approach than the paradigm used in the Apriori algorithm. (Ririanti, 2014). FP-Growth is an alternative algorithm that can be used to determine the most frequent set of data sets in a data set. The FP-Growth algorithm is a development of the Apriori algorithm. FP-growth is a method that is often a mining itemset without a candidate Generation. It builds a very dense data structure (FP-tree) to compress the original transaction database.

3. DATA ANALYSIS

Characteristics of FP-Growth algorithm is the data structure used is a tree called FP-Tree. Using FP-Tree, FP-growth algorithm can directly extract frequent Itemset from FP-Tree. Frequent itemset excavation using FP-Growth algorithm will be done by generating data tree structure or called FP-Tree. FP-Growth method can be divided into 3 main stages as:

1. Generation conditional pattern base generation stage,
2. Phase-building phase of FP-Tree, and
3. Stage of frequent itemset search.

The third stage is a step that will be done to get a frequent itemset.

Input: FP-Tree Tree

Output : Rt Sekumpulan lengkap pola frequent

Method : FP-Growth (Tree, null)

Procedure : FP-Growth (Tree, α)

{ For each combination notated Generate patterns For each get up
01: if Tree contain single path P;
02: then For each combination (Notade β) dari node-node dalam path do
03: Generate Patterns β α with support From nodes In path do β;
04: else For Each a1 In header From tree do }
05: Generate Patterns
06: Build β = a1 α With support = a1 support
07: if Tree β =

4. PREPROCESSED

Table 1 Frequency of Each Item Transaction

TID	ITEM
1	D1,B1
2	C1,A3,B1
3	A3,C2
4	D1,A4,B1,C2
5	D1,C1,A4
6	C1,A3
7	D1,C1,A3
8	D1
9	D1,C1,A3
10	D1,C1,A4
11	D1,C1,A4
12	A3,B1.C2
13	D1,C1
14	C1
15	D1,C1,A3
16	C1,A3
17	D1,C1,A3
18	-
19	D1,A4,C2
20	D1,A4,B1,C2
21	A3,B1,C2
22	A3,B1,C2
23	D1,B1
24	D1
25	D1,C1,A4,B1
26	D1,C1,A4,B1
27	D1,C1,A4
28	D1,C1,A3
29	C2
30	D1,A4,C2

So the next step is to form the FP-Tree tree by looking at the earlier table. The figure below illustrates the formation of FP-tree after reading Table 1.

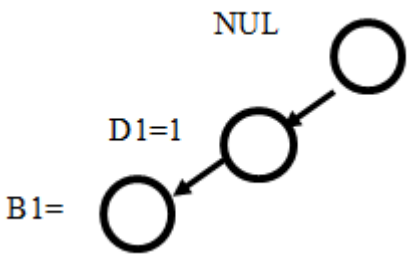

Figure 1 Result of FP-tree Formation After TID reading 1

Figure 4.1 is an explanation of the formation of FP-Tree after the reading is obtained after performing TID 1, ie it contains: NULL-D1 (Information System) = 1 - (B1) Medan = 1.

