Applied Soft Computing Journal 94 (2020) 106420

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

IECT: A methodology for identifying critical products using purchase transactions

Ping-Yu Hsu, Chen-Wan Huang [*]

Department of Business Administration, National Central University, No.300, Zhongda Rd., Zhongli Dist., Taoyuan City 320, Taiwan

ARTICLE INFO

Article history:
Received 15 August 2019
Received in revised form 11 April 2020
Accepted 21 May 2020
Available online 29 May 2020

Keywords:
Critical products
Data mining
Vertical database
Frequent itemsets
RFM

ABSTRACT

Identifying critical products and key customers to strengthen company performance is vitally important in the digital transformation era. Critical products are the itemsets that are preferred by vip customers and yet not popular among ordinary customers. As a result, critical products should be kept on the shelf despite its sales volumes may be lower than other popular items. However, few studies have considered identifying critical products or their potentially valuable patterns. Therefore an innovative algorithm taking advantage of vertical databases to identify critical products was designed. The proposed algorithm is applied to a transaction database of a midsize supermarket to verify the performance. The result showed that precision can reach 80.55% and 82.15% for two different filtering criteria. To the best of our knowledge, this study is the first to apply the concept of critical products to real retail industry transaction records.

1. Introduction

Identifying critical products and customers is an important task for retail stores because consumers purchase different products according to their individual situations and product promotion. Modern retailers all over the world need to adopt digital and data mining methodology to better understand products and customers.

For example, the Safeway retail chain once planned to remove several cheese brands from their shelves. Luckily, IBM was requested to investigate the product sales history before executing the plan, and identified that although cheese sales volumes were low, 25% of high spending customers often purchased them. Hence, removing these products could lead to higher spending customers shopping at other stores. Consequently, the cheese products were retained, even though they did not generate sufficient sales volume directly. Products such as these are called critical products, due to their specific appeal to high spending customers [1].

There are many benefits to identifying critical products, discovering important customers, and detecting different product combination patterns by customer segmentation. These provide reference information for commodity placements on shopping shelves, the store can provide targeted promotional programs for important customers when the critical products appears in a transaction, e.g. offering club membership if they are not already enrolled. Thus, researching critical products is an important topic.

Market basket analysis has been intensively studied for various applications due to its importance and the availability of transaction logs. Applications include identifying popular product combinations, high product utility, cyclic patterns, etc. [2–5]. Others investigated frequent patterns that also involve the combination of frequent and infrequent items in patterns [6–9]. And to the best of our knowledge, all the above mentioned research analysis transaction logs are among the same datasets of users.

However, this research calls for an algorithm to identify critical products that can attract high spending customers and yet does not have the same charm to ordinary customers. Therefore, an algorithm is required to simultaneously investigate transaction logs for different user datasets.

The contributions of this study are as follows:

- Propose a method and criteria to identify critical products and product combinations.
- The proposed method is memory based and only requires scanning the database once.
- Two filtering methods for the dataset achieved 80.55% and 82.15% precision, respectively, for customer transaction data from a midsize retail company.

This remainder of this article is organized as follows. Section 2 reviews previous market basket studies and vertical database mining algorithms. Section 3 describes the proposed method, and Section 4 discusses experimental data collection, cleaning, and the results. Finally, Section 5 summarizes and concludes the paper.

[*] Corresponding author.
E-mail addresses: pyhsu@ncu.edu.tw (P.-Y. Hsu), derek838@gmail.com (C.-W. Huang).

https://doi.org/10.1016/j.asoc.2020.106420
1568-4946/© 2020 Elsevier B.V. All rights reserved.

2. Literature review

Identifying critical products will assist retail companies to find their important customers. However few previous studies have considered critical products, and to the best of our knowledge, the current study is the first to apply the concept of the critical product to real retail industry transaction records.

2.1. Retail transaction logs analysis

Although previous studies have not directly considered critical product identification, several recent interesting research topics have included retail products and product combinations. Liu et al. [2] devised a bundle graph to represent customer preference, where each node represented a product attribute the consumer purchased and edges connecting nodes represented shared features. They developed an algorithm to derive the bundle graph from transaction logs. Subsequent studies have defined proprietary thresholds to find frequent itemsets for retail products [3].

Lismont et al. [5] predicted the next purchase period for a product based on transaction logs. Products and customers were organized into a bipartite network as nodes on the network with edges connecting nodes showing where customers purchased the products. To predict the next purchase period for a product based on transaction logs. And also constructed a unipartite network of products, displaying calculated product features such as centrality, betweenness, closeness, measure degrees, and node rank score. Product features along with customer recency, frequency, and monetary (RFM) characteristics were used as input data to devise models and to predict subsequent purchase periods.

Although these previous studies used retail product datasets, sources, partitioning, and pattern analytics differ significantly for the current study. We extracted the dataset from a real industry and huge database segregated by product attributes and identified critical products from more than twelve thousand products in total. Thus, the proposed approach simultaneously identified attributes for several distinct datasets, in particular identifying critical products and important customers.

2.2. Infrequent pattern mining

The proposed methodology identifies significant frequency patterns within the datasets. Traditional market basket mining methods identify product purchasing patterns using frequent itemset mining, or infrequent itemset mining to explore rare patterns [6–9].

However, the proposed approach not only incorporates frequent and infrequent itemset mining concepts, but also discovers valuable relationship ratios for several datasets. These meaningful numerical ratios help to extract critical products and combinations for different customer datasets.

Ding et al. [3] defined a pattern T to be infrequent or rare if

$$\forall t \in T \ support\ of\ t \leq min\_support$$

$$and\ \exists X, Y, X \cup Y = T\ and\ interest\,(X, Y) \geq min\_interest,$$

where

$$interest\,(X, Y) = support\,(X \cup Y) - support\,(X) * support\,(Y)\,.$$

Dong et al. [7] studied infrequent and frequent itemsets based on multiple level minimum supports and minimum correlation strength, defining infrequent itemset mining as

$$\frac{supp\,(X \cup Y) + interest\,(X, Y) - minsupport + mininterest + 1}{|supp\,(X \cup Y) - minsupport| + |interest\,(X, Y) - mininterest| + 1}.$$

And a pattern T be defined as infrequent if $support(T) < min\_support$, and proposed an algorithm based on tidsets to mine infrequent and rare itemsets [8]. Bakariya et al. [9] defined an algorithm for infrequent itemset mining where the candidate itemset was less than or equal to a pre-defined minimum threshold.

However, these previous studies did not find significant ratios or patterns, most exploring different infrequent itemset mining models. In contrast, the proposed approach not only considers infrequent itemset patterns, but also identifies significant and valuable ratios and coefficients for several customer segmented datasets. We focused on extracting business insights and management implications from frequent patterns and infrequent ratios and coefficients, including filtering criteria to identify critical products combined with customer value.

2.3. Vertical database mining algorithm

Itemset mining is an important subfield of data mining, deriving interesting and useful patterns in transaction databases. Traditional frequent itemset mining aims to identify itemsets that appear frequently together in customer transactions. Although itemset mining was designed for market basket analysis, it can be viewed more generally as identifying frequently co-occurring attribute value groups, which has been extensively studied [4,10–13].

Itemset mining method can process by the horizontal and vertical database structure. The Eclat algorithm is almost universally employed for vertical mining due to its fast performance. The underlying concept for Eclat is that the candidate itemset support can be most efficiently calculated by intersecting tidsets of suitably chosen subsets [14,15].

