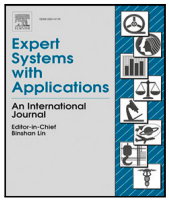
Expert Systems With Applications 237 (2024) 121449

Contents lists available at ScienceDirect

Expert Systems With Applications

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



RFM model customer segmentation based on hierarchical approach using FCA

Chongkolnee Rungruang [a], Pakwan Riyapan [b], Arthit Intarasit [b], Khanchit Chuarkham [c], Jirapond Muangprathub [d,*]

[a] College of Digital Science, Prince of Songkla University, Songkla 90110, Thailand
[b] Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand
[c] Faculty of Commerce and Management, Prince of Songkla University, Trang Campus, Trang 92000, Thailand
[d] Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani 84000, Thailand

ARTICLE INFO

Dataset link: https://archive.ics.uci.edu

Keywords:
Customer segmentation
Formal concept analysis
Hierarchical concept
RFM model
Clustering

ABSTRACT

Nowadays, every business focuses on customer relationship management (CRM) to deliver their customers better services and to establish a competitive advantage over their competitors. Significantly, customer insights with solid customer relationships improve customer retention and satisfaction, thereby contributing to profit. Thus, customer segmentation based on cluster analysis is critical to customer identification in CRM. In addition, it can identify the potential customers and their needs to be matched with marketing strategies. However, unfortunately, this approach has led to a gap between the marketing persons who care about the business implications and clustering output with the data science complexity barrier. Moreover, most clustering methodologies give only groups or segments, such that customers of each group have similar features without customer data relevance. Thus, this work sought to address these concerns by using a hierarchical approach. This research proposes a new effective clustering algorithm by combining Recency, Frequency, and Monetary (RFM) model with formal concept analysis (FCA). This new methodology uses the advantages of FCA in building the knowledge representation; therefore, the obtained construction contains both implicit and explicit knowledge. Explicit knowledge shows information represented in the hierarchical structure model, while implicit knowledge is embedded in the structure with its implication properties. Thus, the knowledge structure from FCA reveals relationships among data points in an easily understood manner. The proposed model was evaluated and compared with K-means clustering and hierarchical clustering using the online retail II dataset from the UCI Machine Learning Repository. The proposed method provides enough and appropriate information for marketers to perceive the value of the clustering results for creating practical marketing strategies in real-world business by offering the marketers both customer segmentation and the relationships in customer data at the same time.

1. Introduction

Segmentation, Targeting, and Positioning (STP) marketing is a core marketing approach for creating superior customer value and supporting the development of products and services (Gupta, Justy, Kamboj, Kumar, & Kristoffersen, 2021; Munusamy & Murugesan, 2020). The first step mainly addresses determining important characteristics to differentiate each market segment, leading to market targeting and product positioning. In competitive markets, businesses need to understand their customers to generate suitable matched marketing strategies. However, it is difficult to understand a massive number of customers clearly (Chen, Zhang, Chu, & Yan, 2019; Deng & Gao, 2020). Because of this, before companies can apply marketing strategies to their customers, they often use customer segmentation (also called market segmentation) to categorize the customers. Customer segmentation divides customers into groups according to their similarities in needs, characteristics, or behaviors, to maintain customer relationships and increase profit. Moreover, customer segmentation is a tool of customer identification that is the primary process of customer

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: chongkolnee.r@psu.ac.th (C. Rungruang), pakwan.r@psu.ac.th (P. Riyapan), arthit.i@psu.ac.th (A. Intarasit), ckhanchit@hotmail.com (K. Chuarkham), jirapond.m@psu.ac.th (J. Muangprathub).

https://doi.org/10.1016/j.eswa.2023.121449
Received 2 April 2023; Received in revised form 2 September 2023; Accepted 2 September 2023
Available online 15 September 2023

relationship management (CRM) (Deng & Gao, 2020; Dolnicar, Grün, & Leisch, 2018a). Notably, customer segmentation supports businesses in enhancing customer retention and loyalty and helps identify the value of a customer (Ballestar, Grau-Carles, & Sainz, 2018; Nandapala & Jayasena, 2020; Wu & Liu, 2020).

There are several customer segmentation methods, but most of them are based on customers' behavioral, psychographic, geographic, and demographic information. However, the customers' behavioral information based on RFM analysis is emphasized because of using a small set of features to segment customers surveyed in Alves Gomes and Meisen (2023). These factors will be used to segment customers with many algorithms. Segmentation algorithms can be divided into four main groups. They include association algorithms (e.g., Apriori, FP growth, ECLAT partition), clustering (e.g., hierarchical clustering, K-means clustering, fuzzy C Means clustering, density-based clustering, affinity propagation clustering), classification algorithms (e.g., KNN, Naïve Bayes, SVM, C4.5, Decision Tree) and regression algorithms (e.g., Logistic regression) (Tsiptsis & Chorianopoulos, 2011). Clustering is an unsupervised learning approach and can be divided into two categories: hard clustering and soft clustering (Singh & Srivastava, 2020). In hard clustering, each object is assigned to only one cluster, while soft clustering can also be done in an overlapping manner in which an object can be part of more than one cluster. After we have generated a partition using a clustering algorithm, we need to evaluate the validity or quality of this partition. If several partitions are generated (e.g., with different numbers of clusters), we need ways to compare them before we conclude the clustering result. There are three broad categories for cluster validity: internal, external, and relative indices (Tsiptsis & Chorianopoulos, 2011). The main approaches to assess these include graphical representations and internal indices. The silhouette plot is an advanced graphical representation that provides visual information about the quality of the partition. Internal indices measure a partition's "intrinsic" quality (how well-separated the clusters are). We have seen that the mean silhouette value can be used as an internal index. There exist many other internal indices. The Calinski–Harabasz, the Davies–Bouldin, and the Dunn indices are the three that are most widely used (Gagolewski, Bartoszuk, & Cena, 2021). Fuzzy Clustering is an example of soft clustering, and cluster validity indices have also been defined for this approach, one such index being the fuzzy silhouette index.

The K-means clustering algorithm is the most popular and is commonly used (Fränti & Sieranoja, 2018; Khalili-Damghani, Abdi, & Abolmakarem, 2018). However, it is sensitive to cluster centroid initialization and outliers, both impacting the eventual results (Deng & Gao, 2020; Meng et al., 2020). Unlike K-means, hierarchical clustering can produce a robust result without assigning initial values. Furthermore, the hierarchical clustering method is the most intuitive way of grouping data (Dolnicar, Grün, & Leisch, 2018b). It allows a human to track the task of dividing a set of n observations (customers) into k clusters (segments) represented in a hierarchy graphically shown as a dendrogram. In contrast, the result of K-means is unstructured non-overlapping clusters; therefore, the hierarchical clustering result is more interpretable and informative for marketers.

Creating effective marketing strategies in real-world business requires business people to have both implicit and explicit knowledge. However, hierarchical clustering only provides customer groups or segments in a hierarchy, lacking some knowledge such as the relevance of customer data. To solve this problem, this study proposes a new clustering algorithm that uses the RFM model and FCA to build knowledge representation and segmentation. The resulting construction contains both implicit and explicit knowledge. Explicit knowledge is information represented in the hierarchical structure model, while implicit knowledge is embedded in the structure with its implication properties. The FCA knowledge structure reveals relationships among data points in an easily understood manner. Additionally, a customer data representation based on FCA can uncover latent issues in customer information. The main contribution of this study is the use of FCA to build a knowledge structure for customer segmentation. FCA is ideal for identifying groups of customers with specific common properties or features. The presented structure has advantages in discovering both explicit and implicit knowledge. The marketers need these types of knowledge to create practical marketing strategies in real-world business. Namely, the explicit knowledge shows a group of customers with the same behavior, while the implicit knowledge shows the behavior associated with using the service in their business. To address this problem, this current study proposes a new effective clustering algorithm using the advantages of the RFM model and FCA to build knowledge representation and segmentation. The obtained construction contains both implicit and explicit knowledge. Explicit knowledge shows information represented in the hierarchical structure model, while implicit knowledge is embedded in the structure with its implication properties. Thus, the knowledge structure from FCA reveals relationships among data points in an easily understood manner. Above all, it is emphasized that a customer data representation based on FCA can be used to discover latent issues in customer information. Practically, the presented knowledge structure provides a decision-maker's approach to exact both customer segmentation and discover their behavior relevance.

This new method's clustering was compared with the most popular method, K-means clustering, and hierarchical clustering, using the online retail II dataset from the UCI Machine Learning Repository. After the segmentation, business implications and interpretation are discussed. Finally, conclusions are shared.

2. Background

2.1. Customer segmentation

Customer segmentation is the process of dividing the customer group into subgroups according to similarities (Choi, Choi, Yoon, & Joung, 2020; Deng & Gao, 2020; Singh & Mittal, 2021; Zeybek, 2018). A simple customer segmentation approach includes geographic, demographic, psychographic, and behavioral segmentation. Behavioral segmentation, being the most commonly used method, includes the following main steps: (1) business understanding and design of the segmentation process; (2) data understanding, preparation, and enrichment; (3) identification of the segments with cluster modeling; (4) evaluation and profiling of the revealed segments; and (5) deployment of the segmentation solution, design, and delivery of the differentiated strategies (Chorianopoulos, 2016; Tsiptsis & Chorianopoulos, 2011).

Segmentation can use several alternative algorithms, including an association algorithm, clustering algorithm, classification algorithm, or regression algorithm. Among these, clustering is the most precise and effective method for customer segmentation (Tsiptsis & Chorianopoulos, 2011).

Customer segmentation with a clustering algorithm uses machine learning to classify customer data based on similarity. All clustering algorithms aim to minimize the distances within clusters and maximize the distances between them. Moreover, the different clustering algorithms and outcomes are related to the objective (cost) function to measure the quality of clustering, the underlying structure assumed, the similarity measure, and how to consider the number of clusters.

Above all, most clustering techniques give only groups or segments without customer data relevance, whereas the marketing persons need an effective customer segmentation that gives appropriate knowledge in an easily understood format.

2.2. RFM model

RFM model is one of the most prevalent behavioral segmentations (Alves Gomes & Meisen, 2023; Hosseini, Abdolvand, & Harandi, 2022; Khalili-Damghani et al., 2018; Peker, Kocyigit, & Eren, 2017; Wang, Tsai, & Ciou, 2020). This model groups existing customers with their recency, frequency, and monetary values and does not focus on attracting new customers but identifies the best customers to perform targeted marketing. Recency is the number of days since the last purchase. Frequency refers to the total number of purchases. Meanwhile, monetary is the total purchase value during a specific period (Dedi, Dzulhaq, Sari, Ramdhan, Tullah, & Sutarman, 2019; Sokol & Holỳ, 2021).