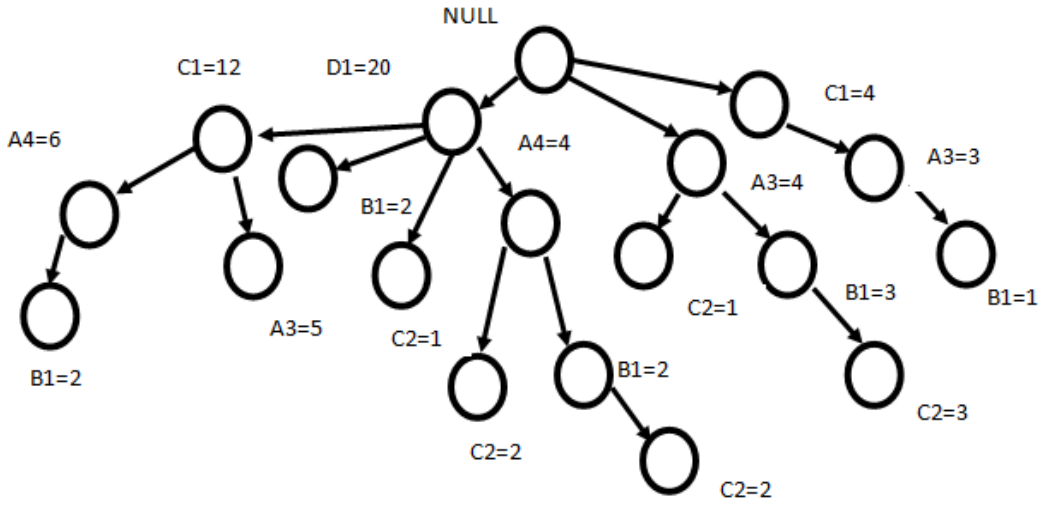

Figure 2. Result of FP-tree Formation After TID Reading 30

The image is obtained after the summed TID 30, which contains Null-Information System (D1) = 20 -IPA (C1) = 16 -SMA Country (A3) = 12 -SMA Private (A4) = 10- Medan (B1) = 10 -IP (C2) = 9. To find the Frequent itemset of table 4.2, it is necessary to first determine the path ending with the smallest support count, ie C2 followed by A4, A3, C1, D1 and ending D1.

Table 2. Frequent Itemset List Sorted by End Ties

Suffix	Frequent Itemset
C2	{C2},{C2,A4}{C2,A3}{C2,D1}{C2,B1,A3},{C2,B1,A4}
B1	{B1},{B1,D1},{B1,A3},{B1,A3,C1},(B1,A4,D1},{B1,A4,C1,D1}
A4	{A4},{A4,C1,D1}
A3	{A3},{A3,C1},{A3,C1,D1}
C1	{C1},{ C1,D1}
D1	{D1}

From the frequent itemset obtained from Fp-Tree and FP-Growth discontinuation it can be calculated the value of Support and Confidence as follows:

Support = (Information System, IPA, Private High School, Medan) = Count (Information System, IPA, Private High School, Medan) / Number of Transactions = 1/30.

Support = (Information System, IPA, PUBLIC HIGH SCHOOL) = Count (Information System, IPA, SMA) / Number of Transactions = 1/30.

Support = (Information System, Medan) Count (Information System, Medan) / Number of Transactions = 1/30.

Support = (Information System, Private High School, IPS) Count (Information System, Private High School, IPS) / Number of Transactions = 1/30.

Support = (Information System, Private High School, Medan, IPS) Count (Information System, Private High School, Medan, IPS) / Number of Transactions = 1/30.

Support = (SMA, IPS) Count (SMA, IPS) / Number of Transactions = 1/30.

Support = (SMA, Medan, IPS) Count (PUBLIC HIGH SCHOOL, Medan, IPS) / Number of Transactions = 1/30.

Support = (IPA, PUBLIC HIGH SCHOOL, Medan) Count (IPA, PUBLIC HIGH SCHOOL, Medan) / Total Transaction = 1/30.

As for the confidence or trust value is as follows:

Confidence = (SMA, IPS) = Count (PUBLIC HIGH SCHOOL, IPS) / Count Medan = 1/9.

Confidence = (Information System, Private High School, Medan, IPS) = Count (Information System, Private High School, Medan, IPS / Count Medan = 1/9.

Confidence = (PUBLIC HIGH SCHOOL, Medan, IPS) Count (PUBLIC HIGH SCHOOL, Medan, IPS / Count Medan = 1/9.

Confidence = (Information System, Private High School, IPS) Count (Information System, Private High School, IPS) / Count Medan = 1/19.

Confidence = (Information System, Private High School, Medan, IPS) Count (Information System, Private High School, Medan, IPS) / Count Medan = 1/9.

Confidence = (PUBLIC HIGH SCHOOL, Medan, IPS) Count (PUBLIC HIGH SCHOOL, Medan, IPS) / Count Medan = 1/9.