Vertical database mining has been extensively studied. Zaki et al. [16] proposed an efficient parallel method to discover relevant itemsets and devised an approach to overcome I/O costs and search efficiency problems and limitations. And proposed a vertical tidlist intersection concept, i.e., the Eclat algorithm, to find frequent itemsets and attributes [17].

Subsequent studies leveraged an extended Eclat algorithm performance benefits to address problems with huge candidate generation and multiple database scans, calculating tidlist intersections where the resulting number of tids is the frequency of each itemset. They also provided a useful critical review of Eclat related methods [8].

Gao et al. [18] devised an algorithm to solve frequent itemset mining performance problems, using a non-frequent two itemset approach to reduce a large number of useless candidate sets. The employed lost support values to reduce the time to count support and provide faster results for frequent sets.

This research leverage the benefits of tidset easy tracing and intersecting the tidsets of suitably selected itemsets, to process the computing and novel filtering in different user databases. Therefore, we adopted the vertical data structure to combine pattern filtering using a proposed improved equivalence class transformation algorithm to assist in identifying critical products and important customers.

3. Methodology development

The research framework is depicted in Fig. 1. The raw data are transaction data of each member in a supermarket. The customers are clustered with RFM information into vip and non-vip databases, which are organized into vertical databases.

Items low transaction supports are filtered. The frequent items are composed with each other to form candidate critical itemsets. The candidate sets are examined against minimum support of

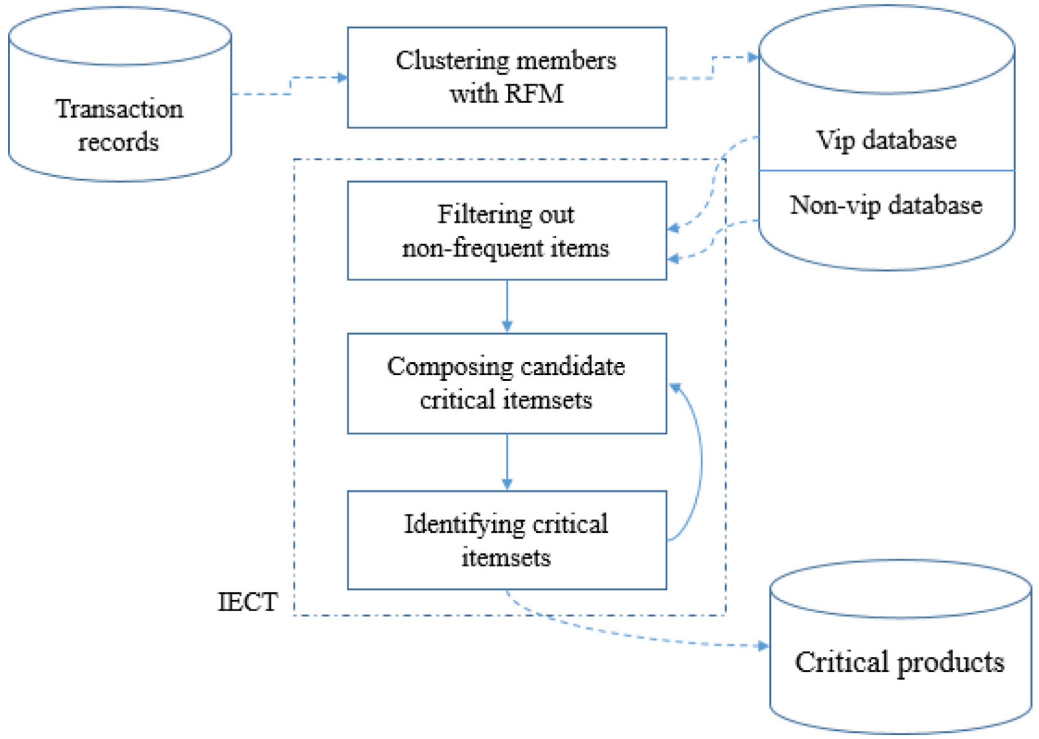

Transaction records
Clustering members with RFM
Vip database
Non-vip database
Filtering out non-frequent items
Composing candidate critical itemsets
Identifying critical itemsets
IECT
Critical products

Fig. 1. Method framework.

the vip database and a minimum ratio of the vip and non-vip database. The itemsets that satisfy both criteria are critical products.

We propose a memory based vertical database approach to identify critical products. Before delving into the methodology details we define and explain the required data structures.

3.1. Data structures

Horizontal database stores data in transaction records, which keep purchased items being recorded in each transaction. A sample horizontal database is shown in Table 1a. On the other hand, in a vertical database, each record corresponds to an item and store the transaction ids that include the item. An example of a vertical database is shown in Table 1b.

Table 1a
Horizontal database.

tid	Items
1	a, f, g
2	a, b, f, g
3	a, b, e, g
4	a, b, e, f, g
5	a, f, g
6	a, b, f
7	a, b, g
8	b, e, f, g
9	a, f

Table 1b
A sample vertical database.

I	a	b	e	f	g
D_1	1	2	3	1	1
	2	3	4	2	2
	3	4	8	4	3
	4	6		5	4
	5	7		6	5
	6	8		8	7
	7			9	8
	9				

Definition 1 (*Vertical Database and Transactions*).

Let $I = \{i_1, i_2, \ldots, i_n\}$, be a set of items, $T = \{t_1, t_2, \ldots, t_m\}$, be a set of transactions.

D be a vertical database of I and T. iff

$$\forall\left(\langle i_k, T_k\rangle \in D\right) \to i_k \in I, T_k \subseteq T, \forall t \in T_k, k\ being\ purchased\ in\ t.$$

P be an itemset of I, i.e., a vertical database of I and T, $P \subseteq I$.

Given D and P, $S_D(P)$ returns the set of transactions containing P in D, i.e.,

$$S_D(P) = \cap\{T_i | i \in P, \langle i, T_i\rangle \in D\}. \quad (1)$$

For instance $S_D(a) = \{1, 2, 3, 4, 5, 6, 7, 9\}$ and $S_D(b) = \{2, 3, 4, 6, 7, 8\}$.

With the customer value evaluation method and clustering algorithm, customers are divided into N+1 groups, where one group is vip and the N groups are non-vip. N is a parameter of this methodology and the value of N is supplied by researchers or market experts in the experiments.

The vip customers contribute considerably more revenue than the others. Non-vip customers were also partitioned into N groups according to the sales contribution to the retail. Table 2 shows a sample D_v, D_{n_1}, D_{n_2} where D_v comprised vip customer transactions. And $D_{n_1} and\ D_{n_2}$ are two non-vip customer groups, respectively

Critical products are itemsets frequently purchased by vip customers but significantly less commonly by non-vip customers. Several relationships are defined below to express the criteria.

A *prefix* is a frequent itemset in the vip database.

Definition 2 (*Prefix*).

Let I be a set of items, $P \subseteq I$.

P is a prefix if $|\cap_{i\in P}\left\{T|\langle i, T\rangle \in D_{vip}\right\}| \geq minsup$ (2)

Table 2
Example vertical databases of multiple datasets.

I	a	b	e	f	g
D_v	1 2 3 4 5 6	2 3 4 6	3 4	1 2 4 5 6	1 2 3 4 5
D_{n_1}	7	7 8	8	8	7 8
D_{n_2}	9			9	

Besides being frequently purchased by vip customers. A critical product must be rarely purchased by non-vip customers. In the current study, two criteria to distill the prefix for critical products are introduced. More factors can be included if required in future studies.

