System Design for Extensive Exploratory Causal Association Odds Ratio Rule Analysis in a Temporary Database Environment

Ting-Yan Lin[1], Meng-Feng Tsai[1], Chi-Sheng Huang[2], Liang-Han Lin[1], Jiun-Yi Tsai[1]*

[1]Department of Computer Science and Information Engineering, National Central University, Taiwan
[2]Department of Computer Science and Information Engineering,
National Taichung University of Science and Technology, Taiwan
tingyan@g.ncu.edu.tw, mftsai@csie.ncu.edu.tw, vcshuang@nutc.edu.tw,
109522083@cc.ncu.edu.tw, 109522069@cc.ncu.edu.tw

Abstract

This study will undertake extensive exploratory analysis within a temporary database environment (or data lake), positioned between raw datasets, often sourced from relational databases, and a multidimensional data warehouse supporting diverse analytical perspectives. The primary objective is to generate causal association odds ratio rules that go beyond traditional association rules, aiming to enhance decision-support systems with insights into underlying causal relationships. Such a design allows for a focused interpretability on causal factors influenced by computation across highly complex scenarios, offering effective dimensional references for constructing multidimensional data warehouses and aiding in the selection of valuable validation analyses.

Keywords: Causal association odds ratio rule mining, Exploratory analysis, Staging database, Data warehouse, Data lake

1 Introduction

Recent advancements in various fields, such as e-commerce, social networks, AR/VR, and video surveillance, have greatly improved efficiency and convenience, driven by information and communication technologies. These innovations have led to a rapid increase in the quantity, form, and interrelationships of electronic data, surpassing expectations [1-2]. This growth has fueled the development of big data, cloud services, data science, and machine learning, both technically and academically [3-4]. Technological advancements typically prioritize domains with clear objectives, large data volumes, and manageable interdependencies, where relationships are not overwhelmingly complex.

From the perspective of traditional decision support systems, the potential for diverse analytical topics has grown exponentially, potentially outpacing the growth in data volume [5-6]. This is due to the rapid expansion of possible areas of focus, which grow in proportion (or faster) to the combinations and permutations of the data. The increase in these combinations is exponential relative to the data size, making analysis more complex [7].

Traditional association rules primarily capture correlations without addressing causality, limiting their ability to guide decision-making. In contrast, causal association odds ratio rules reveal causal relationships, provide insights that better support decision-making. For example, while a traditional association rule might suggest that "gender = male" correlates with "salary = low," causal association odds ratio analysis can reveal whether gender is a causal factor for salary differences.

Effectively focusing the scope of analysis has become a critical challenge in both practical applications and research. This study takes a unique approach by focusing on exploratory analysis within a staging database environment, akin to a "data lake" [8]. Unlike fully structured environments or raw datasets, this staging space allows for the transposition and collection of denormalized data. While not fully structured, this environment highlights the challenge of effectively focusing analysis. The goal is to apply causal association odds ratio rule analysis, which better aligns with decision support needs, to identify and concentrate on the most relevant features. This process ultimately supports the development of a foundational data warehouse environment for practical analysis.

Many studies limit their scope to manageable spaces to ensure falsifiability and repeatability, often resulting in specialized solutions that work well in specific contexts but lack broader applicability. This study, however, aims to provide a generalized solution through exploratory analysis in expansive intermediate data spaces. The intent is to enhance the applicability of scientific analysis, offering a novel approach distinct from mainstream methodologies in the field.

2 Related Work

2.1 Staging Database and Data Lake

A staging database acts as an intermediary between the data source and target, facilitating processes like data cleaning, transformation, integration, and reformatting. It temporarily stores raw data before it's loaded into a final data warehouse or analytical system.

*Corresponding Author: Chi-Sheng Huang; Email: vcshuang@nutc.edu.tw
DOI: https://doi.org/10.70003/160792642026032702005

A data lake, on the other hand, is a central repository that stores raw data in its original format [9-10], without predefined structures. Data is only processed when needed for analysis. While similar to a staging database, a data lake operates on a larger scale, housing diverse data that can be processed according to specific analytical requirements, as illustrated in Figure 1.