The RFM scoring process is carried out to demonstrate RFM analysis using the quintile method. The customers are split into quintiles (five equal groups), and each customer is given a score based on which quintile he belongs to Christy, Umamakeswari, Priyatharsini, and Neyaa (2021). Accordingly, the scores assigned in the first quintile with the highest values for frequency (F) is 5, while the other quintiles are scored with 4, 3, 2, and 1 in rank order. This process is also undertaken for monetary (M). In contrast, the last quintile with the smallest recency (R) values is coded as 5. Next, we sequentially give the other quintiles based on recency values the scores 4, 3, 2, and 1. Finally, all the customers are ranked using R, F, and M values. Accordingly, the best customer group is 5-5-5, and the worst is 1-1-1.

However, the other approaches to the RFM model use the actual values of each RFM factor as the segmentation variables. The original RFM values are determined using various clustering techniques with required data preprocessing. Especially, K-means is the most popular and most commonly used method (Ernawati, Baharin, & Kasmin, 2021). Unfortunately, it is sensitive to cluster centroid initialization and outliers, so these affect the clustering results (Christy et al., 2021). In addition, it gives only complete separation of customer groups without customer data relationships. To ensure the accuracy of the K-means clustering results, the RFM values undergo preprocessing to identify and eliminate outliers. These outliers are then treated separately from the rest of the data (Chen, Sain, & Guo, 2012). Therefore, we utilize the RFM scoring technique with the quintile method to avoid the need for valid outlier removal. This study sought to create a practical approach to address these problems by using the RFM model, whose advantage is using a very small number of variables (only three variables), with clustering based on FCA. However, we can incorporate new variables into the RFM model to improve the accuracy and gain more information. An example is the RFMT model introduced by Zhou, Wei, and Xu (2021), which includes interpurchase time (T) to enhance customer segmentation. Other models, such as LRFM (Chang & Tsay, 2004), RFMTC (Yeh, Yang, & Ting, 2009), and RFMD (Noori, 2015), also include additional variables.

2.3. Formal concept analysis

Formal concept analysis (FCA) is a mathematical theory of concept formation based on lattice theory, applied in classification and concept discovery to organize information and discover relationships (Ganter & Wille, 2012; Wille, 2009). Moreover, FCA follows a human-centered approach and supports exploration operations through the concept lattice to organize information and discover relationships embedded in the binary relations between a pair of sets (called objects and attributes, respectively). A node in the concept lattice is an objects/attributes pair, called a (formal) concept. A concept consists of the extent (all objects belonging to the concept) and the intent (attributes describing the concept).

The core knowledge of the FCA approach is based on a simple data representation: a binary table called a formal context that is transformed into a mathematical structure called a concept lattice (Castellanos, Cigarrán, & García-Serrano, 2017). $\mathbb{K} := (G, M, I)$ is a formal context in which G represents a set of objects, M represents a set of attributes and $I \subseteq G \times M$ represents a set of is-a or has-a relationships between G and M, defined by gIm, which is read as the object g has the attribute m. I is the incidence relation of the context (G, M, I).

From this formal context, a set of formal concepts can be generated. To define a formal concept, the following derivation operations are needed. For any subsets A and B, $A \subseteq G$ and $B \subseteq$ M:

$$A \mapsto A' := \{m \in M \mid gIm \quad \forall g \in A\}$$
$$B \mapsto B' := \{g \in G \mid gIm \quad \forall m \in B\}$$

Thus, a formal concept is a pair (A, B) where $A \subseteq G$ is a set of objects (the extent of the formal concept), and $B \subseteq M$ is a set of attributes (the intent of the formal concept), which has the following properties:

- If all objects a in A are tagged with an attribute b, then b must be included in B (i.e., $B = A'$ the intent of the formal concept includes all the attributes shared by the objects in the extent).
- Conversely, if an object a is tagged with all the attributes in B, then a must be included in A (i.e., $A = B'$: the extent of the formal concept includes all those objects filtered out by the intent).

Formal concepts can be ordered according to their extents by applying the partial order relationships (Benavent, Castellanos, de Ves, García-Serrano, & Cigarrán, 2019) in Eq. (1), where a formal concept (A, B) with extent A is considered a sub-concept of another formal concept (C, D) with an extent C when the objects in A are contained into the objects in C.

$$(A, B) \leq (C, D) \Leftrightarrow (A \subseteq C \Leftrightarrow D \subseteq B). \tag{1}$$

The organization that results from this order relationship can be proven to be a lattice, a concept lattice, associated with the formal context denoted by $\mathfrak{B}(G, M, I)$. Since concept lattices are ordered sets, they can be naturally displayed in Hasse diagrams.

Frequency (support) is one of the most popular measures in the theory of pattern mining (Kuznetsov & Makhalova, 2018). Frequency arises from the assumption that the most "interesting" concepts are frequent ones:

$$\text{supp}(A, B) = \frac{|A|}{|G|} \tag{2}$$

The support provides an efficient level-wise algorithm of semilattice computing:

$$B_1 \subset B_2 \rightarrow \text{supp}\left(B_1\right) \geq \text{supp}\left(B_2\right). \tag{3}$$

In this study, we say that a set of attributes is frequent if its support exceeds a certain threshold. Thus, the frequency of an attribute set means that it is frequent.

Practically, clustering based on the FCA method contains four steps (Zhang, Zhao, & Yan, 2018). First, the dataset is pretreated, and feature items are extracted from data elements. Then, a formal context is constructed by taking data elements as objects and feature items as attributes. Next, the concept of formal contexts is extracted, and the concept lattice of the formal contexts is constructed. Finally, a Hasse diagram as a highly informative visualization can be generated.

It is emphasized that the FCA provides a well-defined mathematical framework to discover implicit and explicit knowledge in an easily understood format by using formal context and a Hasse diagram that clearly represents the concepts' generalization/specialization relationships (Ganter & Wille, 2012). In addition, FCA can construct an informative concept hierarchy providing valuable information on various specific domains. Therefore, we propose performing RFM model customer segmentation based on FCA.

Table 1
Study of RFM-based customer segmentation.

Resource	Dataset	Variables	Clustering techniques
Anitha and Patil (2022)	Online retail (Chen et al., 2012) Real-time retail	RFM	K-means
Hosseini et al. (2022)	Internet Banking	RFM	K-means and DBSCAN
Christy et al. (2021)	Online retail (Chen et al., 2012)	RFM	K-means, RM K-means, and Fuzzy C-means
Frasquet, Ieva, and Ziliani (2021)	Offline and Online retail	RFM + Sex + Age	Latent Class Analysis (LCA)
Rahim, Mushafiq, Khan, and Arain (2021)	Online retail (Chen et al., 2012)	RFM	K-means
Zhou et al. (2021)	Online retail	RFMT	Hierarchical
Shokouhyar, Shokoohyar, and Safari (2020)	Retail	RFM	K-means
Monalisa, Nadya, and Novita (2019)	Distribution and retail	RFM	Fuzzy C-means
Nakano and Kondo (2018)	Multichannel Retail	RFM	Latent Class Analysis (LCA)

2.4. Association rule mining

An association rule is just a statement about the conditional sample probability (called confidence) of an event w.r.t. another one, together with the statement of the joint sample probability of the two events (called support), where both events are described in terms of attribute sets. In FCA terms, for two subsets of attributes (called itemsets in data mining) Y_1 and $Y_2 \subseteq M$, the association rule $Y_1 \rightarrow Y_2$ has support $\frac{\left|(Y_1 \cup Y_2)'\right|}{|G|}$ and confidence $\frac{|(Y_1 \cup Y_2)'|}{\left|Y_1'\right|}$. The rule $Y_1 \rightarrow Y_2$ is called frequent if $\text{supp}\left(Y_1 \rightarrow Y_2\right) \geq \text{minsupp}$ for some threshold minsupp. In FCA, association rules are known as partial implications (Kuznetsov & Poelmans, 2013). This work applies the obtained FCA knowledge structure to extract implication rules to discover knowledge relationships in customer segmentation.

3. Related work

RFM model-based customer segmentation has been demonstrated using various clustering techniques, briefly summarized in Table 1. Most of these studies have applied K-means clustering to explicit customer segmentation surveyed in Hiziroglu (2013) and Alves Gomes and Meisen (2023). K-means is popular in the application areas because it is simple to understand, interpret, and apply. For example, Chen et al. applied K-means clustering and decision tree induction to segment customers from the online retail dataset of customer transactions in the UCI repository (Chen et al., 2012). In the same way, many studies have deployed dataset segmentation by using the K-means algorithm (Anitha & Patil, 2022, 2022; Chen et al., 2012; Christy et al., 2021). These studies have shown that data preprocessing is a crucial and time-consuming process before segmentation, especially before K-means clustering. For this type of clustering, data preprocessing needs to remove outliers, scale the ranges of data, and solve any long tails problems by data transformation (Tavakoli et al., 2018). In addition, we found that the number of segments chosen was between 2 and 10 groups. The number of segments applied should not be too large because it will make it difficult for the marketing analyst to interpret and design marketing strategies for selected customer segments. The Silhouette index is the most widely used validation index. The online retail dataset of customer transactions introduced by Chen et al. (2012) and placed in the UCI repository is the most often used. Therefore, this work uses this dataset for benchmarking and comparing the proposed model with K-means clustering and hierarchical clustering.

Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of observations (customers) into k groups (segments). Moreover, this is one of the most frequently used methods and can produce a robust result. Unlike other unsupervised methods, such as K-means, for hierarchical clustering, it is not necessary to assign any initial values (Zhou et al., 2021). Zhou et al. (2021) used hierarchical clustering for customer segmentation by web content mining. They used Calinski–Harabasz index and Davies–Doulbin index to determine the optimal cluster number. In addition, Chen, Zhang, and Zhao (2017) developed an approach to analyze customer behavior in mobile app usage and identify associations between functions by combining three data mining techniques: RFM analysis, link analysis based on graph theory, and association rule learning. Their proposed approach can be used to visualize and gain the advantages of these associations. Moreover, it provides insights into customer behaviors and function usage preferences. Remarkably, the prior work based on hierarchical clustering shows that segments and relationships are important and valuable knowledge for analyzing customer behaviors. Thus, this work proposes customer segmentation based on a hierarchical structure using FCA.

Poelmans, Ignatov, Kuznetsov, and Dedene (2013) utilized FCA's visualization abilities to discover the main research topics of papers on FCA that were published between 2003 and 2011. Their findings showed that FCA-based techniques were used by researchers for knowledge discovery and ontology engineering in various application domains, including software mining, web analytics, medicine, biology, and chemistry data. Moreover, it is worth noting that FCA possesses visualization capabilities. However, research on applications to CRM in customer segmentation is lacking, while some studies, such as the one by Marchetto (2005), have focused on software mining to extract concerns from web applications using MDSOC Hyperspaces definition and FCA. Another example is the work by Ravi, Ravi, and Prasad (2017), which proposed a hybrid model comprising fuzzy formal concept analysis and concept-level sentiment analysis (FFCA+SA). That research aimed to conduct opinion mining for CRM within the financial services sector. However, customer segmentation was not the main focus. Hence, there is still much to be explored in this area.