The first criteria is that the ratio of the non-vip transactions divide the vip transactions must be smaller than the minimum ratio threshold. The criteria is to guarantee that non-vip customers purchased only a limited ratio of the products.

Definition 3 (*First Criteria — Minimum Ratio*).
Let $P \subseteq I, D_v$ be a vip database, $D_{n_1} \ldots D_{n_N}$ be N non-vip databases and minR be a minimum ratio. Then

$$cr^1_{D_v,D_{n_1},\ldots,D_{n_N}}(P) = \begin{cases} true, & \frac{\Sigma_{i=1\ldots N} |S_{D_{n_i}}(P)|}{|S_{D_v}(P)|} \le minR \\ false, & o/w. \end{cases} \tag{3}$$

For example in Table 2, $\left|S_{D_{n_i}}(a)\right| / \left|S_{D_v}(a)\right| = 0.33$ and $\left|S_{D_{n_i}}(b)\right| / \left|S_{D_v}(b)\right| = 0.5$, hence if the criteria is $minR \le 0.4$, {a} meet the criterion whereas {b} does not.

Suppose customers in D_{n_N} are not vips, but are close, and hence their behavior is similar to vips. Therefore, we propose the second criteria to examine if customers in some group purchase more target *prefix* than other non-vip customers by checking the skewness coefficient [19] of transactions containing the target *prefix*.

Definition 4 (*Second Criteria — Filtering Non-vip Customer Groups*).
Let $P \subseteq I, D_{n_1} \ldots D_{n_N}$ be non-vip databases, then

$$\mu = \frac{\sum_{i=1\ldots N} \left|S_{D_{n_i}}(P)\right|}{N}., \sigma = \frac{\sqrt{\sum_{i=i\ldots N} \left(\left|S_{D_{n_i}}(P)\right| - \mu\right)^2}}{N}. \tag{4}$$

$$cr^2_{D_{n_1},\ldots,D_{n_N}}(P) = \begin{cases} true, & \frac{\Sigma_{i=1\ldots(N-1)} \left(\left|S_{D_{n_i}}(P)\right| - \mu\right)^3}{\sigma^3(N-1)} < 0 \\ & \vee \left(S_{D_{n_1}}(P) = \cdots S_{D_{n_N}}(P) = \varnothing\right) \\ false, & o/w \end{cases} \tag{5}$$

Special attention is required for $S_{D_{n_1}}(P) = \cdots S_{D_{n_N}}(P) = \varnothing$, i.e., where no non-vip customers purchase the *prefix*, hence *Skewness* =0/0, which is undefined mathematically. However, since there is no non-vip customer purchases, the *prefix* is also a perfect candidate critical products. Therefore we set $cr^2_{D_{n_1},\ldots,D_{n_N}}(P)$ = true for this case.

Hence a *prefix* is an itemset with supports in the vip database exceeding minimum support, and a *suffix* is a stack of items that could potentially expand the *prefix*. To expedite computation and facilitate vertical databases, the transaction sets supporting both *prefix* and target items are also recorded in the stack [20]. In this study, the items along with supported transactions are defined as vertical relation, *vr*.

Definition 5 (*Suffix and Vertical Relationship*).
Let vr be a relation of
$I \rightarrow 2^{T_v} \times 2^{T_{n_1}} \ldots, 2^{T_{n_N}}$,
where T_v, T_{n_1},..., T_{n_N} are transactions sets for vip and non-vip customers, respectively.
Then

$$vr(i) = \left\langle S_{D_v}(i), S_{D_{n_1}}(i), \ldots, S_{D_{n_N}}(i) \right\rangle \tag{6}$$

Let a suffix be a stack of vr entries.

For example, the vr (ab) = ⟨{2, 3, 4, 6} , {7} , {}⟩

3.2. Proposed improved equivalence class transformation algorithm

The proposed IECT algorithm is memory based and scans the database only once. The scan initially examines each item and keeps those that occur frequently in the vip database along with their *vr* into *stack_of_items*. The *prefix* initially is empty, and we use recursion on the stack in a depth first manner.

Items in *stack_of_items* are popped up one by one to examine if they meet the criteria. Target items that meet the criteria are concatenated with the *prefix* to form a critical product.

We loop through items remaining in *stack_of_items* to further expand the *prefix*. Each matching item and corresponding *vr* are pushed into *suffix*. Finally, a new IECT instance is initiated to examine if the current *prefix* can be further expanded. Initially IECT ([] , *stack_of_items*, *minsup*, *minR*) is instantiated. The returned value of IECT is the set of critical products.

3.3. IECT Example

For the example dataset in Table 2, initially *stack_of_items* = ({a}, [{1, 2, 3, 4, 5, 6}, {7}, {9}]), ({f}, [{1, 2, 4, 5, 6}, {8}, {9}]), ({g}, [{1, 2, 3, 4, 5}, {7, 8}, ∅), ({b}, [{2, 3, 4, 6}, {7, 8}, ∅])].

Fig. 2 shows the *prefix* set and *stack_of_items* for the first IECT instance, for $minsup \ge 4$ and $minR \le 0.4$.

In the first iteration, {b} is popped up and placed in the *prefix*. However, since $cr^1_{D_v,D_{n_1},D_{n_2}}(b) = false$, {b} is not a critical product. To process the other items {a} is a *suffix*, since only support of $(a) \ge minsup$.

Fig. 3(a) shows the parameters after the second call to IECT. In this IECT instance, {a} is popped up, and {ba} is in *prefix*. Since $cr^1_{D_v,D_{n_1},D_{n_2}}(ba) \wedge cr^2_{D_{n_1},D_{n_2}}(ba)$ are both true, {ba} is a critical product. Since the stack is empty (Fig. 3(b)), IECT instance {b} is terminated.

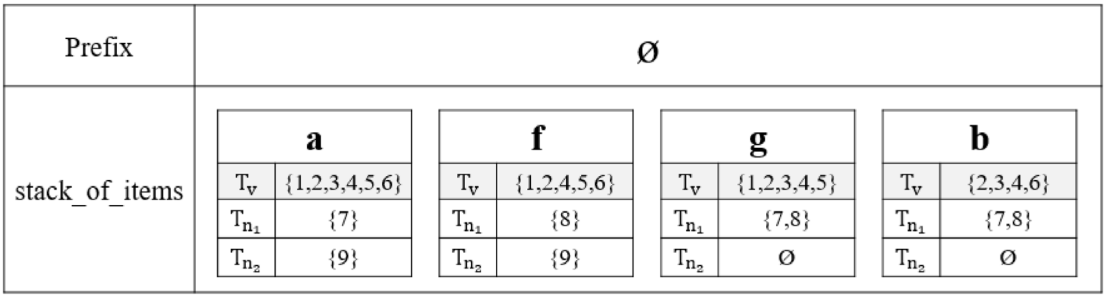
In the next iteration, {g} is popped up, and placed in the *prefix*. Since $cr^1_{D_v,D_{n_1},D_{n_2}}(\mathrm{g}) \wedge cr^2_{D_{n_1},D_{n_2}}(\mathrm{g})$ are both true, {g} is a critical product. Since support of $(a)\, and\, (f) \ge minsup$, {a} and {f} both are the *suffix* to process other items in *stack_of_items*.

Fig. 4(a) shows the parameters for the next IECT call. In this IECT instance, {f} is popped up, and {gf} is in *prefix*. Since $cr^1_{D_v,D_{n_1},D_{n_2}}(gf) \wedge cr^2_{D_{n_1},D_{n_2}}(gf)$ are both true, {gf} is a critical product. Since support of (a) $\ge minsup$, {a} is a *suffix*.