Confidence = (Information System, IPS) Count (Information System, IPS) / Count Medan = 1/9.

Confidence = (Information System, Private High School, IPS) Count (Information System, Private High School, IPS) / Count Medan = 1/9.

Having obtained the value of support and confidence of the overall combination of data with the calculation of FP-Tree and FP-Growth then obtain the highest and accurate support and confidence of the combination. (D1, C1, A3) {Information System, IPA, PUBLIC HIGH SCHOOL} which has a support value: 5/30 = 0.16 and the value of confidence: 5/20 = 0.25.

Lift Ratio is an important parameter besides support and confidence in association rule. Lift Ratio measures how important a rule is formed based on the value of support and confidence. Lift Ratio is a value that indicates the validity of the transaction process and provides information on whether product A is purchased together with product B.

Improvement Ratio can be calculated by the formula

$$\frac{Support(A \cap B)}{SupportA * SupportB}$$

To find valid value of rule is if have value of Lift Ratio> 1 by way of Support Lift Ratio = Support containing value A and value B divided support A * Support B support value containing value A and value B Is result of minimum support divided by Item (D1, C1, A3) {Information System, IPA, SMA} On every occurrence transacted. =

$$\frac{5/30}{20/30*16/30*12/30} = \frac{0{,}16}{0{,}66*0{,}53*0{,}4} =$$

$$\frac{0{,}16}{0{,}1392} = 1{,}1494$$

So from the search results Lift Ratio value> 1 then we can determine a valid rule of the many rules that run with Rapidminer software.

The results of the most influencing rule are: If he was originally a high school private school, the address of Medan and the origin of the Joints IPS then he chose Prodi Information System with 100% confidence level and supported 6% of the overall data. The results of the rule will be targeted in promoting education.

5. CONCLUSIONS

Methods in search Frequent Items set decision tree using FP-Growth algorithm works very well in doing Frequent Items set with FP-Tree formation process by generating a rule from new student sample data. Determination of data variables greatly determines the accuracy of FP-Growth made and the percentage in determining the drink support and minimum confidence is influenced by variable data used to find frequent itemset that is interconnected to find data variable that will be made strategy in education promotion. From research done on some attributes not used in the resulting rule, so the selection of attributes in the dataset is very important.

REFERENCES

[1] S. Aryza, M. Irwanto, Z. Lubis, A. P. U. Siahaan, R. Rahim, and M. Furqan, "A Novelty Design of Minimization of Electrical Losses in A Vector Controlled Induction Machine Drive," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 300, no. 1.

[2] R. Rahim *et al.*, "Insecure Whatsapp Chat History, Data Storage and Proposed Security," *Int. J. Pure Appl. Math.*, vol. 119, no. 16, pp. 2481–2486, 2018.

[3] A. P. U. Siahaan and R. Rahim, "Dynamic Key Matrix of Hill Cipher Using Genetic Algorithm," *Int. J. Secur. Its Appl.*, vol. 10, no. 8, pp. 173–180, Aug. 2016.

[4] V. N. S. Lestari, H. Djanggih, A. Aswari, N. Hipan, and A. P. U. Siahaan, "Technique for order preference by similarity to ideal solution as decision support method for determining employee performance of sales section," *Int. J. Eng. Technol.*, vol. 7, no. 2.14 Special Issue 14, 2018.

[5] G. Gunawan *et al.*, "Mobile Application Detection of Road Damage using Canny Algorithm," *J. Phys. Conf. Ser.*, vol. 1019, p. 012035, Jun. 2018.

[6] R. Rahim *et al.*, "TOPSIS Method Application for Decision Support System in Internal Control for Selecting Best Employees," *J. Phys. Conf. Ser.*, vol. 1028, p. 012052, Jun. 2018.

[7] M. J. A. Berry and L. G. S., *Data Mining Techniques For Marketing, sales, Customer Relationship Management, Second Edition,*. Wiley Publishing, Inc., 2004.

[8] K. Saputra and A. P. U. Siahaan, "Klasifikasi Data Minuman Wine Menggunakan Algoritma K-Nearest Neighbor."