In Lei, Yan, Han, and Jiang (2018), association rules were extracted from concept lattices generated by FCA. This study proposed a feasible and effective method based on concept lattice and attribute analysis to reduce the number of association rules. Furthermore, they controlled the number of concepts by using rough attribute values and improved the concept's quality. Above all, this study emphasized that association rules from concept lattices can provide latent and valuable knowledge.

However, in practice, a concept lattice has many formal concepts that lead to the high complexity of conceptual clustering. Nguyen, Tran, Quan, Nguyen, and Le (2019) introduced a framework named MarCHGen (Malware Concept Hierarchy Generation) to generate a malware concept hierarchy. In this framework, they established frequent concepts on a concept lattice. A frequent concept has object sets larger than a defined threshold. Furthermore, only the frequent concepts are kept on the lattice. Therefore, a frequent lattice will be pruned to be less complex than the original lattice.

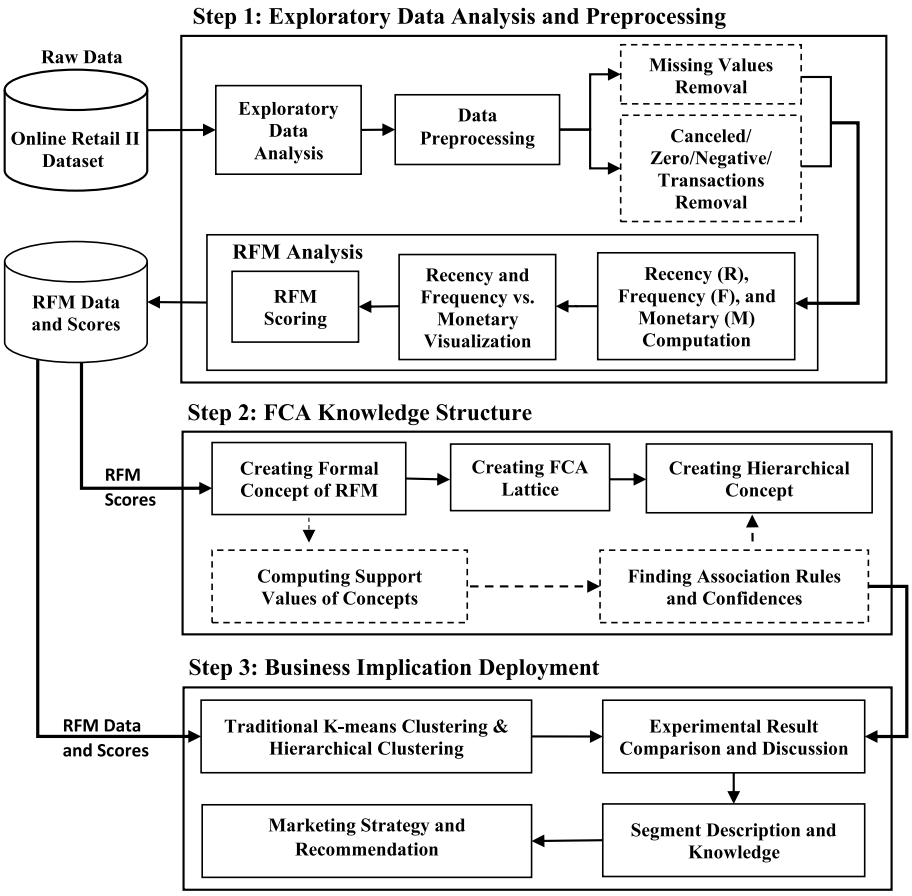


Fig. 1. Tasks and techniques of the proposed approach.

Above all, hierarchical clustering gives only groups or segments of customers in a hierarchy lacking some knowledge in the form of implications such as the relevance of customer data. To counter this problem, this research proposes a new effective clustering algorithm using the advantages of the RFM model and FCA to build knowledge representation and segmentation. Moreover, this appears to be the first time that the RFM model has been combined with the FCA.

4. Research methodology

The main contribution of this study is in applying FCA to build a knowledge structure for customer segmentation. FCA is ideal for identifying groups of customers with certain common properties (or features). The advantages of the presented structure are the discovery of explicit and implicit knowledge. The explicit knowledge is derived from identifying the co-appearance of attributes, while the implicit knowledge stems from the implication rules in relationships inside the concept lattice structure generated by the FCA. This section describes the process of building a hierarchical structure using FCA and its knowledge acquisition. Moreover, the traditional clustering approaches, K-means, and agglomerative hierarchical clustering are compared with the proposed approach.

The proposed methodology can be divided into three main steps, as shown in Fig. 1. The first step involves data exploration and preparation, which is critical for eventual accurate insights. Afterwards, the FCA knowledge structure provides explicit and implicit knowledge in customer segmentation and the implication relationships, respectively. The final step focuses on deploying the acquired knowledge in the customer segmentation and the discovered knowledge in each segment. The corresponding details are explained in the subsections that follow.

4.1. Exploratory data analysis and preprocessing

Exploratory data analysis (EDA) is the primary data exploration to extract and understand the data patterns by using statistics and graphical representations. In this research, we use the online retail II dataset used in Chen et al. (2012), Chen, Guo, and Li (2019), Christy et al. (2021), and Rahim et al. (2021) from the UCI Machine Learning Repository. This dataset contains all the transactions over two years, occurring for a UK-based and registered, non-store online retail, from 1 December 2009 to 9 December 2011. The company mainly sells unique all-occasion giftware. Many customers of the company are wholesalers. The customer transaction dataset contains 1,067,371 records with eight variables, as shown in Table 2.

Practically, we identify the unique customers, details of data, inconsistent data, incomplete data (or missing values), and noisy data. We prepare the dataset by cleaning steps, such as removing missing values and canceled/zero/negative transactions. Afterwards, the obtained dataset must be preprocessed for the required RFM model-based clustering analysis. The main steps and relevant tasks involved in the data preparation are as follows:

After this data preparation, there were 805,549 transactions remaining in the target dataset. Next, the RFM analysis concept, which is a popular and significant customer segmentation approach in the retail industry, will be used to reduce the data before generating a knowledge structure. The recency of transactions, frequency, and amount the customer spent are determined to create RFM values. Moreover, recency, frequency, and monetary values are visualized to gain knowledge and understanding of these data. Next, the day of the last purchase was decided as the reference date to calculate recency. At last, each customer's RFM score was created.

Table 2
Variables in the customer transaction dataset (1,067,371 instances).

Variable name	Data type	Description; typical values and meanings
Invoice	Nominal	Invoice number; A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
StockCode	Nominal	Product (item) code; A 5-digit integral number uniquely assigned to each distinct product.
Description	Nominal	Product (item) name;
Quantity	Numeric	The quantities of each product (item) per transaction.
InvoiceDate	Numeric	Invoice date and time; The day and time when a transaction was generated.
Price	Numeric	Unit price; Product price per unit in sterling (Â£).
Customer ID	Nominal	Customer number; A 5-digit integral number uniquely assigned to each customer.
Country	Nominal	Country name; The name of the country where a customer resides.

Table 3
Variables in the RFM scores and data (5878 instances).

Variable name	Data type	Description
CustomerID	Nominal	Corresponding to each distinct customer id
Recency	Numeric	Recency in days
Frequency	Numeric	Frequency of purchase per customer
Monetary	Numeric	Monetary or total amount spent per customer
R_score	Numeric	Recency scores [1, 5]
F_score	Numeric	Frequency scores [1, 5]
M_score	Numeric	Monetary scores [1, 5]

To calculate the RFM score, we can see that all the quantities calculated here (recency, frequency, and monetary) have different ranges. All the customers are ranked by considering their recency, frequency, and monetary values, and R-F-M codes represent them. Thus, we first convert these quantities to scores based on the quintiles in the target dataset. This step of the RFM scoring was computed as follows.

1. Create an aggregated variable named $Total$ by multiplying $Quantity$ with $Price$, which gives the total amount of money spent per product/item in each transaction.
2. Create three essential aggregated variables $Recency$, $Frequency$, and $Monetary$ to calculate the values of these variables for every customer:
   - Recency for variable $Recency$: the difference between the analysis date and the most recent date that the customer has shopped in the store. The analysis date has been taken as the maximum date available for the variable $InvoiceDate$.
   - Frequency for variable $Frequency$: The number of transactions performed by each customer.
   - Monetary for variable $Monetary$: Total money spent by every customer in the store.
3. Calculate RFM scores. All the values for frequency and monetary in the first quintile are given 1 for F_score and M_score, given 2 for the second quintile, and so on. For recency, a more recent customer will have less recency value than a customer who has not shopped in a while. Therefore, the recency values in the first quintile are given 5 for R_score, 4 for the second quintile values, and so on.

The result from data preparation with RFM analysis consists of 5878 instances summarized in Table 3. This prepared dataset is provided to build the knowledge structure in the next step.

4.2. FCA knowledge structure

This step uses RFM scores as the data to generate a formal context. We first transform the RFM scores (with three attributes and five levels in each) into a formal context containing binary data using 15 attributes to analyze and discover the latent concepts addressed in the customers and their relationships. An example of RFM score formal context is shown in Table 4. The first customer in the first row of this table has RFM scores of 2-2-5. Afterwards, we create the formal concepts of RFM scores and then create the FCA lattice following Eq. (1) of the theory in Section 2.3.

The hierarchical concept of this lattice will be built to provide a knowledge base of the proposed system. At the same time, the support values of the concepts will be computed to determine the partial implication association rules and their confidences. To construct the less complex hierarchical concept lattice with a high degree of visualization, we determine only the concepts whose number of objects is larger than a certain threshold.

From the hierarchical concept lattice, the customer segmentation is derived from the sublattice that considers the top view of this lattice. The relationship of customer behavior in each segmentation is derived from the implication rules following Eqs. (2)–(3) of the theory in Section 2.3. Those data will be prepared as described in the following subsection.

4.3. Business implication deployment

This subsection begins with experimenting with traditional types of clustering, namely K-means and agglomerative hierarchical clustering, by using RFM data and scores dataset.

As is well-known, the K-means clustering algorithm is very sensitive to outliers (data anomalies) or variables of incomparable scales or magnitudes. To ensure optimal parameters for the K-means model, we followed the guidelines in Tavakoli et al. (2018) by removing outliers using Interquartile Ranges (IQR). Following the removal of outliers, only 5633 customers remained after this screening. We then addressed the Long Tail problem by applying Log Transformation to recency, frequency, and monetary. After that, we used Max-Min scaling to scale the ranges of frequency and monetary. Finally, we determined the optimal number of segments (best K for K-means). To find the best parameters for the K-means model, we proceeded with the following steps:

1. Removing the outliers: Some customers show unusual behavior in almost all businesses. Therefore, we had to remove these data for high-quality data analysis and apply machine learning methods. The critical point is that we used IQR to remove the outliers. To calculate the IQR, the dataset is divided into quartiles or four rank-ordered even parts via linear interpolation. These thresholds separating these quartiles are denoted by Q1 (also called the lower quartile), Q2 (the median), and Q3 (also called the upper quartile). Thus, we used the formula (4) below to find the thresholds serving as upper and lower bounds of recency, frequency, and monetary features.

$$\begin{aligned} \mathrm{IQR} &= \mathrm{Q3} - \mathrm{Q1} \\ \text{lower bound} &= \mathrm{Q1} - 1.5 \times \mathrm{IQR} \\ \text{upper bound} &= \mathrm{Q3} + 1.5 \times \mathrm{IQR} \end{aligned} \tag{4}$$

   Accordingly, we removed the values more extreme than the lower or the upper bound in each of recency, frequency, and monetary. Finally, there were only 5633 customers left by this screening.