Fig. 4(b) shows the parameters for the next IECT call. In this IECT instance, {a} is popped up, and {gfa} is in *prefix*. Since $cr^1_{D_v,D_{n_1},D_{n_2}}(gfa) \wedge cr^2_{D_{n_1},D_{n_2}}(gfa)$ are both true, {gfa} is a critical product. The stack is empty, as shown in Fig. 4(c). Another *suffix* {a} is popped up, and {ga} is in *prefix*. Since $cr^1_{D_v,D_{n_1},D_{n_2}}(ga) \wedge cr^2_{D_{n_1},D_{n_2}}(ga)$ are both true, {ga} is a critical product. Since the stack is empty (Fig. 4(d)), IECT instance {g} is terminated.

In the next iteration, {f} is popped up, and {f} is in *prefix*. However, since $cr^2_{D_{n_1},D_{n_2}}(f)$ *is false*, {f} is not a critical product. Since

```
Algorithm: IECT (Improved Equivalence Class Transformation)
    Input: prefix, stack_of_items, minsup, minR
    Output: CP (critical products)
    while $stack\_of\_items \neq \emptyset$
        $\langle i, T_v, T_{n_1,\ldots,}T_{n_N} \rangle \leftarrow pop.stack\_of\_items$
        $suffix = [\,]$
        if $\left(cr^1_{T_v, T_{n_1},\ldots,T_{n_N}}(i)\right) \leq minR \wedge \left(cr^2_{T_{n_1},\ldots,T_{n_N}}(i)\right) = True$
            $CP \leftarrow CP \cup \{append(prefix, i)\}$
        for $\langle i_2, T_{2_v}, T_{2_{n_1},\ldots,}T_{2_{n_N}} \rangle \in stack\_of\_items$
            if $|T_v \cap T_{2_v}| \geq minsup$
                push $\langle i_2, \{T_v \cap T_{2_v}\}, \{T_{n_1} \cap T_{2_{n_1}}\}, \ldots, \{T_{n_N} \cap T_{2_{n_N}}\} \rangle$ to $suffix$
        IECT $(append(prefix, i), suffix, minsup, minR)$
    return CP
```



Prefix	∅
stack_of_items	see item tables

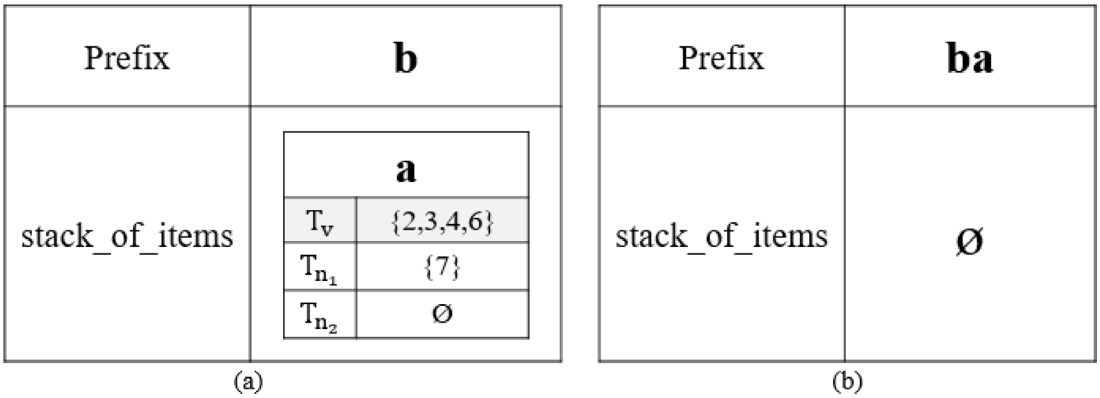
a	
T_v	{1,2,3,4,5,6}
T_{n_1}	{7}
T_{n_2}	{9}

f	
T_v	{1,2,4,5,6}
T_{n_1}	{8}
T_{n_2}	{9}

g	
T_v	{1,2,3,4,5}
T_{n_1}	{7,8}
T_{n_2}	∅

b	
T_v	{2,3,4,6}
T_{n_1}	{7,8}
T_{n_2}	∅

Fig. 2. Initial *stack_of_items*.



(a)

Prefix	b
stack_of_items	see item table

a	
T_v	{2,3,4,6}
T_{n_1}	{7}
T_{n_2}	∅

(b)

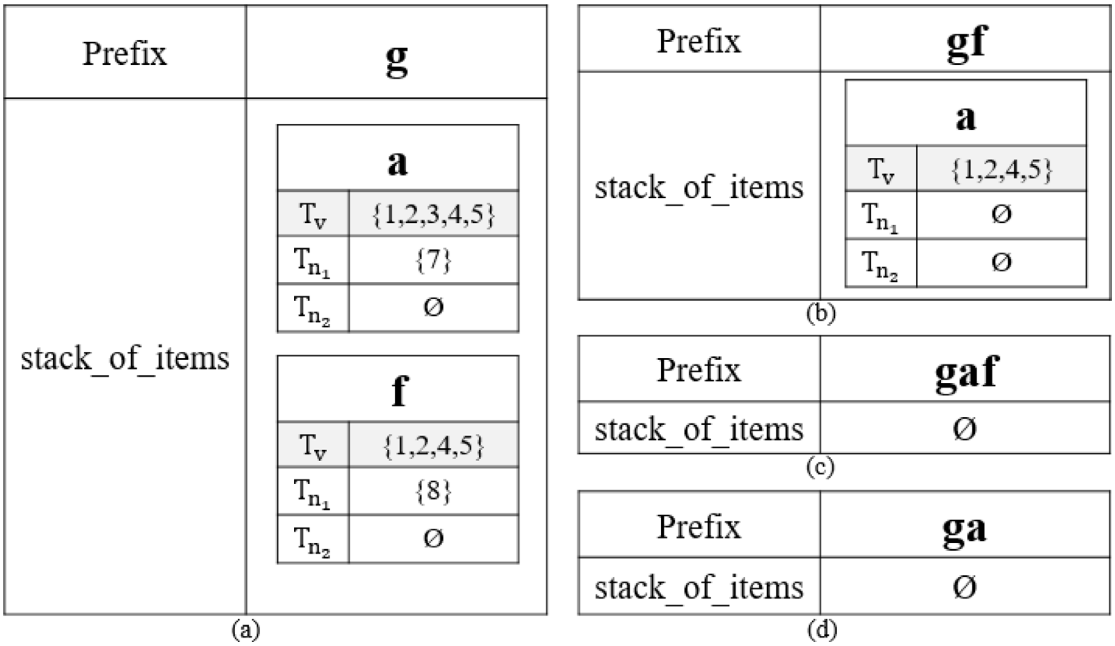
Prefix	ba
stack_of_items	∅

Fig. 3. IECT input for instance b.

support of $(a) \geq$ *minsup*, {a} is a *suffix*. Fig. 5(a) shows the parameters for the next IECT call. In this IECT instance, {a} is popped up, and {fa} is in *prefix*. However, since $cr^2_{D_{n_1},D_{n_2}}$ *(fa) is false*, {fa} is not a critical product. The stack is empty and IECT instance {f} is terminated.

In the final iteration, {a} is popped up, and {a} is in *prefix*. However, since $cr^2_{D_{n_1},D_{n_2}}$ (a) *is false*, {a} is not a critical product. Since the stack is empty (Fig. 5(b)), IECT instance {a} is terminated.