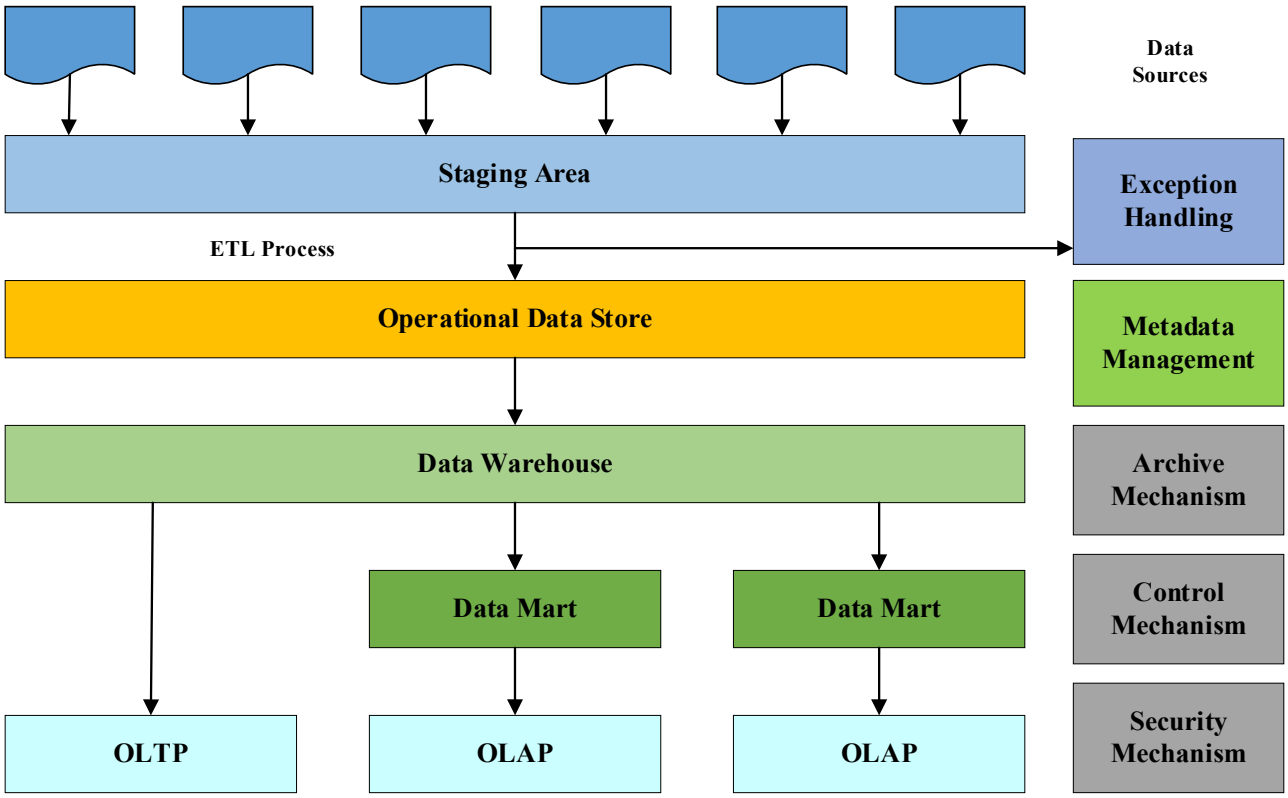

Figure 1. Staging database structure

2.2 Association Rules and Causal Association Odds Ratio Rules

Association rule analysis, originally developed for business applications, identifies patterns in sales data to understand relationships between items. This approach has been widely used in market basket analysis, customer behavior analysis, and other business strategies. For instance, discovering that customers who buy beer also purchase diapers reveals a correlation that businesses can leverage for bundling or promotional activities. Today, such techniques have found applications across diverse fields [11].

Causal association odds ratio rules are widely used in fields like medicine, social sciences, and statistics to identify cause-and-effect relationships, such as the factors driving diseases or the impacts of climate change [12-13]. Unlike association rules, which only measure correlations, causal rules focus on the strength and direction of cause-and-effect relationships, reflecting the principle that "correlation does not imply causation" [14-15].