Table 4
Example of RFM score formal context.

R1	R2	R3	R4	R5	F1	F2	F3	F4	F5	M1	M2	M3	M4	M5
0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
0	0	1	0	0	0	0	1	0	0	0	0	0	1	0
0	0	0	0	1	0	0	0	1	0	0	0	0	0	1
0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
0	1	0	0	0	0	0	1	0	0	0	0	1	0	0

2. Solving the Long Tail problem in recency, frequency, and monetary: The histograms of RFM values can be used to support our clustering step, but it leads to problems. Thus, we applied the Log transformation to change the distribution of our data from a long-tailed shape in histograms of RFM values to a more Normal distribution.
3. Scaling the range of frequency and monetary to have appropriate data for clustering methods: Since the range of values in our frequency was utterly different from monetary values, we had to apply normalization to both frequency and monetary data. We used Max-Min Scaling, which helped us scale our data to an appropriate range for clustering.
4. Finding the best Number of Segments (best K for K-means)

The above K-means clustering process results will be compared with our work. Unlike K-means, hierarchical clustering can produce a result without assigning initial values. In this work, we also used agglomerative hierarchical clustering to compare with our approach. Usually, the hierarchical clustering method is the most intuitive way of grouping data (Dolnicar et al., 2018b). The hierarchical clustering applied Ward's method for similarity measures. However, the number of targeted clusters is a crucial parameter that should be determined before applying both the K-means and the hierarchical clustering algorithm. The Silhouette index and Davies–Boulbin index were chosen to assess the performances of clustering models. A higher Silhouette index or a lower Davies–Bouldin index is preferred for an optimal cluster number.

After the traditional clustering experimental results, the customer segmentation will be discussed by creating a descriptive profile for each customer segment with different characteristics based on the results of the proposed hierarchical concept and knowledge reported in the previous section. Finally, to design and deliver differentiated marketing strategies, the segmentation solution with customer relationship management and marketing implications are recommended for each customer group.

5. Results and discussion

This section presents the first to final steps and examines the three main steps mentioned above. These steps include (1) analyzing and preparing the data, (2) creating a knowledge structure using FCA, and (3) deploying and its business implications. Finally, we conclude with an overall discussion.

5.1. Findings from exploratory data analysis and preprocessing

After preparing the dataset, RFM values were visualized in the first step, shown in Fig. 2. This figure highlights the presence of outliers in the RFM values and indicates that the RFM values are not normally distributed. According to the data at hand, it seems that most of our clients have completed their latest transaction in the last 100 days. Furthermore, the majority of these individuals have made fewer than 100 orders overall, with most of them having spent less than £1,000.

In addition, the three variables are not on comparable scales, and the value ranges are pretty different: Recency [0, 738], Frequency [1, 12890], and Monetary [0.0, 608821.7]. For this reason, we used the Spearman correlation coefficients to evaluate the relationships between the variables in the RFM model, as this is appropriate for non-normally distributed data and is robust against outliers (Rebekić, Lončarić, Petrović, & Marić, 2015; Schober, Boer, & Schwarte, 2018). Fig. 3 shows the correlation matrices, while Fig. 4 displays the distribution of RFM values. These figures underscore the strong correlation between the variables Frequency and Monetary.

Moreover, Fig. 5 illustrates the distribution among the different RFM scores. After analyzing the data, customers with a recency and frequency value of 5 seem to have the highest monetary value (designated as Monetary value 5). This implies that individuals who have made frequent and recent purchases are more likely to be valuable customers. On the other hand, customers who have made purchases infrequently and a long time ago tend to spend the least money, with a monetary value of 1.

Finally, the results of RFM data and scores consisting of 5878 instances are provided for the next step.

Table 5
The first 11 formal concepts from all 208 concepts.

Concept#	Attributes	Support	Number of customers
1	{}	1.000000	5878
2	{M5}	0.200068	1176
3	{M4}	0.199898	1175
4	{M3}	0.200068	1176
5	{M2}	0.199898	1175
6	{M1}	0.200068	1176
7	{F5}	0.200749	1180
8	{F5, M5}	0.143756	845
9	{F5, M4}	0.043722	257
10	{F5, M3}	0.012759	75
11	{F5, M2}	0.000510	3

5.2. FCA knowledge structure results

The second step involves the FCA knowledge structure. We used the fcaR package (Cordero, Enciso, López-Rodríguez, & Mora, 2022) to perform FCA with R. The resulting FCA contains 208 concepts. The first 11 formal concepts are shown in Table 5, and some concepts forming a sublattice are shown in Fig. 6. This table and figure show a group of customers sharing the same attributes (behaviors).

In addition, we extract association rules from the concept lattice. The general form is "⟨N⟩P = [C] => ⟨N′⟩C′", where N is the number of objects satisfying the premise, P is a precondition, C is the confidence of association rule, N′ is the number of objects meeting the premise, and C′ is the conclusion. For example, the first implication in Table 6, i.e., ⟨521⟩ R5 M5 = [81%] => ⟨422⟩ F5, means that there are 521 customers with a recency score of 5, and the monetary score of 5, and 422 customers among them have frequency score of 5. Thus, the confidence of this implication is 81%.

However, the number of formal concepts, namely 208, is excessively large. The concept lattice is partially ordered according to the extents or objects by set inclusion. Therefore, we use the top-ranked concepts

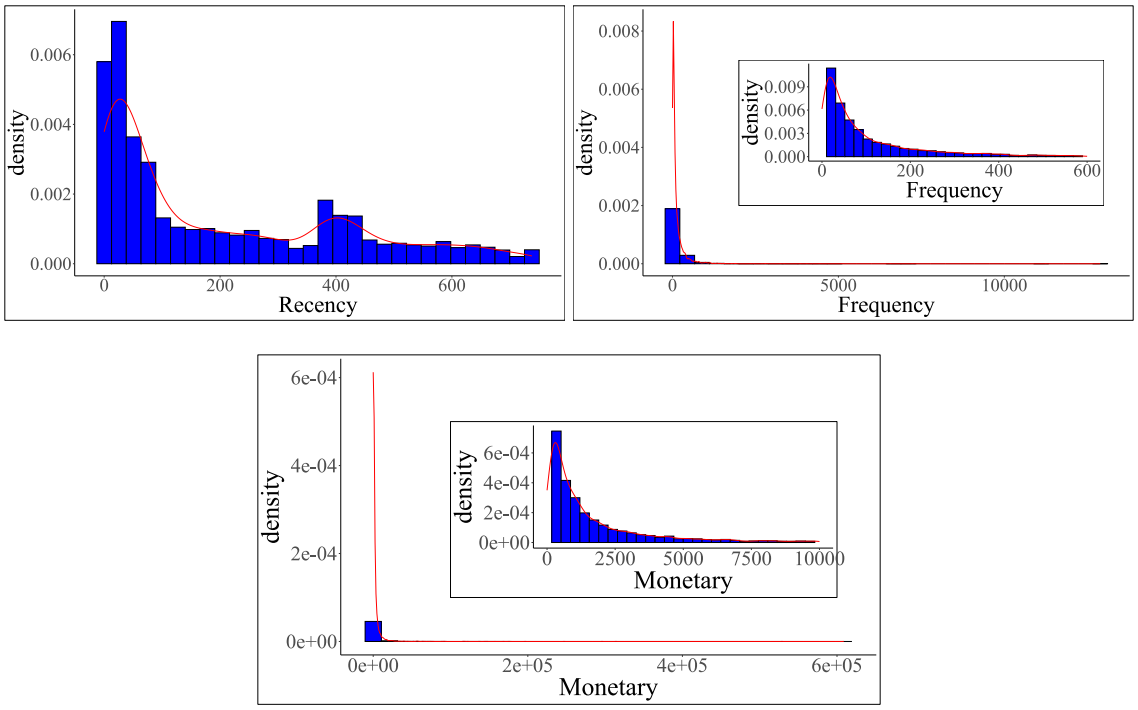


Fig. 2. Histograms of RFM values.

Table 6
List of association rules from concept lattice with confidence ≥ 60%.

No.	Rules	No.	Rules
1	<521 >R5 M5 = [81%] = ><422 >F5	10	<284 >R1 F1 = [68%] = ><192 >M1
2	<524 >R5 F5 = [81%] = ><422 >M5	11	<517 >R1 M1 = [67%]=><345 >F1
3	<450 >R1 F1 = [77%] = ><345 >M1	12	<170 >R3 M1 = [66%] = ><113 >F1
4	<329 >R4 M5 = [75%] = ><246 >F5	13	<1176 >M1 = [66%]=><775 >F1
5	<115 >R4 F1 = [73%] = ><84 >M1	14	<294 >R2 M1 = [65%] = ><192 >F1
6	<1176 >M5 = [72%]=><845 >F5	15	<131 >R4 M1 = [64%]=><84 >F1
7	<1180 >F5 = [72%]=><845 >M5	16	<64 >R5 M1 = [64%] = ><41 >F1
8	<350 >R4 F5 = [70%] = ><246 >M5	17	<197 >R3 F5 = [62%] = ><123 >M5
9	<1106 >F1 = [70%] = ><775 >M1	18	<188 >R3 F1 = [60%] = ><113 >M1

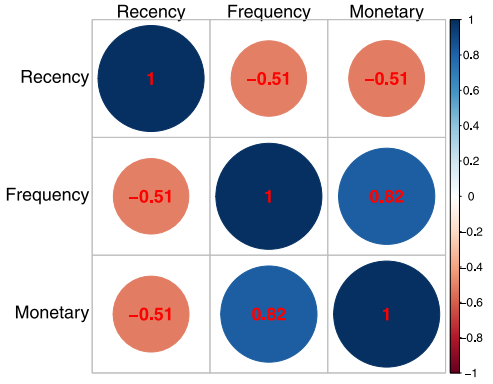


Fig. 3. Matrices of correlation with Spearman coefficients between the variables Recency, Frequency, and Monetary.

as the primary customer clusters. In this case, we include the concepts {M5}, {F5}, {M4}, {M3}, {M2}, and {M1}. Then, we considered the other concepts relating to these concepts.

The RFM scoring applies the quintile method by splitting data into five quintiles, and thus, in this case, we use only concepts whose support exceeds 0.04. We have a dataset of 5878 customers. A concept is considered if this concept consists of at least $0.04 \times 5878 = 235$ customers (20% of main clusters). Therefore, Fig. 7 depicts the customer hierarchical concept.