Fig. 6 shows the recursive steps for the sample database, and the final critical product sets = {ba, g, gf, gfa, ga}.



<table>
<tr><td>Prefix</td><td><b>g</b></td></tr>
<tr><td>stack_of_items</td><td>
<table>
<tr><th colspan="2"><b>a</b></th></tr>
<tr><td>T_v</td><td>{1,2,3,4,5}</td></tr>
<tr><td>T_{n_1}</td><td>{7}</td></tr>
<tr><td>T_{n_2}</td><td>Ø</td></tr>
</table>
<table>
<tr><th colspan="2"><b>f</b></th></tr>
<tr><td>T_v</td><td>{1,2,4,5}</td></tr>
<tr><td>T_{n_1}</td><td>{8}</td></tr>
<tr><td>T_{n_2}</td><td>Ø</td></tr>
</table>
</td></tr>
</table>

(a)

<table>
<tr><td>Prefix</td><td><b>gf</b></td></tr>
<tr><td>stack_of_items</td><td>
<table>
<tr><th colspan="2"><b>a</b></th></tr>
<tr><td>T_v</td><td>{1,2,4,5}</td></tr>
<tr><td>T_{n_1}</td><td>Ø</td></tr>
<tr><td>T_{n_2}</td><td>Ø</td></tr>
</table>
</td></tr>
</table>

(b)

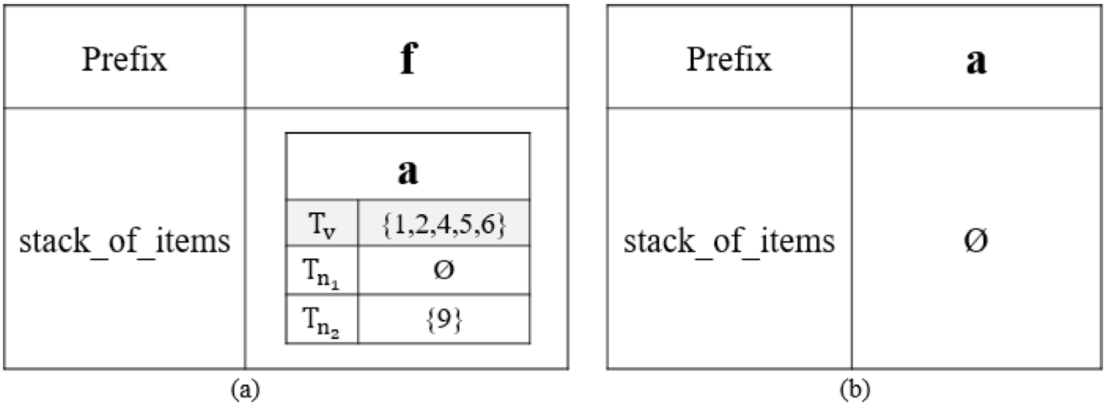
Prefix	**gaf**
stack_of_items	Ø

(c)

Prefix	**ga**
stack_of_items	Ø

(d)

Fig. 4. IECT input for instance g.



<table>
<tr><td>Prefix</td><td><b>f</b></td></tr>
<tr><td>stack_of_items</td><td>
<table>
<tr><th colspan="2"><b>a</b></th></tr>
<tr><td>T_v</td><td>{1,2,4,5,6}</td></tr>
<tr><td>T_{n_1}</td><td>Ø</td></tr>
<tr><td>T_{n_2}</td><td>{9}</td></tr>
</table>
</td></tr>
</table>

(a)

Prefix	**a**
stack_of_items	Ø

(b)

Fig. 5. IECT input for instance f and a.

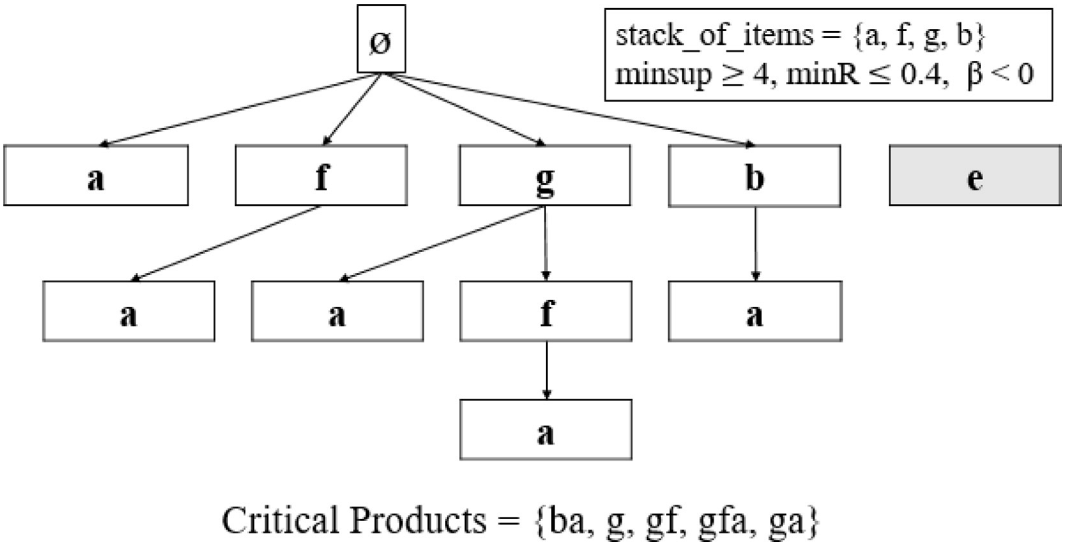

Ø
stack_of_items = {a, f, g, b}
minsup ≥ 4, minR ≤ 0.4, β < 0
a
f
g
b
e
a
a
f
a
a
Critical Products = {ba, g, gf, gfa, ga}

Fig. 6. Recursive steps for the sample example.

Table 3
Transaction statistic information.

Description	Statistic
Membership	3,819
Product categories	12,152
Total number of transactions	636,977
Average transaction amount per member	$10,482
Average purchased quantity per member	166

Table 4
RFM excerpted result.

ID	Date	Transactions	Spending	Recency	Frequency	Monetary
M0065	May.28	2	106	1	3	1
M0075	Oct.30	21	12,705	4	8	6
M0080	Nov.05	4	2,562	4	4	6
M0262	Nov.27	173	74,724	9	9	9
M0283	Nov.16	83	24,826	5	9	5

4. Experiment

4.1. Data collection and cleaning

The data resource for this study was extracted from a real retail industry transaction database, including membership transaction records (membership ID, transaction date, product purchasing record, etc.). Table 3 summarizes the 636,977 product transaction records remaining after data cleaning, including 3,819 membership records.

4.2. Customers evaluation and segmentation

Various of RFM methods are being used to summarize customer behavior and customer values to organizations [21–25]. RFM is an especially popular method in the retail industry because it captures customer behavior with three formula whose values can be calculated with transaction records [26–29].

The RFM method could extract customer features by using a small number of criteria and reducing the complexity of analysis. And the customer analysis is categorized by three variables of the recency (R), frequency (F), and monetary (M). The detailed definitions are as follows:

- Recency: Duration between the last purchase time and the specific survey time.
- Frequency: Purchase frequency refers to the transaction quantity in a specific time period.
- Monetary: the amount of consumption by the customer within a certain period.