For instance, an association rule might suggest that increased ice cream sales are linked to higher drowning rates. However, causal analysis identifies "temperature" as the underlying factor influencing both behaviors. Higher temperatures lead to more ice cream sales and more swimmers, indirectly increasing drownings, as illustrated in Figure 2. Causal rules thus clarify true relationships and uncover actual causes that association rules alone cannot [16-17].

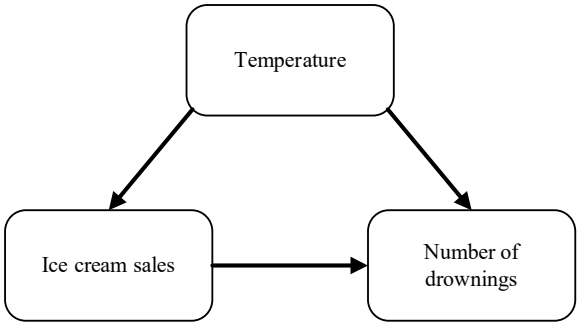

Figure 2. The relationship between ice cream sales, number of drownings, and temperature

Before making decisions, it is essential to assess their potential impacts and rely on data that clearly demonstrates expected outcomes. A scientific approach is needed to establish causal relationships, showing not just correlations but how changes in one variable drive changes in others. Compared to association rules, causal association odds ratio rules provide greater value by focusing on cause-and-effect dynamics.

However, current studies often use association rules with high support and confidence to approximate causal rules, leading to inconsistencies that limit their utility. Identifying data combinations that truly reflect causal relationships can offer a stronger foundation for decision-making, enabling more accurate evaluations and providing robust support for effective decisions.

2.3 Exploratory Analysis

Exploratory Data Analysis (EDA) [18], introduced by John Tukey, involves examining a dataset to uncover its key characteristics, often using visualizations. Unlike Confirmatory Factor Analysis (CFA) [19], which tests predefined hypotheses or inferences, EDA focuses on observation and exploration, enabling the generation of new insights and hypotheses from the data.

3 Method

3.1 The Experiment Flow

This study adopts a single-machine architecture with a data lake as the data source. Unlike a data warehouse, which stores pre-processed, high-quality data in a structured, multidimensional format, a data lake serves as a central repository for raw, unprocessed data without predefined purposes. In a warehouse, data is often transformed across various granularities, such as from department to faculty level or from semester grades to annual grades.

The study aims to mine causal association odds ratio rules directly from the data lake and, in the future, apply these rules to a multidimensional data warehouse for extraction at different levels of granularity. The system process, illustrated in Figure 3, begins by mining multivariable association rules from the data lake to identify candidate causal association odds ratio rules. This step significantly narrows the scope for further mining. Through cohort studies and the creation of a fair dataset from the data lake, the odds ratio is calculated to verify if the candidate rule represents a true causal relationship. Among various statistical methods for verification, this study focuses on using the odds ratio as the primary indicator to evaluate the correctness of causal association odds ratio rules.