5.3. Business implications

For business implication deployment, the concept lattice in Fig. 6 and the hierarchical concept in Fig. 7 emphasize that the monetary score is the most significant variable in dividing customers into groups. Thus, we can simply split the customers with their monetary scores. Moreover, we also must monitor the customers having a frequency score of 5 in the highest quintile. The concept {F5} is associated with concept {F5, M5} and concept {F5, M4} belonging to the first and second quintile of monetary, respectively. These two concepts cover 19% of all customers. Accordingly, we know that the more frequently the customer purchases, the more profit the company gains. The most frequent spenders with recent purchases are customers with the highest spending amount. In contrast, customers who made their transactions long ago and less frequently contributed low monetary scores.

From Table 8, the total purchase value of concept {M5} is 77.26% of the overall purchase value. It conforms to the Pareto principle (Craft & Leake, 2002; Kim, Singh, & Winer, 2017) that "80% of sales comes from 20% of customers". Moreover, Table 8 shows that customers in concept {F5} spent 66.72% of company purchase value.

To analyze the clusters' characteristics, we refer to the information presented in Fig. 7 and Table 7.

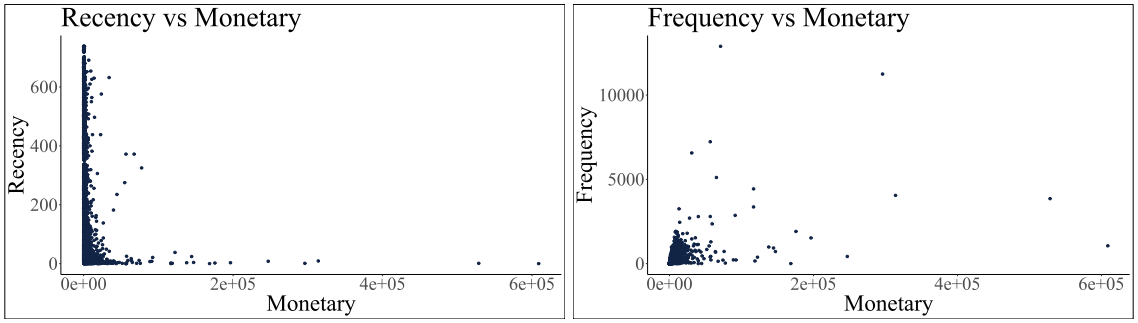


Fig. 4. Distribution of (a) Recency vs. Monetary and (b) Frequency vs. Monetary values.

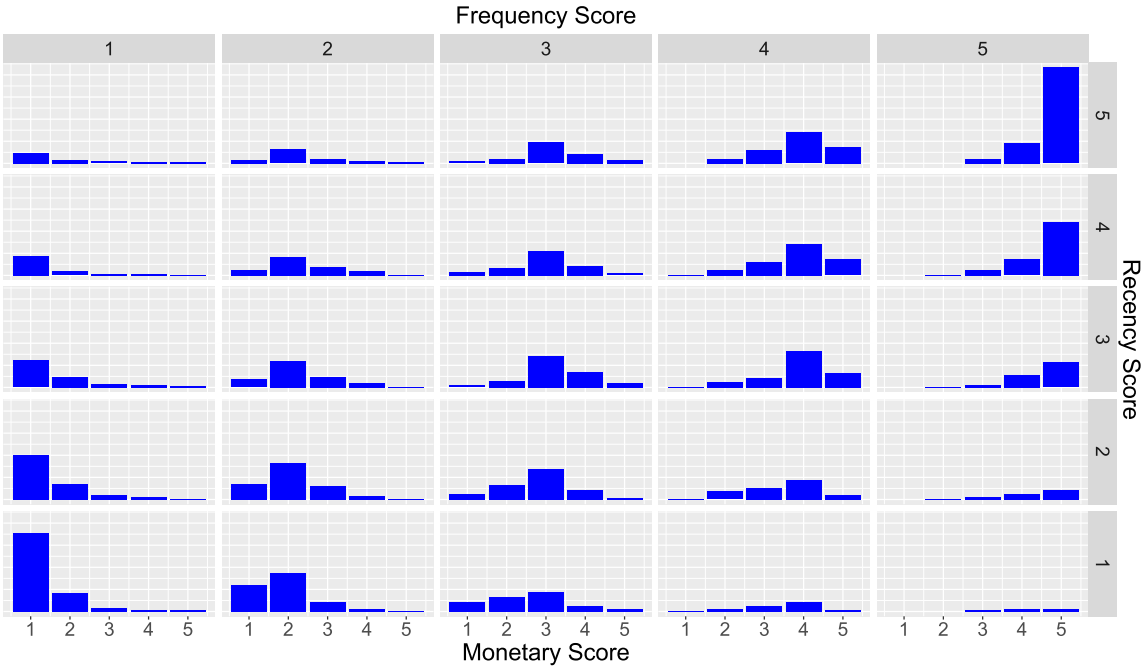


Fig. 5. RFM bar chart.

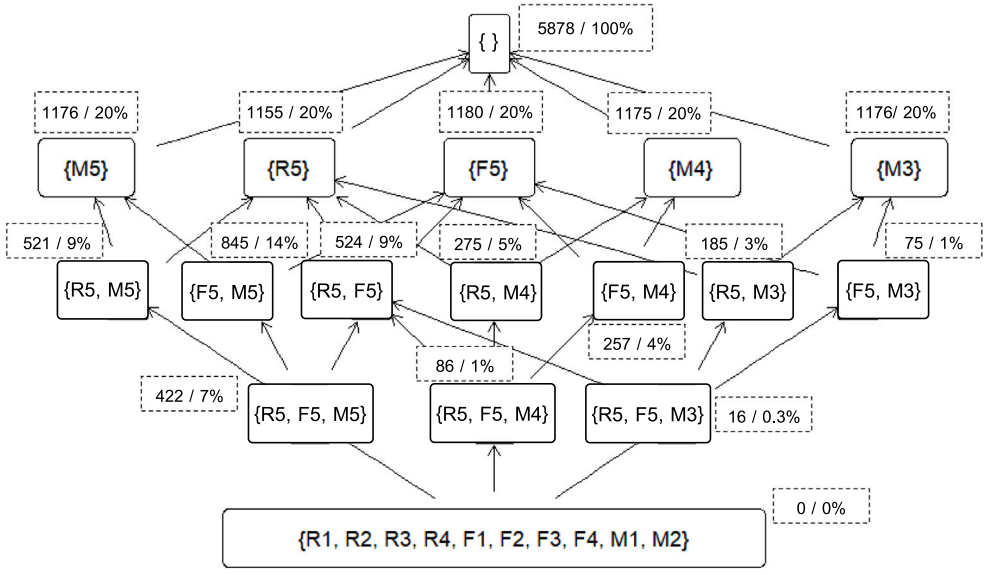


Fig. 6. Sublattice of concept 1, 2, 3, 4, 7, 8, 36, and 43 (non-specific object).

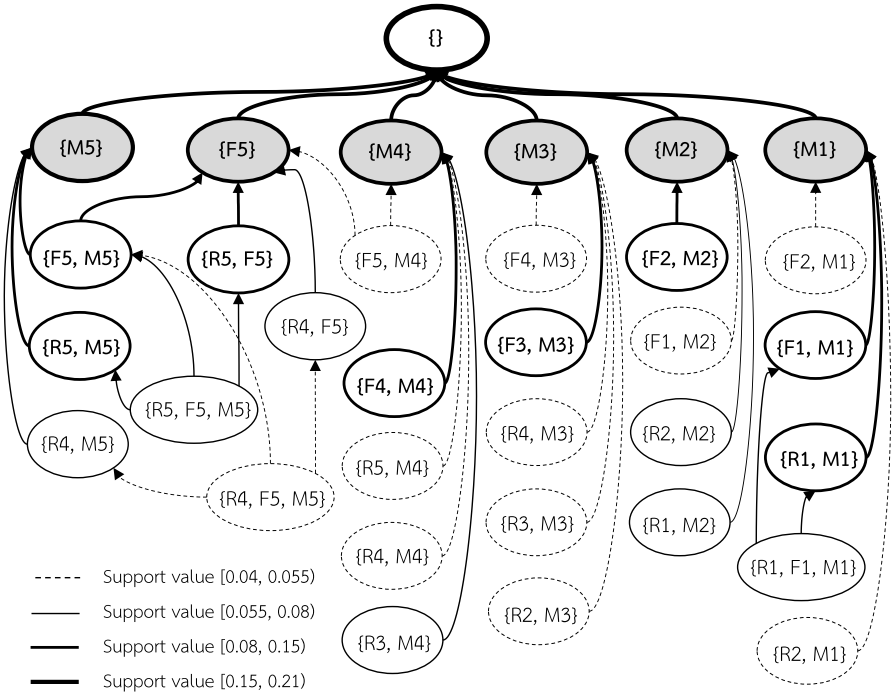


Fig. 7. RFM-based customer hierarchical concept.

Table 7
Description of the concepts related to Fig. 7.

Concepts	Support	Number of customers	Concepts	Support	Number of customers
{M5}	0.200068	1176	{R5, F5}	0.089146	524
{M4}	0.199898	1175	{R5, F5, M5}	0.071793	422
{M3}	0.200068	1176	{R4, M5}	0.055971	329
{M2}	0.199898	1175	{R4, M4}	0.049337	290
{M1}	0.200068	1176	{R4, M3}	0.043552	256
{F5}	0.200749	1180	{R4, F5}	0.059544	350
{F5, M5}	0.143756	845	{R4, F5, M5}	0.041851	246
{F5, M4}	0.043722	257	{R3, M4}	0.055291	325
{F4, M4}	0.099013	582	{R3, M3}	0.045253	266
{F4, M3}	0.042021	247	{R2, M3}	0.048826	287
{F3, M3}	0.098503	579	{R2, M2}	0.056992	335
{F2, M2}	0.105478	620	{R2, M1}	0.050017	294
{F2, M1}	0.048826	287	{R1, M2}	0.058013	341
{F1, M1}	0.131848	775	{R1, M1}	0.087955	517
{R5, M5}	0.088636	521	{R1, F1, M1}	0.058693	345
{R5, M4}	0.046785	275			

Cluster 1 ‘High Rollers’ ({M5}): This group comprises 1176 customers or 20% of the total customers. It includes customers who have spent considerable money, showcasing their ability to make substantial purchases and potentially contributing significantly to the business revenue. The 845 customers, or 72% of total customers in this cluster, have frequency scores in the top quintile (F5). Concept {F5, M5} is a sub-concept of this concept {M5}. Concept {R5, M5} is also a sub-concept of this concept {M5}. Consequently, it demonstrates that most of the customers in this group purchased products most recently. Regarding the recency scores, 521 customers, or 44% of cluster members, have recency scores in the top quintile (R5). Concept {R5, M5} is a sub-concept of this concept/cluster. Furthermore, 329 customers, or 28% of cluster members, are in the fourth quintile (R4) of recency values. For this reason, most customers in this cluster most recently bought products. Above all, the best customers with RFM scores 5-5-5, and 4-5-5 in sub-concept {R5, F5, M5} and {R4, F5, M5} include 668 customers or 11% of the whole population. These customers are exceptional in their spending habits and significantly impact the business’s success. They are regarded with prestige and recognition for their loyalty and high purchasing power. The implication is that they are highly valued customers.