Customer segmentation is mostly analyzed from two perspectives: demographics/geographic characteristics and actual purchasing behavior [24,28,29]. Demographics/geographic characteristics will need to collect relevant statistical data, such as race, gender, age, and geographic characteristics, etc. And customer behaviors are mostly analyzed with RFM [21–23,26,27]. Since this research collected only transaction data, RFM is utilized to cluster customers. However, the proposed method can work with any other customer segmentation method, as long as data are available. Therefore we adopt the RFM method in this research. The excerpted RFM result is shown in Table 4.

In this research, we predefined five customer groups according to Pareto's "The 80/20 Principle" [30], which states that 20% of customers contribute 80% of revenue. Accordingly, we divided the customer dataset into five segments, $D_v, D_{n_1}, D_{n_2}, D_{n_3}$ *and* D_{n_4}, ordered from highest contribution to lower. D_v is the vip customers and $D_{n_1}, D_{n_2}, D_{n_3}$ *and* D_{n_4} are non-vip. Table 5 shows the initial and final cluster results. And Table 6 shows the customer segmentation information with RFM scores.

Table 5
Cluster centers information.

Cluster centers information					
Initial	D_{n_3} 9	D_{n_1} 21	D_{n_2} 15	D_v 27	D_{n_4} 3
Final	D_{n_3} 9.95	D_{n_1} 19.91	D_{n_2} 15.11	D_v 24.9	D_{n_4} 5.18

Table 6
Customer segmentation information with RFM scores.

ID	Transactions	Spending	Recency	Frequency	Monetary	Segment
M0273	464	$29,602	8	9	8	D_v
M0288	438	$38,561	8	9	9	D_v
M0589	65	$10,577	5	5	6	D_{n_1}
M1249	24	$3,322	3	3	4	D_{n_2}

The proposed IECT methodology in this research can collaborate with any clustering algorithms, such as K-means and hierarchical clustering [22,31,32]. K-means is just a sample method utilized in this study. The reason k-means be adopted is for its simplicity and relatively low time and space complexity [1,14,21,23,24,27,28,33–36].

In the study, k is the number of databases, while R, F, M of each customer are the features for conducting clustering.

4.3. Critical products analytics

4.3.1. Bipartite segmentation

Customers along with their transaction data are randomly divided into 3 subsets. In each iteration, one subset is used for validation, while the other two are used for training. First, we integrated non-vip customers as a single group and investigated together with vip customer parameters. The initial minimum support threshold of frequent itemset mining is must greater or equal to 30. The minimum ratio between non-vip and vip is must less or equal to 0.143. Then obtained critical products for each dataset and validated using the testing dataset.

Second, we investigated different support thresholds based on minimum support and minimum ratio. The study partition intervals of purchased quantity as 30 to 39, 40 to 49 and 50 above, and further divide the ratio degrees as 0.143, 0.122, and 0.077, respectively. Table 7 shows the final experiment precision with an average precision of 80.55%, from the definition of minimum support and minimum ratio threshold.

From Section 4.3.1, this study found a significant pattern from the frequent purchasing of important customers in three intervals of different quantity thresholds. From the viewpoint of the purchased quantity, it is not purchasing a lot, to reach the highest precision rate. The highest precision at 82.77% (Fig. 7) that is obtained from the purchased quantity interval of 40 to 49 in this store. And observe the procurement ratio analysis in different customer groups, the experiment reveals the smallest ratio will reach the highest precision rate.

4.3.2. Multiple segmentation

We not only applied the ratio between non-vip and customers but also considering further filtering for non-vip customer groups that including the N$^{\text{th}}$ group of non-vip customers that purchased more itemsets than other non-vip customers. We retained the itemset as critical products if the distribution skewed to high contribution non-vip customers or the skewness criteria satisfied. Otherwise, the itemset was pruned. Table 8 shows the average precision result reached 82.15%.

Table 7
Precision result of ratio criteria.

Experiments	Threshold	1st Fold		2nd Fold		3rd Fold		TOTAL		Precision
		True	All	True	All	True	All	True	All	
50 above By ratio	**50 above total**	1,313	1,698	1,172	1,405	987	1,381	3,472	4,484	77.43%
	0.122<ratio<=0.143	111	148	769	902	268	342	1,148	1,392	82.47%
	0.077<ratio<=0.122	1,185	1,506	353	418	670	972	2,208	2,896	76.24%
	ratio<=0.077	17	44	50	85	49	67	116	196	59.18%
40 to 49 By ratio	**40 to 49 total**	472	568	976	1,179	699	847	2,147	2,594	82.77%
	0.122<ratio<=0.143	322	399	360	423	126	169	808	991	81.53%
	0.077<ratio<=0.122	110	126	589	728	267	335	966	1,189	81.24%
	ratio<=0.077	40	43	27	28	306	343	373	414	90.10%
30 to 39 By ratio	**30 to 39 total**	2,086	2,501	2,206	2,651	1,109	1,451	5,401	6,603	81.80%
	0.122<ratio<=0.143	311	377	685	813	285	370	1,281	1,560	82.12%
	0.077<ratio<=0.122	554	668	846	1,059	497	662	1,897	2,389	79.41%
	ratio<=0.077	1,221	1,456	675	779	327	419	2,223	2,654	83.76%
TOTAL	**Sub total**	3,871	4,767	4,354	5,235	2,795	3,679	11,020	13,681	**80.55%**
	0.122<ratio<=0.143	744	924	1,814	2,138	679	881	3,327	3,943	82.09%
	0.077<ratio<=0.122	1,849	2,300	1,788	2,205	1,434	1,969	5,071	6,474	78.33%
	ratio<=0.077	1,278	1,543	752	892	682	829	2,712	3,264	83.09%

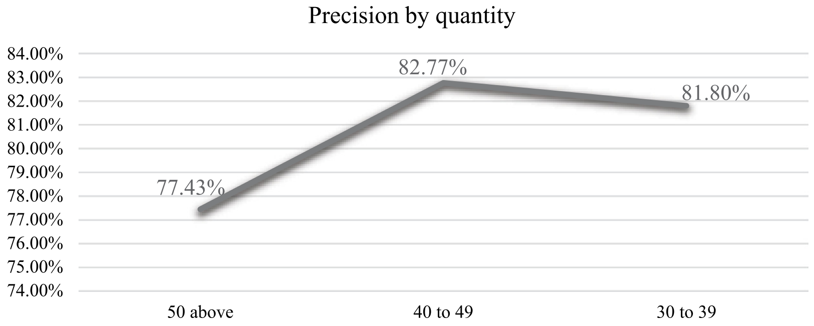


Fig. 7. Precision by quantity.

The results show an interesting pattern regarding the highest precision rate at 83.06% (Table 8) also come from interval 40 to 49 purchased quantity. Besides, the purchasing ratio between non-vip and vip customers obtained precision results at 81.63%, 81.80%, and 83.19%, respectively from a larger to smaller ratio. The higher precision arises from lower purchasing ratios and shows in Fig. 8.

Compare the results of Section 4.3.1 and 4.3.2 experiments. Processing further filtering in non-vip customer groups, the precision rate increased from 80.55% to 82.15%. This valuable finding will support the store to apply further criteria to get higher precision. The 1.96% precision improvement is important to identify more critical products and important customers, but also to assist business development and management.

According to the research observation, the highest precision rate results from the purchasing quantity of 40 to 49 purchased quantity. Table 9 shows the critical products by one item and two items that this study discovered a base on the interval of 40 to 49 purchased quantity. This finding assists in marketing promotion according to different product quantities and combinations.

4.3.3. Sensitivity analysis of value N

The experiment of this research is based on segmenting five customer groups (N+1=5). To evaluate if the methodology is sensitive to the value of N, more experiments are conducted, where N+1 is set as 4 and 6, respectively.