[9] E. Hajizadeh, H. D. Ardakani, and J. Shahrabi, "Application of Data Mining Techniques in Stock Markets: A survey," *J. Econ. Int. Financ.*, vol. 2, no. 27, pp. 109–118, 2010.

[10] T. Larose D, *Data Mining Methods and Models*. New Jersey: Jhon Wiley & Sons, Inc., 2006.

[11] Supiyandi, M. I. Perangin-angin, A. H. Lubis, A. Ikhwan, Mesran, and A. P. U. Siahaan, "Association Rules Analysis on FP-Growth Method in Predicting Sales," *Int. J. Recent Trends Eng. Res.*, vol. 3, no. 10, pp. 58–65, 2017.

[12] W. Fitriani and A. P. U. Siahaan, "Comparison Between WEKA and Salford System in Data Mining Software," *Int. J. Mob. Comput. Appl.*, vol. 3, no. 4, pp. 1–4, 2016.

[13] L. Marlina, Muslim, and A. P. U. Siahaan, "Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms)," *Inte rnational J. Eng. Trends Technol.*, vol. 38, no. 7, pp. 380–383, 2016.

[14] M. H. Dunham, *Data Mining Introductory and Advanced Topics*. New Jersey: Prentice Hall, 2003.

[15] M. Dharma Tuah Putra Nasution *et al.*, "Decision Support Rating System with Analytical Hierarchy Process Method," *Int. J. Eng. Technol.*, vol. 7, no. 2.3, pp. 105–108, Mar. 2018.

[16] E. Turban, J. E. Aronson, and T. Liang, *Decision Support Sistems and Intelligent Systems*. Yogyakarta: Andi, 2005.

[17] Khairul, M. Simaremare, and A. P. U. Siahaan, "Decision Support System in Selecting The Appropriate Laptop Using Simple Additive Weighting," *Int. J. Recent Trends Eng. Res.*, vol. 2, no. 12, pp. 215–222, 2016.

[18] Y. Rossanty, D. Hasibuan, J. Napitupulu, M. D. T. P. Nasution, and R. Rahim, "Composite performance index as decision support method for multi case problem," *Int. J. Eng. Technol.*, vol. 7, no. 2.29, pp. 33–36, 2018.

[19] C. H. Primasari, R. Wardoyo, and A. K. Sari, "Integrated AHP, Profile Matching, and TOPSIS for selecting type of goats based on environmental and financial criteria," *Int. J. Adv. Intell. Informatics*, vol. 4, no. 1, p. 28, Mar. 2018.

[20] Y. Rossanty, S. Aryza, M. D. T. P. Nasution, and A. P. U. Siahaan, "Design service of QFD and SPC methods in the process performance potential gain and customers value in a company," *Int. J. Civ. Eng. Technol.*, vol. 9, no. 6, 2018.

[21] M. D. R. Pérez-Salazar, N. F. Mateo-Díaz, R. García-Rodríguez, C. E. Mar-Orozco, and L. Cruz-Rivero, "A genetic algorithm to solve a three-echelon capacitated location problem for a distribution center within a solid waste management system in the northern region of Veracruz, Mexico," *DYNA*, vol. 82, no. 191, pp. 51–57, Jun. 2015.

[22] Umamaheswari R, Siva Purnima S and Dr. S. Saravana Mahesan , Customer Preservence for an Organisation using Data Mining . International Journal of Civil Engineering and Technology, 8(10), 2017, pp. 933 – 938 .

[23] Mashael Saeed Alqhtani and M. Rizwan Jamee l Qureshi , Data Mining Approach for Classifying Twitter's Users . International Journal of Computer Engineering & Technology , 8(5), 2017, pp. 42 – 53 .

[24] Debarka Banerjee and Biswajit Roy , Application of Data Mining with R Programming to Implement ERP for Digitization of Valuation Reports on Commercial Vehicles and Passenger Cars . International Journal of Management , 8 (6), 2017 , pp. 10 9 – 1 29 .