Cluster 2 ‘Big Spenders’ ({M4}): This segment contains 1175 customers or 20% of all the customers. The customers in this group spent a significant amount of money but did not reach the same level as the High Rollers. Nonetheless, they have a significant impact on the business’s earnings. Moreover, the customers in this segment with frequency scores of F5 and F4 contribute 4% and 10% of the cluster members in sequence.

Cluster 3 ‘Moderate Spenders’ ({M3}): This cluster includes 1176 members or 20% of the total population. This group of customers spent an average amount of money, not as much as the High Rollers or Big Spenders, but still contributed significantly to the business’s revenue. In this particular group, 49% of customers possess frequency scores of 3 (F3) and monetary scores of 3 (M3).

Cluster 4 ‘Low Spenders’ ({M2}): This group includes 1175 members or 20% of the population. This group spent less than the previous groups, but still brought value to the business and should be given attention to maintain engagement and prevent them from leaving while encouraging spending. It is evident that the majority of customers in this particular group, accounting for 53% of cluster members, exhibit frequency scores of 2 (F2) and monetary scores of 2 (M2).

Table 8
Customer segmentation with FCA and total purchase value by cluster.

Cluster/Concept	Total purchase value	% of total purchase value	% of customer number
{M1}	193,389.6	1.09%	20%
{M2}	505,586.8	2.85%	20%
{M3}	1,063,264.1	5.99%	20%
{M4}	2,272,762.6	12.81%	20%
{M5}	13,708,426.1	77.26%	20%
{F5}	11,837,646.0	66.72%	20%

Cluster 5 'Lowest Spenders' ({M1}): This segment includes 1176 members or 20% of the population. This customer group makes the smallest financial contributions among the segments formed. Most customers in this group, with 66% of cluster members, have frequency scores of 1 (F1) and monetary scores of 1 (M1). They have the lowest potential to become loyal customers. In addition, 44% of cluster members have recency scores of 1 (R1) and monetary scores of 1 (M1). In the sub-concept of {R1, F1, M1}, 30% of all customers have RFM scores of 1-1-1, indicating that this group has the highest chance of being lost. These customers are considered the worst customers.

Cluster 6 'Frequent Buyers' ({F5}): Out of all the customers, 20% belong to this group, which has 1180 members. These customers make the most frequent purchases. Within this group, 72% of the members have frequency scores of 5 (F5) and monetary scores of 5 (M5). Additionally, 44% of these customers have recency scores of 5 (R5) and frequency scores of 5 (F5), indicating that they have made their most recent and frequent purchases of products.

We will showcase the effectiveness of our approach by presenting some examples. Fig. 7 is particularly useful for business people interested in customer retention. It highlights that customers in the groups {R3, M4}, {R3, M3}, and {R2, M3} spent a significant amount of money but were inactive for a while. These customers will likely get lost, resulting in a substantial financial loss for the business. As a result, it is necessary to take action affecting these customers. For another example, we have a considerable number of customers in the categories {F4, M4} and {F3, M3} who made multiple purchases involving significant amounts of money. By incentivizing them to make more frequent and higher-value purchases, we can potentially turn them into loyal advocates of our brand.

5.4. Discussion

To compare and discuss traditional K-means clustering and hierarchical clustering, the results of using the Silhouette index and Davies–Boulbin index to assess the performance of the traditional clustering model are shown in Fig. 8. We found that an appropriate choice is 4 – 6 clusters because they are meaningful and valuable in business implications. The number of customer segments should not be too large because it will be challenging to interpret and design marketing strategies. In addition, the choice in most of the prior works is 4 – 5 groups (Ernawati et al., 2021), and we need to compare the results with those studies. Therefore, we divided the customers into six main concepts based on FCA. The cluster profiles and results summarizing traditional K-means clustering and hierarchical clustering are provided in Figs. 9 and 10, and Tables 9 and 10.

After conducting experiments, it was discovered that outliers do not impact FCA. Therefore, it is unnecessary to eliminate outliers during data preprocessing. On the other hand, the K-means algorithm requires outlier management and data normalization as described in Section 4.3. The hierarchical clustering approach is also sensitive to outliers, which may result in imbalanced clustering. Our proposed technique leverages the RFM quintile method to generate RFM scores for all customers, which were then used to create a formal context. As a result, our approach simplifies the preprocessing stage compared to the other two methods.

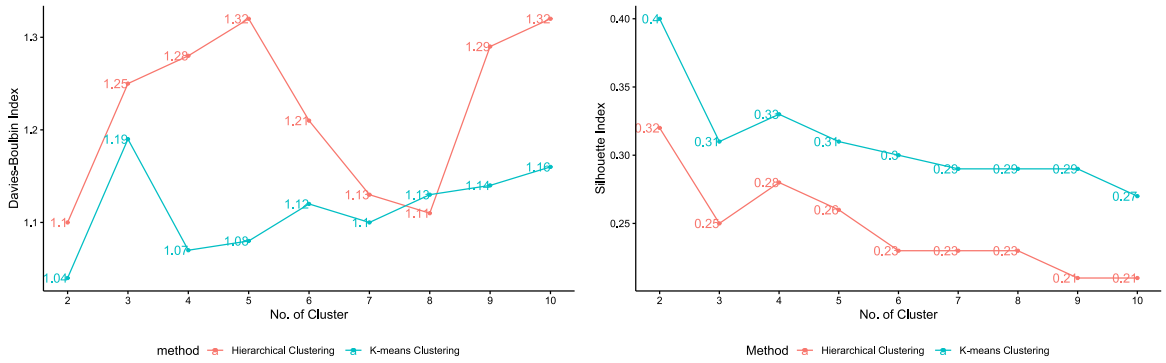
This new methodology combines RFM analysis and FCA; therefore, the relationships can be visualized in a Hasse diagram for the hierarchical concept. Moreover, this approach applies the obtained FCA knowledge structure to extract implication rules to discover knowledge relationships in customer segmentation. K-means clustering gives only completely separate customer groups without their relationships and implicit knowledge. For the hierarchical clustering technique, the results are a complete separation of customer groups with a hierarchically structured dendrogram. FCA also has a strong mathematical theory as support. Thus, the results of FCA can be assessed using mathematical theory. The results show the actual incidents that occur in the dataset. This new approach mainly creates hierarchical overlapping clusters. Therefore, a customer possibly (usually) belongs to more than one hierarchical concept or cluster.

A novel approach that combines RFM analysis and FCA has been introduced, enabling customer relationships to be visualized through a Hasse diagram for a hierarchical concept. This methodology utilizes the FCA knowledge structure to extract implication rules, facilitating the discovery of knowledge relationships in customer segmentation, in contrast to K-means clustering, which only provides distinct customer groups without their relationships and implicit knowledge. The hierarchical clustering technique also completely separates customer groups and is associated with a hierarchically structured dendrogram. This innovative methodology primarily generates hierarchical overlapping clusters, meaning a customer may belong to more than one hierarchical concept or cluster. Moreover, FCA is supported by a strong mathematical theory, allowing for rigorous evaluation of the results. The outcomes of our approach indicate the actual incidents that occur in the dataset. The results of our proposed approach are presented in a clear and understandable way with the help of different visual aids, such as the sublattice in Fig. 6 and the hierarchical concepts in Fig. 7.

From Section 5.3, the business implication deployment examples highlight how our new approach can be effectively implemented and visualized in the context of real incidents. This is in contrast to the limited representation of each customer group presented by K-means clustering and hierarchical clustering techniques, as shown in Table 9 and Table 10.

For this reason, business decision-makers have improved opportunities to deliver the customers an appropriately matched marketing strategy and increase customer retention and satisfaction. In addition, the FCA approach makes decision-makers able to clearly visualize the real incidence of customer purchase behavior. Above all, this new approach provides ease of processing and understanding of the results. The comparison summary is shown in Table 11.

To deploy in marketing strategy and recommendations, the total purchase value of the most valuable customers is almost 80 percent of total sales during the data accumulation period. On the other hand, the company gains only 1 percent of the total purchases from the worst customer group. The best customers are high-potential targets for new products. They will help the company to promote its products and brand. Therefore, the company should reward these customers and maintain close relationships to increase revenue. In addition, the company should understand why the worst customers or lost customers did not buy the products anymore. For customers with average frequency and good total sales, the company should offer promotions and recommend products for upselling. The company should provide special offers to increase customer visits, leading to more purchases.



(a) Davies-Boulbin index (b) Silhouette index

Fig. 8. Davies–Boulbin index and Silhouette index of hierarchical and K-means clustering at different cluster numbers.

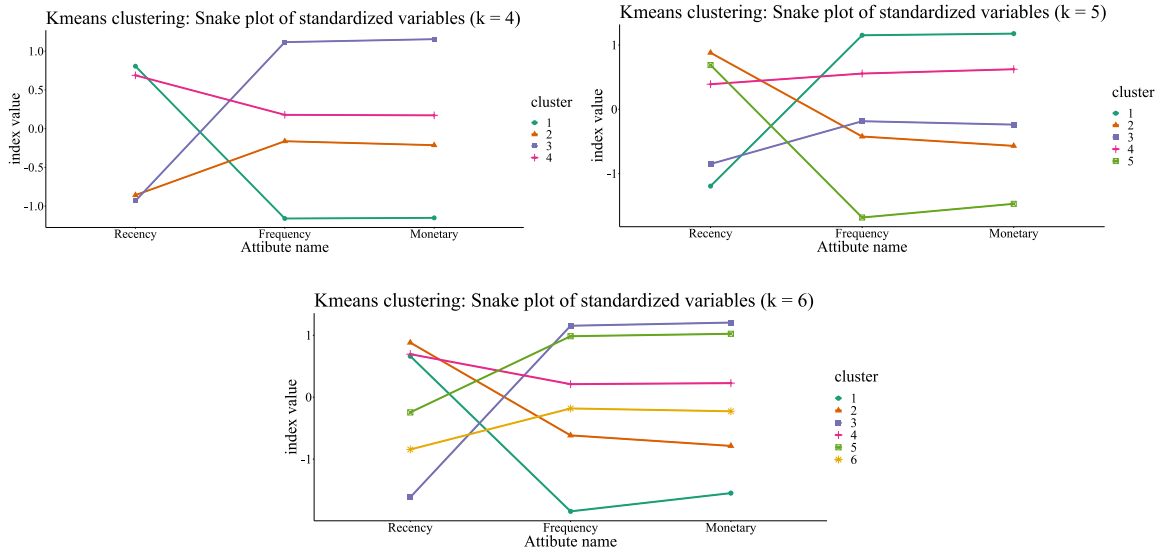


Fig. 9. Snake plots of standardized variables of K-means clustering output.