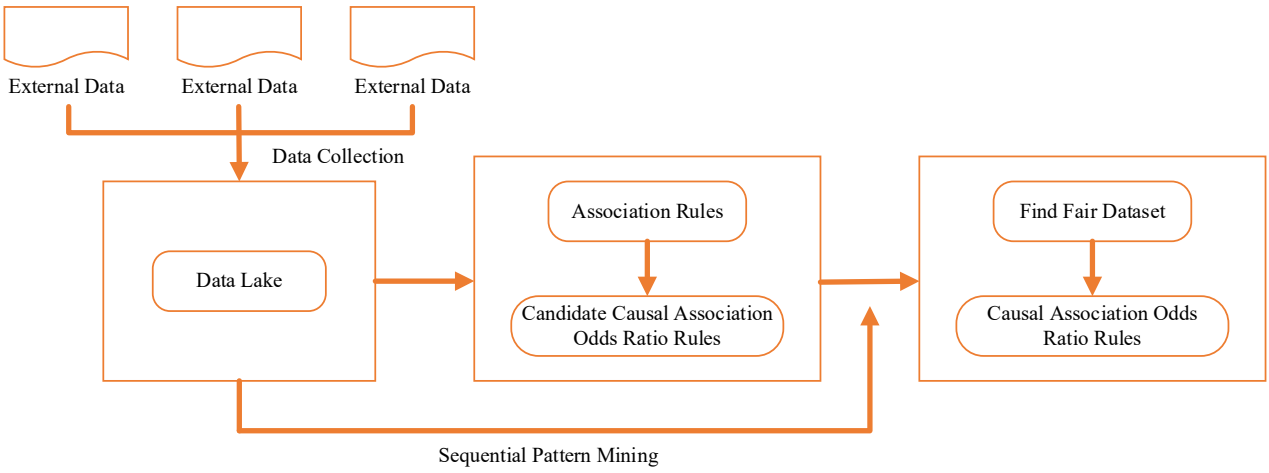

Figure 3. System flow diagram

Table. 1. Dataset field table

Gender	Age	Postal code	Marital status
Business District	Membership Level	Activity Expenditure	Category A Product Expenditure
Category B Product Expenditure	Category C Product Expenditure	Category D Product Expenditure	Category E Product Expenditure
Category F Product Expenditure	Category G Product Expenditure	Category H Product Expenditure	Category I Product Expenditure
Category J Product Expenditure	Category K Product Expenditure	Category L Product Expenditure	Category M Product Expenditure
Category N Product Expenditure	Category O Product Expenditure	Category P Product Expenditure	Category Q Product Expenditure
Category R Product Expenditure	Category S Product Expenditure	Total Expenditure	Total Purchase Frequency

3.2 Data Description and Data Preprocessing

3.2.1 Data Description

This study uses customer transaction data from two cosmetic brands, Brand A and Brand B, with around 50,000 records and 28 fields per record. The field names are listed in Table 1. The dataset includes customer information, store data, and transaction records. Due to the significant price differences between the two brands and the frequent product updates, which may alter customer spending habits, the dataset is divided into 8 subsets based on the brand and the transaction year. Each subset represents the total transaction data of a specific customer for one brand in a given year.

3.2.2 Data Preprocessing

Extract relevant data tables from the data lake and perform initial normalization or numeric transformation, such as converting raw grades to PR values to ensure fairness. For example, grading discrepancies between different classes should be addressed. Additionally, join the data tables based on relevant associations to facilitate subsequent analysis.

3.3 Causal Association Odds Ratio Rule Mining

3.3.1 Causal Association Odds Ratio Rule Mining Process

The goal of mining causal association odds ratio rules is to support decision-making, identify root causes, and improve future plans [20-21]. Given the vast number of possible data combinations, evaluating all subsets for causal relationships would be resource-intensive. Therefore, it is essential to first identify candidate causal rules to filter out irrelevant combinations, significantly reducing the time and resources needed for analysis.

Using sequential pattern mining techniques [22], highly correlated data combinations are identified by calculating support [23], confidence, and lift [24]. Confidence, in particular, is key for detecting causal association odds ratio rules. By removing one cause from combinations and calculating the decrease in confidence, potential causal rules can be identified. Below are examples of this process:

Example 1: {Did not join a club, Information and Electrical Engineering College, Bachelor's degree, No punishment} → {Currently unemployed}

Example 1 presents a multivariable association rule, showing that students who did not join a club, belong to the Information and Electrical Engineering College, hold a Bachelor's degree, and have no recorded punishment are highly correlated with being unemployed one year after graduation. The confidence is 87%, indicating that nearly 90% of students with these attributes will be unemployed a year after graduation.

Example 2: {Did not join a club, Information and Electrical Engineering College, Bachelor's degree, No punishment} → {Currently unemployed}

Example 2 is similar to Example 1, but with the "Bachelor's degree" variable removed. The confidence drops to 42%, indicating that without a Bachelor's degree, the likelihood of unemployment one year after graduation remains around 40%, a 50% decrease in confidence. This suggests that the "Bachelor's degree" is a strong candidate for a causal association odds ratio rule influencing employment status after graduation.