In the two experiments, validation evaluation is conducted with the hold out method with 70% of data being utilized for training and the remaining 30% of data are treated as testing data. The precision of experiments of four groups (N+1=4) and six groups (N+1=6) are 79.76% and 79.89%, respectively. Fig. 9 shows the precision comparison of the three kinds of segmentation.

4.4. Experiment summary

Based on the proposed methodology and setting the initial minimum support threshold of frequent itemset mining is must greater or equal to 30, the minimum ratio between non-vip and vip is must less or equal to 0.143. Table 10 summarizes the statistics for the identified 85 critical products.

From the analysis and observation, to define the minimum support threshold as quantity interval of 40 to 49, and also consider the further filtering skewness criteria in non-vip customer groups. We found 19 critical products, including 5 single products and 14 combinations of two products.

The proposed approach could found critical products and related important customers with high precision. Consequently, the store could utilize these outcomes to find their important customers by identifying critical products. Besides, it provides significant store benefits since vip customers comprised only 20% of total store membership. The store can adopt prompt action to the vip customer.

5. Conclusions

Identifying critical products is essential for all retail outlets in the digital transformation era. Critical products are purchased primarily by important customers, providing a facile mechanism

Table 8
Precision result of ratio and skewness filtering.

Experiments	Threshold	1st Fold		2nd Fold		3rd Fold		TOTAL		Precision
		True	All	True	All	True	All	True	All	
50 above By ratio	**50 above total**	778	923	900	1,069	658	873	2,336	2,865	81.54%
	0.122<ratio<=0.143	111	148	674	783	268	342	1,053	1,273	82.72%
	0.077<ratio<=0.122	662	752	176	201	373	500	1,211	1,453	83.34%
	ratio<=0.077	5	23	50	85	17	31	72	139	51.80%
40 to 49 By ratio	**40 to 49 total**	465	556	880	1,058	699	847	2,044	2,461	83.06%
	0.122<ratio<=0.143	322	399	360	423	126	169	808	991	81.53%
	0.077<ratio<=0.122	103	114	493	607	267	335	863	1,056	81.72%
	ratio <=0.077	40	43	27	28	306	343	373	414	90.10%
30 to 39 By ratio	**30 to 39 total**	1,858	2,238	1,875	2,251	1,044	1,332	4,777	5,821	82.06%
	0.122<ratio<=0.143	172	217	584	705	272	353	1,028	1,275	80.63%
	0.077<ratio<=0.122	465	565	616	767	445	560	1,526	1,892	80.66%
	ratio<=0.077	1,221	1,456	675	779	327	419	2,223	2,654	83.76%
TOTAL	**Total**	3,101	3,717	3,655	4,378	2,401	3,052	9,157	11,147	**82.15%**
	0.122<ratio<=0.143	605	764	1,618	1,911	666	864	2,889	3,539	81.63%
	0.077<ratio<=0.122	1,230	1,431	1,285	1,575	1,085	1,395	3,600	4,401	81.80%
	ratio<=0.077	1,266	1,522	752	892	650	793	2,668	3,207	83.19%

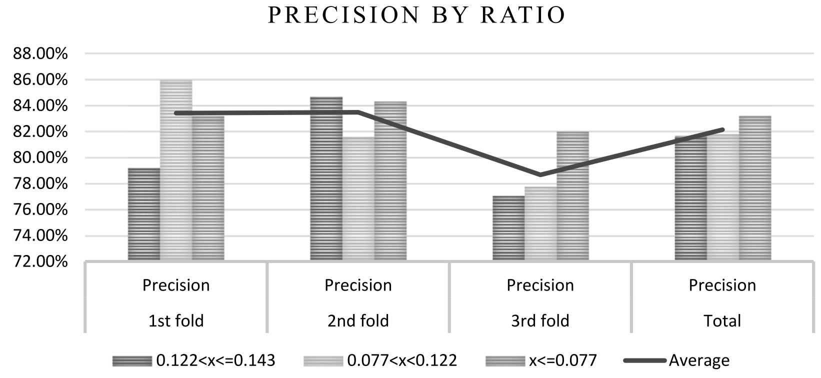


Fig. 8. Precision by ratio.

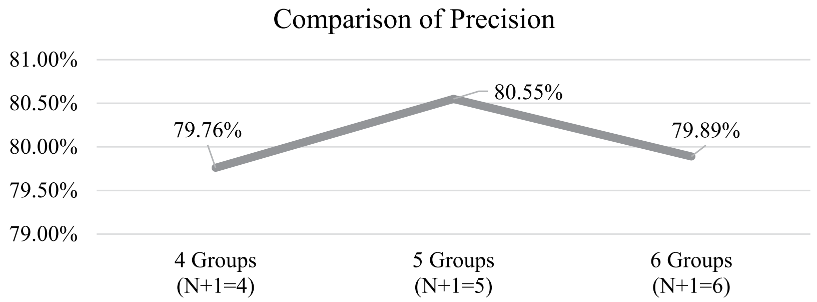


Fig. 9. Comparison of precision.

for the store to identify these customers and target them appropriately. However few studies have considered critical products, and most previous research only used transaction logs among the same datasets of users. In contrast, the current study identified critical products from real retail industry databases by analyzing different user datasets.

We proposed a memory based vertical database approach to identify critical products, scanning the database once only, analyzing frequent itemsets, significant ratio pattern, and skewness coefficient from the different datasets.

The proposed approach achieved an 80.55% precision rate by the innovative ratio definition and improved to 82.15% precision rate according to our novel methodology of further filtering in non-vip customers. We also identified 85 critical products, including 17 single and 68 two-item combinations from the original database.

These findings not only offer more marketing strategies for retail globally, but also provide management assistance to understand different consumer procurement behaviors, shelf and allocation planning, and membership program expansion. The

Table 9
Critical products details of purchase quantity 40 to 49.

Quantity 40 to 49	1st fold	2nd fold	3rd fold	Total
One-item	3	3	2	8
Two-items	4	7	5	16

Table 10
Final result of critical products.

Threshold	0.122<ratio<=0.143	0.077<ratio<=0.122	Ratio<=0.077	Total
50 above	1	2	3	6
40 to 49	7	6	6	19
30 to 39	13	17	30	60
Total	21	25	39	85

store could use the proposed process to identify critical products and valuable customers and hence increase company sales.

With its merit, the study is only the first step toward critical products discovery. Research is encouraged to verify the effectiveness of the methodologies with other retail transaction logs specifically. Future research can extend the model with a sequence of transaction logs, so as to research different aspects of critical products and important customers.

CRediT authorship contribution statement

Ping-Yu Hsu: Conceptualization, Methodology, Resources, Writing - review & editing, Supervision. **Chen-Wan Huang:** Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research was partially supported by the Ministry of Science and Technology (MOST 107-2410-H-008-037-MY2).

References

[1] M. Kantardzic, DATA MINING: Concepts, Models, Methods and Algorithms, second ed., John Wiley and Sons, New Jersey, 2011, http://dx.doi.org/10.1002/9781118029145.

[2] G. Liu, Y. Fu, G. Chen, H. Xiong, C. Chen, Modeling buying motives for personalized product bundle recommendation, ACM Trans. Knowl. Discov. Data 11 (3) (2017) http://dx.doi.org/10.1145/3022185.

[3] J. Ding, S.S.T. Yau, Tcom, an innovative data structure for mining association rules among infrequent items, Comput. Math. Appl. 57 (2) (2009) 290–301, http://dx.doi.org/10.1016/j.camwa.2008.09.044.