Table 9
K-means clustering output summary related to Fig. 9.

k = 4					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	High (394.26)	Low (13.66)	Low (248.20)	1472	Lost customers
2	Low (32.91)	Medium (48.10)	Medium (765.57)	1121	Potential loyalists
3	Low (37.75)	High (251.59)	High (4205.91)	1432	Champions
4	High (311.10)	Medium (75.13)	Medium (1299.57)	1608	Hibernating customers
k = 5					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	Low (20.95)	High (264.09)	High (4369.32)	1141	Champions
2	High (399.47)	Medium (30.88)	Medium Low (480.56)	1450	Hibernating customers
3	Low (32.73)	Medium (46.31)	Medium (737.69)	1061	Potential loyalists
4	Medium High (216.01)	Medium High (122.99)	Medium High (2173.25)	1263	At risk
5	High (363.34)	Low (6.31)	Low (176.06)	718	Lost customers
k = 6					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	High (360.09)	Low (5.00)	Low (168.14)	565	Lost customers
2	High (407.47)	Medium (23.20)	Medium Low (349.47)	1238	Hibernating customers
3	Low (8.91)	High (273.09)	High (4631.04)	683	Champions
4	High (308.96)	Medium High (73.98)	Medium High (1279.63)	1105	At risk
5	Medium (82.73)	High (208.63)	High (3479.89)	990	Loyal customers
6	Medium Low (32.58)	Medium (46.34)	Medium (747.53)	1052	Potential loyalists

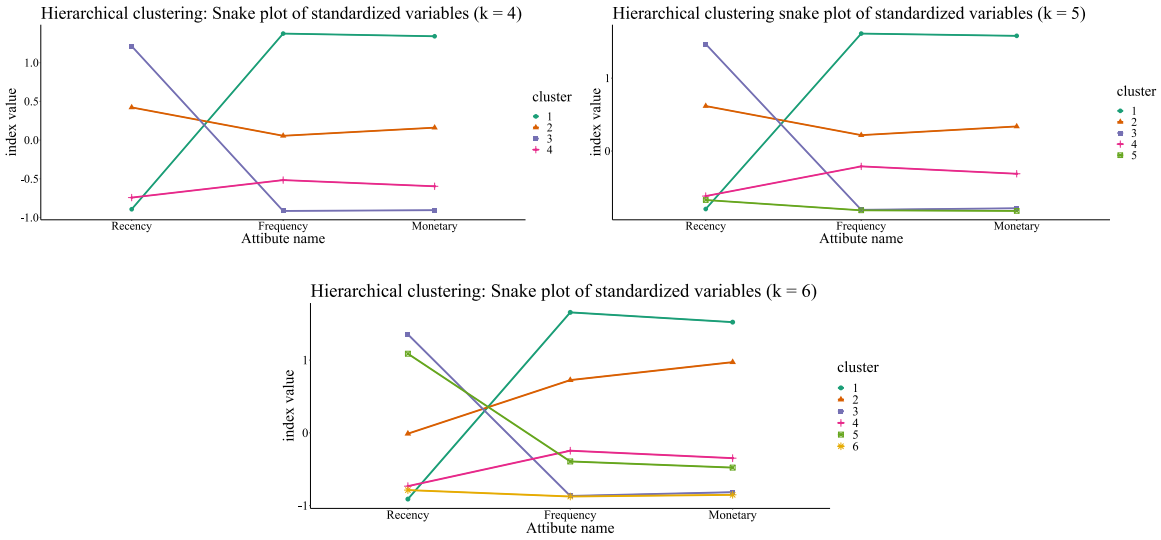


Fig. 10. Snake plots of standardized variables of hierarchical clustering output.

Table 10
Hierarchical clustering output summary related to Fig. 10.

k = 4					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	Low (14.01)	High (265.24)	High (4302.61)	913	Champions
2	Medium High (258.88)	Medium (122.63)	Medium (2223.53)	1594	Hibernating customers
3	High (406.21)	Low (17.78)	Low (346.06)	1688	Lost customers
4	Low (42.029)	Medium (60.81)	Medium (889.69)	1438	Potential loyalists
k = 5					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	Low (14.01)	High (265.24)	High (4302.61)	913	Champions
2	Medium High (258.88)	Medium (122.63)	Medium (2223.53)	1594	Hibernating customers
3	High (406.21)	Low (17.78)	Low (346.06)	1688	Lost customers
4	Low (44.75)	Medium Low (78.76)	Medium Low (1137.65)	1020	Potential loyalists
5	Low (35.39)	Low (17.03)	Low (284.63)	418	New customers
k = 6					
Cluster#	**Recency (Mean)**	**Frequency (Mean)**	**Monetary (Mean)**	**Count**	**Description**
1	Low (14.01)	High (265.24)	High (4302.61)	913	Champions
2	Medium (169.74)	Medium High (173.95)	Medium High (3371.74)	848	Loyal customers
3	High (406.21)	Low (17.78)	Low (346.06)	1688	Lost customers
4	Low (44.75)	Medium Low (78.76)	Medium Low (1137.65)	1020	Potential loyalists
5	High (360.20)	Medium Low (64.29)	Medium Low (918.33)	746	Hibernating customers
6	Low (35.39)	Low (17.03)	Low (284.63)	418	New customers

Table 11
FCA-based, K-means and hierarchical clustering comparison.

Issues	FCA	K-means	Hierarchical
Preprocessing	Less	More	More
Outlier	Not sensitive to outliers	Sensitive to outliers and outlier management needed	Sensitive to outliers leading to imbalanced clustering results
Data normalization	RFM quintile method used	Data normalization needed	Data normalization needed
Relationship information among segments	Relationship demonstration with Hasse diagram and hierarchical concept	Complete separation of customer groups	Complete separation of customer groups with hierarchically structured dendrogram
Mathematical theory support	Strong support	Cluster centroid initialization sensitivity	Imbalanced clustering results.
Implicit Knowledge	Included	Not included	Not included
Clustering type	Soft clustering	Hard clustering	Hard clustering
Clustering result	Hierarchical overlapping clusters	Unstructured non-overlapping clusters	Hierarchical structured non-overlapping clusters
Visualization	Easy and various	More difficult	More difficult

6. Conclusion

This research proposes a new effective clustering algorithm using the advantages of FCA to build a knowledge representation. This model combines the RFM model with FCA. Thus, the construction contains both implicit and explicit knowledge. Explicit knowledge shows cluster visualized information represented in the hierarchical structure model, while implicit knowledge is embedded in the structure with its implication properties. Thus, the knowledge structure from FCA reveals relationships among data points and easily understood results. Afterwards, the proposed model was compared with K-means clustering and hierarchical clustering using the online retail II dataset from the UCI Machine Learning Repository. In conclusion, the proposed method provides enough and appropriate information for marketers to perceive the value of the clustering results for creating practical marketing strategies in real-world business. This approach offers marketers both customer segmentation and relationships in customer data simultaneously. The advantage of the RFM model is the use of a very small number of variables (only three variables) to reduce the complexity of the model. However, we suggest adding new variables to the RFM model to increase the accuracy and gain more information in future studies. In addition, we will modify and improve this model by representing RFM values in a non-binary formal context in future studies using fuzzy sets. Then, FCA will be considered and compared with the results from this alternative approach. Therefore, we recommend employing it in the context of other industries in future research. This proposed method can also be applied to businesses other than online retail because the characteristics of the retail dataset are similar to our experiment. In addition, visualizations play a crucial role in achieving success in data-driven decision-making with this approach regarding information systems and application development. We suggest the application should provide interactive visualization options that enable users to add or choose features and adjust for better visualization effortlessly.

CRediT authorship contribution statement

Chongkolnee Rungruang: Methodology, Software, Writing – original draft, Visualization, Data curation, Validation. **Pakwan Riyapan:** Supervision, Validation, Writing – review & editing. **Arthit Intarasit:** Supervision, Validation, Writing – review & editing. **Khanchit Chuarkham:** Supervision, Validation, Writing – review & editing. **Jirapond Muangprathub:** Conceptualization, Methodology, Software, Supervision, Validation, Writing – original draft.

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.

Data availability

We use dataset from UCI repository at: https://archive.ics.uci.edu.

Acknowledgments

The authors are deeply grateful to the Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Thailand. This research was financially supported by the Research and Development Office, Prince of Songkla University, Thailand, under grant No. SIT6502069S. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation. The authors also gratefully acknowledge the helpful check of the English language by Assoc. Prof. Dr. Seppo Karrila.

References

Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, http://dx.doi.org/10.1007/s10257-023-00640-4.

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785–1792. http://dx.doi.org/10.1016/j.jksuci.2019.12.011, URL https://www.sciencedirect.com/science/article/pii/S1319157819309802.

Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2018). Customer segmentation in e-commerce: Applications to the cashback business model. *Journal of Business Research*, *88*, 407–414. http://dx.doi.org/10.1016/j.jbusres.2017.11.047, URL https://www.sciencedirect.com/science/article/pii/S0148296317304939.

Benavent, X., Castellanos, A., de Ves, E., García-Serrano, A., & Cigarrán, J. (2019). FCA-based knowledge representation and local generalized linear models to address relevance and diversity in diverse social images. *Future Generation Computer Systems*, *100*, 250–265. http://dx.doi.org/10.1016/j.future.2019.05.029, URL https://www.sciencedirect.com/science/article/pii/S0167739X18307271.

Castellanos, A., Cigarrán, J., & García-Serrano, A. (2017). Formal concept analysis for topic detection: A clustering quality experimental analysis. *Information Systems*, *66*, 24–42. http://dx.doi.org/10.1016/j.is.2017.01.008, URL https://www.sciencedirect.com/science/article/pii/S030643791730087X.

Chang, H., & Tsay, S. (2004). Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation. *Journal of Information Management*, *11*(4), 161–203, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-79951576359&partnerID=40&md5=b5597c45dbdb9db7076022549b658424.

Chen, D., Guo, K., & Li, B. (2019). Predicting customer profitability dynamically over time: An experimental comparative study. In *Iberoamerican congress on pattern recognition* (pp. 174–183). Springer.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197–208. http://dx.doi.org/10.1057/dbm.2012.17.

Chen, H., Zhang, L., Chu, X., & Yan, B. (2019). Smartphone customer segmentation based on the usage pattern. *Advanced Engineering Informatics*, *42*, Article 101000. http://dx.doi.org/10.1016/j.aei.2019.101000, URL https://www.sciencedirect.com/science/article/pii/S1474034619305737.

Chen, Q., Zhang, M., & Zhao, X. (2017). Analysing customer behaviour in mobile app usage. *Industrial Management & Data Systems*.