Example 3: {Did not join a club, Information and Electrical Engineering College, Bachelor's degree, No punishment} → {Currently unemployed}

Example 3 removes two attributes ("Did not join a club" and "Bachelor's degree") and calculates the resulting decrease in confidence. This helps assess how the combination of these two factors affects the probability of unemployment one year after graduation. It highlights the potential of exploring causal association odds ratio rule candidates through multiple factors, compensating for the limitations of analyzing single variables. By considering multiple variables, causal association odds ratio rules that might otherwise be overlooked can be uncovered.

This method allows for the exploration of multivariable causal association odds ratio rules, helping to identify relationships between factors that influence outcomes in more complex ways than single-variable analysis.

3.3.2 Judging the Credibility of Causal Association Odds Ratio Rules

In the previous section, analysts filtered causal association odds ratio rule candidates based on the selected confidence decrease ratio. The next step is to assess the credibility of these rules. This study employs a "manipulation" approach, similar to a retrospective cohort study, to identify and verify causal association odds ratio rules through data manipulation.

A cohort study, or generational study, is a longitudinal research method widely used in fields such as medicine, social sciences, and ecology [25]. It observes a population over time to study risk factors and determine the likelihood of developing a specific outcome based on correlations. In a cohort study, participants share common characteristics or experiences within a defined time frame. They are divided into subgroups based on exposure to a suspected factor, and outcomes are tracked and compared across these subgroups. This method helps identify causal relationships between exposure factors and outcomes, as illustrated in Figure 4.

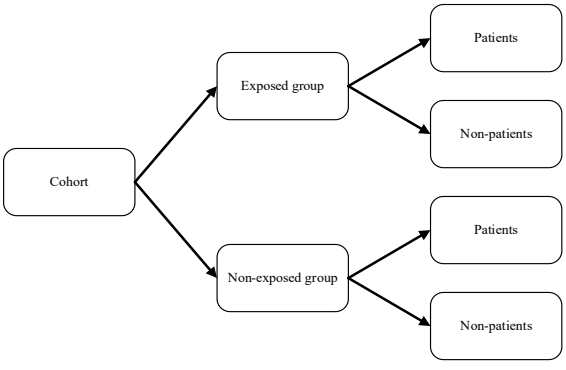

Figure 4. Cohort study model

To test the rule $P \rightarrow Z$, where P is the causal variable (exposure variable) and Z is the outcome variable (response variable), we manipulate the control set, ensuring that its attributes remain consistent. This isolates the effect of P on Z, unaffected by other variables.

In the dataset, we identify pairs of records: one with the exposure variable P and the other without it, while keeping the control set variables constant. By observing the response variable Z in these paired records, we can verify if the presence or absence of P influences Z, as illustrated in the following example.

Suppose a related rule $P \rightarrow Z$, with a control group $C = (A, B, D)$, $(P = 1, A = 1, B = 0, D = 1, Z = 1)$ and $(P = 0, A = 1, B = 0, D = 1, Z = 0)$. It can be observed that by fixing the control group variables to prevent influence from other control group variables, the presence of P leads to the occurrence of Z. This method can also be applied to multiple causal association odds ratio rules, such as the rule $\{P1, P2\} \rightarrow Z$, $(P1 = 1, P2 = 1, A = 1, B = 0, D = 1, Z = 1)$ and $(P1 = 0, P2 = 0, A = 1, B = 0, D = 1, Z = 0)$.

3.3.3 Fair Data Sets

The odds ratio quantifies the association between two events [26], defining a causal rule as "a change in attribute A affects attribute B." Common in medical and social research, it helps identify causal links. For example, sequential pattern mining may reveal $\{Gender = Male\} \rightarrow \{Salary = Low\}$, as shown in Table 2.