[4] U. Yun, D. Kim, Mining of high average-utility itemsets using novel list structure and pruning strategy, Future Gener. Comput. Syst. 68 (2017) 346–360, http://dx.doi.org/10.1016/j.future.2016.10.027.

[5] J. Lismont, S. Ram, J. Vanthienen, W. Lemahieu, B. Baesens, Predicting interpurchase time in a retail environment using customer-product networks: An empirical study and evaluation, Expert Syst. Appl. 104 (2018) 22–32, http://dx.doi.org/10.1016/j.eswa.2018.03.016.

[6] L. Zhou, S. Yau, Efficient association rule mining among both frequent and infrequent items, Comput. Math. Appl. 54 (6) (2007) 737–749, http://dx.doi.org/10.1016/j.camwa.2007.02.010.

[7] X. Dong, C. Liu, Mining interesting infrequent and frequent itemsets based on multiple level minimum supports and minimum correlation strength, Int. J. Serv. Technol. Manag. 21 (2015) 301–317, http://dx.doi.org/10.1504/IJSTM.2015.073941.

[8] M. Man, W.A.W. Abu Bakar, M.M. Abd. Jalil, J.A. Jusoh, Postdiffset algorithm in rare pattern: An implementation via benchmark case study, Int. J. Electr. Comput. Eng. 8 (6) (2018) 4477–4485, http://dx.doi.org/10.11591/ijece.v8i6.pp.4477-4485.

[9] B. Bakariya, G. Thakur, An efficient algorithm for extracting infrequent itemsets from weblog, Int. Arab J. Inf. Technol. (ISSN: 16833198) 16 (2) (2019) 275–280.

[10] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, ACM SIGMOD Rec. 22 (2) (1993) 207–216, http://dx.doi.org/10.1145/170036.170072.

[11] P. Fournier-Viger, J.C.-W. Lin, B. Vo, T.T. Chi, J. Zhang, H.B. Le, A survey of itemset mining, Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 7 (4) (2017) http://dx.doi.org/10.1002/widm.1207.

[12] Z.-H. Deng, Diffnodesets: An efficient structure for fast mining frequent itemsets, Appl. Soft Comput. J. 41 (2016) 214–223, http://dx.doi.org/10.1016/j.asoc.2016.01.010.

[13] L. Zhang, G. Fu, F. Cheng, J. Qiu, Y. Su, A multi-objective evolutionary approach for mining frequent and high utility itemsets, Appl. Soft Comput. J. 62 (2018) 974–986, http://dx.doi.org/10.1016/j.asoc.2017.09.033.

[14] M.J. Zaki, W. Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, New York, 2014.

[15] C. Zhang, P. Tian, X. Zhang, Z.L. Jiang, L. Yao, X. Wang, Fast eclat algorithms based on minwise hashing for large scale transactions, IEEE Internet of Things J. 6 (2) (2019) 3948–3961, http://dx.doi.org/10.1109/JIOT.2018.2885851.

[16] M.J. Zaki, S. Parthasarathy, M. Ogihara, W. Li, Parallel algorithms for discovery of association rules, Data Min. Knowl. Discov. 1 (4) (1997) 343–373, http://dx.doi.org/10.1023/A:1009773317876.

[17] M.J. Zaki, Scalable algorithms for association mining, IEEE Trans. Knowl. Data Eng. 12 (3) (2000) 372–390, http://dx.doi.org/10.1109/69.846291.

[18] Q. Gao, F.-L. Zhang, R.-J. Wang, Mining frequent sets using fuzzy multiple-level association rules, J. Electron. Sci. Technol. 16 (2) (2018) 145–152, http://dx.doi.org/10.11989/JEST.1674-862X.60408013.

[19] K. Black, Business Statistics for Contemporary Decision Making, sixth ed., John Wiley and Sons, US, 2010.

[20] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, second ed., MIT Press, London, 2001.

[21] M. Song, X. Zhao, H. E, Z. Ou, Statistics-based CRM approach via time series segmenting RFM on large scale data, Knowl.-Based Syst. 132 (2017) 21–29, http://dx.doi.org/10.1016/j.knosys.2017.05.027.

[22] P.A. Sarvari, A. Ustundag, H. Takci, Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis, Kybernetes 45 (7) (2016) 1129–1157, http://dx.doi.org/10.1108/K-07-2015-0180.

[23] A. Dursun, M. Caber, Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis, Tour. Manag. Perspect. 18 (2016) 153–160, http://dx.doi.org/10.1016/j.tmp.2016.03.001.

[24] E. Nikumanesh, A. Albadvi, Customer's life-time value using the RFM model in the banking industry: A case study, Int. J. Electron. Cust. Relationsh. Manag. 8 (2014) 15–30, http://dx.doi.org/10.1504/IJECRM.2014.066876.

[25] S. Khandelwal, A. Mathias, Using a 360° view of customers for segmentation, J. Med. Market. 11 (3) (2011) 215–220, http://dx.doi.org/10.1177/1745790411408853.

[26] T. Tanaka, T. Hamaguchi, T. Saigo, K. Tsuda, Classifying and understanding prospective customers via heterogeneity of supermarket stores, Procedia Comput. Sci. 112 (2017) 956–964, http://dx.doi.org/10.1016/j.procs.2017.08.133.

[27] S. Peker, A. Kocyigit, P.E. Eren, LRFMP Model for customer segmentation in the grocery retail industry: a case study, Market. Intell. Plan. 35 (4) (2017) 544–559, http://dx.doi.org/10.1108/MIP-11-2016-0210.

[28] M. Namvar, S. Khakabimamaghani, M.R. Gholamian, An approach to optimised customer segmentation and profiling using RFM, LTV, and demographic features, Int. J. Electron. Cust. Relationsh. Manag. 5 (2011) 220–235, http://dx.doi.org/10.1504/IJECRM.2011.044688.

[29] A. Hiziroglu, Soft computing applications in customer segmentation: State-of-art review and critique, Expert Syst. Appl. 40 (16) (2013) 6491–6507, http://dx.doi.org/10.1016/j.eswa.2013.05.052.

[30] R. Koch, The 80/20 Principle: The Secret to Achieving more with Less, Nicholas Brealey, London, 1998.

[31] P. Govender, V. Sivakumar, Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), Atmos. Pollut. Res. 11 (1) (1980) 40–56, http://dx.doi.org/10.1016/j.apr.2019.09.009.

[32] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (8) (2010) 651–666, http://dx.doi.org/10.1016/j.patrec.2009.09.011.

[33] D.C.S. Beddows, M. Dall'Osto, R.M. Harrison, Cluster analysis of rural, urban, and curbside atmospheric particle size data, Environ. Sci. Technol. 43 (13) (2009) 4694–4700, http://dx.doi.org/10.1021/es803121t.

[34] P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, Introduction to Data Mining, second ed., Pearson Education, New York, 2019.

[35] Y. Su, J. Reedy, R.J. Carroll, Clustering in general measurement error models, Stat. Sin. 28 (4) (2018) 2337–2351, http://dx.doi.org/10.5705/ss.202017.0093.

[36] L. de la Fuente-Tomas, B. Arranz, G. Safont, P. Sierra, M. Sanchez-Autet, A. Garcia-Blanco, M.P. Garcia-Portilla, Classification of patients with bipolar disorder using k-means clustering, PLoS One 14 (1) (2019) http://dx.doi.org/10.1371/journal.pone.0210314.