Choi, H., Choi, E. K., Yoon, B., & Joung, H. W. (2020). Understanding food truck customers: Selection attributes and customer segmentation. *International Journal of Hospitality Management*, *90*, Article 102647. http://dx.doi.org/10.1016/j.ijhm.2020.102647, URL https://www.sciencedirect.com/science/article/pii/S0278431920301997.

Chorianopoulos, A. (2016). *Effective CRM using predictive analytics*. John Wiley & Sons.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, *33*(10), 1251–1257. http://dx.doi.org/10.1016/j.jksuci.2018.09.004.

Cordero, P., Enciso, M., López-Rodríguez, D., & Mora, Á. (2022). fcaR, formal concept analysis with R. *R J.*, *14*, 341–361.

Craft, R. C., & Leake, C. (2002). The Pareto principle in organizational decision making. *Management Decision*, *40*(8), 729–733. http://dx.doi.org/10.1108/00251740210437699.

Dedi, Dzulhaq, M. I., Sari, K. W., Ramdhan, S., Tullah, R., & Sutarman (2019). Customer segmentation based on RFM value using K-means algorithm. In *2019 fourth international conference on informatics and computing* (pp. 1–7). http://dx.doi.org/10.1109/ICIC47613.2019.8985726.

Deng, Y., & Gao, Q. (2020). A study on e-commerce customer segmentation management based on improved K-means algorithm. *Information Systems and e-Business Management*, *18*(4), 497–510. http://dx.doi.org/10.1007/s10257-018-0381-3.

Dolnicar, S., Grün, B., & Leisch, F. (2018a). *Market segmentation analysis: Understanding it, doing it, and making it useful*. Springer Nature.

Dolnicar, S., Grün, B., & Leisch, F. (2018b). Step 5: Extracting segments. In *Market segmentation analysis* (pp. 75–181). Springer.

Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. *Journal of Physics: Conference Series*, *1869*(1), http://dx.doi.org/10.1088/1742-6596/1869/1/012085.

Fränti, P., & Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. *Applied Intelligence*, *48*(12), 4743–4759. http://dx.doi.org/10.1007/s10489-018-1238-7.

Frasquet, M., Ieva, M., & Ziliani, C. (2021). Online channel adoption in supermarket retailing. *Journal of Retailing and Consumer Services*, *59*, Article 102374. http://dx.doi.org/10.1016/j.jretconser.2020.102374.

Gagolewski, M., Bartoszuk, M., & Cena, A. (2021). Are cluster validity measures (in) valid? *Information Sciences*, *581*, 620–636. http://dx.doi.org/10.1016/j.ins.2021.10.004, URL https://www.sciencedirect.com/science/article/pii/S0020025521010082.

Ganter, B., & Wille, R. (2012). *Formal concept analysis: mathematical foundations*. Springer Science & Business Media.

Gupta, S., Justy, T., Kamboj, S., Kumar, A., & Kristoffersen, E. (2021). Big data and firm marketing performance: Findings from knowledge-based view. *Technological Forecasting and Social Change*, *171*, Article 120986. http://dx.doi.org/10.1016/j.techfore.2021.120986, URL https://www.sciencedirect.com/science/article/pii/S0040162521004182.

Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications*, *40*(16), 6491–6507. http://dx.doi.org/10.1016/j.eswa.2013.05.052, URL https://www.sciencedirect.com/science/article/pii/S0957417413003503.

Hosseini, M., Abdolvand, N., & Harandi, S. R. (2022). Two-dimensional analysis of customer behavior in traditional and electronic banking. *Digital Business*, *2*(2), Article 100030. http://dx.doi.org/10.1016/j.digbus.2022.100030, URL https://www.sciencedirect.com/science/article/pii/S2666954422000102.

Khalili-Damghani, K., Abdi, F., & Abolmakarem, S. (2018). Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing*, *73*, 816–828. http://dx.doi.org/10.1016/j.asoc.2018.09.001, URL https://www.sciencedirect.com/science/article/pii/S1568494618305052.

Kim, B. J., Singh, V., & Winer, R. S. (2017). The Pareto rule for frequently purchased packaged goods: an empirical generalization. *Marketing Letters*, *28*(4), 491–507. http://dx.doi.org/10.1007/s11002-017-9442-5.

Kuznetsov, S., & Makhalova, T. (2018). On interestingness measures of formal concepts. *Information Sciences*, *442–443*, 202–219. http://dx.doi.org/10.1016/j.ins.2018.02.032, URL https://www.sciencedirect.com/science/article/pii/S0020025516315791.

Kuznetsov, S. O., & Poelmans, J. (2013). Knowledge representation and processing with formal concept analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(3), 200–215.

Lei, Y., Yan, Y., Han, Y., & Jiang, F. (2018). The hierarchies of multivalued attribute domains and corresponding applications in data mining. *Wireless Communications and Mobile Computing*, *2018*.

Marchetto, A. (2005). A concerns-based metrics suite for web applications. *INFOCOMP Journal of Computer Science*, *4*(3), 11–22.

Meng, X., Liu, M., Wu, J., Zhou, H., Xu, F., & Wu, Q. (2020). Hierarchical clustering on metric lattice. *International Journal of Intelligent Information and Database Systems*, *13*(1), 1–16.

Monalisa, S., Nadya, P., & Novita, R. (2019). Analysis for customer lifetime value categorization with RFM model. *Procedia Computer Science*, *161*, 834–840. http://dx.doi.org/10.1016/j.procs.2019.11.190, URL https://www.sciencedirect.com/science/article/pii/S1877050919319015 The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

Munusamy, S., & Murugesan, P. (2020). Modified dynamic fuzzy c-means clustering algorithm – Application in dynamic customer segmentation. *Applied Intelligence*, *50*(6), 1922–1942. http://dx.doi.org/10.1007/s10489-019-01626-x.

Nakano, S., & Kondo, F. N. (2018). Customer segmentation with purchase channels and media touchpoints using single source panel data. *Journal of Retailing and Consumer Services*, *41*, 142–152. http://dx.doi.org/10.1016/j.jretconser.2017.11.012.

Nandapala, E., & Jayasena, K. (2020). The practical approach in customers segmentation by using the K-means algorithm. In *2020 IEEE 15th international conference on industrial and information systems* (pp. 344–349). http://dx.doi.org/10.1109/ICIIS51140.2020.9342639.

Nguyen, T. B., Tran, C. D., Quan, T. T., Nguyen, M. H., & Le, T. A. (2019). MarCHGen: A framework for generating a malware concept hierarchy. *Expert Systems*, *36*(5), Article e12445.

Noori, B. (2015). An analysis of mobile banking user behavior using customer segmentation. *International Journal of Global Business*, *8*(2).

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence & Planning*, *35*(4), 544–559. http://dx.doi.org/10.1108/MIP-11-2016-0210.

Poelmans, J., Ignatov, D. I., Kuznetsov, S. O., & Dedene, G. (2013). Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications*, *40*(16), 6538–6560. http://dx.doi.org/10.1016/j.eswa.2013.05.009.

Rahim, M. A., Mushafiq, M., Khan, S., & Arain, Z. A. (2021). RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, *61*, Article 102566. http://dx.doi.org/10.1016/j.jretconser.2021.102566, URL https://www.sciencedirect.com/science/article/pii/S0969698921001326.

Ravi, K., Ravi, V., & Prasad, P. S. R. K. (2017). Fuzzy formal concept analysis based opinion mining for CRM in financial services. *Applied Soft Computing*, *60*, 786–807. http://dx.doi.org/10.1016/j.asoc.2017.05.028, URL https://www.sciencedirect.com/science/article/pii/S1568494617302910.

Rebekić, A., Lončarić, Z., Petrović, S., & Marić, S. (2015). Pearson's or Spearman's correlation coefficient-which one to use? *Poljoprivreda*, *21*(2), 47–54.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768. http://dx.doi.org/10.1213/ANE.0000000000002864.

Shokouhyar, S., Shokoohyar, S., & Safari, S. (2020). Research on the influence of after-sales service quality factors on customer satisfaction. *Journal of Retailing and Consumer Services*, *56*, Article 102139. http://dx.doi.org/10.1016/j.jretconser.2020.102139, URL https://www.sciencedirect.com/science/article/pii/S0969698919313311.

Singh, J., & Mittal, M. (2021). Customer's purchase prediction using customer segmentation approach for clustering of categorical data. *Management and Production Engineering Review*, *12*.

Singh, S., & Srivastava, S. (2020). Review of clustering techniques in control system: Review of clustering techniques in control system. *Procedia Computer Science*, *173*, 272–280. http://dx.doi.org/10.1016/j.procs.2020.06.032, URL https://www.sciencedirect.com/science/article/pii/S1877050920315362 International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020.

Sokol, O., & Holỳ, V. (2021). The role of shopping mission in retail customer segmentation. *International Journal of Market Research*, *63*(4), 454–470.

Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018). Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: A case study. In *2018 IEEE 15th international conference on E-business engineering* (pp. 119–126). http://dx.doi.org/10.1109/ICEBE.2018.00027.

Tsiptsis, K. K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.

Wang, S. C., Tsai, Y. T., & Ciou, Y. S. (2020). A hybrid big data analytical approach for analyzing customer patterns through an integrated supply chain network. *Journal of Industrial Information Integration*, *20*, Article 100177. http://dx.doi.org/10.1016/j.jii.2020.100177, URL https://www.sciencedirect.com/science/article/pii/S2452414X20300522.

Wille, R. (2009). Restructuring lattice theory: an approach based on hierarchies of concepts. In *International conference on formal concept analysis* (pp. 314–339). Springer.

Wu, T., & Liu, X. (2020). A dynamic interval type-2 fuzzy customer segmentation model and its application in E-commerce. *Applied Soft Computing*, *94*, Article 106366. http://dx.doi.org/10.1016/j.asoc.2020.106366, URL https://www.sciencedirect.com/science/article/pii/S1568494620303069.

Yeh, I.-C., Yang, K. J., & Ting, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, *36*(3, Part 2), 5866–5871. http://dx.doi.org/10.1016/j.eswa.2008.07.018, URL https://www.sciencedirect.com/science/article/pii/S0957417408004508.

Zeybek, H. (2018). Customer segmentation strategy for rail freight market: The case of Turkish State Railways. *Research in Transportation Business & Management*, *28*, 45–53. http://dx.doi.org/10.1016/j.rtbm.2018.10.003, URL https://www.sciencedirect.com/science/article/pii/S2210539516301596.

Zhang, Z., Zhao, J., & Yan, X. (2018). A web page clustering method based on formal concept analysis. *Information*, *9*(9), http://dx.doi.org/10.3390/info9090228, URL https://www.mdpi.com/2078-2489/9/9/228.

Zhou, J., Wei, J., & Xu, B. (2021). Customer segmentation by web content mining. *Journal of Retailing and Consumer Services*, *61*, Article 102588. http://dx.doi.org/10.1016/j.jretconser.2021.102588, URL https://www.sciencedirect.com/science/article/pii/S0969698921001545.