Table 2. Salary distribution by gender

Salary	Gender = Male	Gender = Female	Total
Low salary	185	65	250
High salary	120	60	180
Total	305	125	430

The table above shows the number of individuals in each category. For males, the ratio of low salary to high salary is 185/120 = 1.54, while for females, it is 65/60 = 1.08. This means the odds of a male earning a low salary is 1.54, while for a female, it is 1.08. The odds ratio (OR) for males compared to females is 1.54/1.08 = 1.43. This odds ratio indicates the strength of the causal association, showing that "gender = male" and "salary = low" are positively correlated. The odds ratio for the rule X→Y is defined as follows:

$$Odds\ Ratio\left(X \rightarrow Y\right) = \frac{supp\left(xy\right) \times supp\left(\neg x \neg y\right)}{supp\left(\neg xy\right) \times supp\left(x \neg y\right)} \quad \textbf{(1)}$$

The OR compares the probability of $Y = 1$ occurring in two groups: $X = 0$ and $X = 1$. A value greater than 1 indicates a positive correlation, while less than 1 indicates a negative correlation. A higher OR suggests a stronger association.

Determining a causal association odds ratio rule requires more than one data pair. Thus, a fair dataset is needed, containing records that match the rule $P \rightarrow Z$ and its control set. Based on the response variable, four possible types of paired records exist, as shown in Table 3.

Table 3. Four possible combinations of paired records

P=1 \ P=0	Z	$\neg Z$
Z	N_{11}	N_{12}
$\neg Z$	N_{21}	N_{22}

In the table above, N_{11} represents the number of paired sets in which both the exposed group $P = 1$ and the non-exposed group $P = 0$ contain Z. N_{12} represents the number of paired sets where the exposed group contains Z and the non-exposed group contains $\neg Z$. N_{21} represents the number of paired sets where the exposed group contains $\neg Z$ and the non-exposed group contains Z. Lastly, N_{22} represents the number of paired sets in which both the exposed and non-exposed groups contain $\neg Z$.

By calculating the quantity of all paired sets in the dataset, the odds ratio for a fair dataset can be determined. The greater the odds ratio value is above 1, the higher the confidence in the causal association odds ratio rule. To avoid a division-by-zero scenario, when N_{21} equals zero, this study sets the value to 1.

$$Odds\ Ratio\left(P \rightarrow Z\right) = \frac{N_{12}}{N_{21}} \quad \textbf{(2)}$$

Including too many variables in the control group can create an overly large set, reducing the chances of finding a valid fair dataset. Therefore, only variables related to the outcome should be included in the control set. Additionally, mutually exclusive variables, like P (university graduates) and Q (graduate school graduates), should be excluded, as their combinations $(P = 0, Q = 1)$ or $(P = 1, Q = 1)$ are impossible in the dataset.

Using previously identified causal association odds ratio rule candidates, one can determine which attributes are related to the outcome variable. All variables with a potential causal relationship to the outcome variable are included in the control group set, as illustrated in the Figure 5 below.

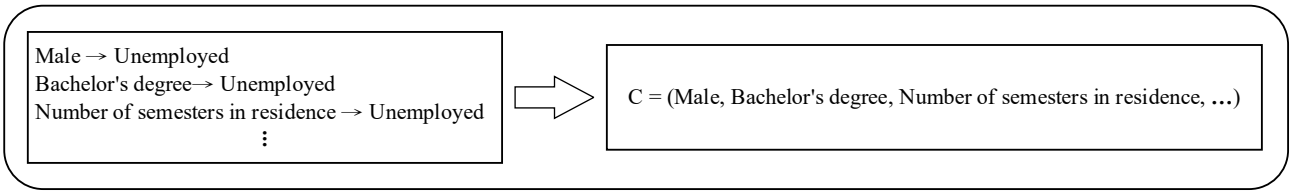

Figure 5. Selection of the control group

3.3.4 The Value of Identifying Causal Association Odds Ratio Rules with Multiple Variables

False negatives can occur when mutually exclusive variables are included in the control group set, leading to empty fair datasets. For example, consider the rule $P \rightarrow Z$. If a variable Q, which is associated with the outcome Z, is also related to the causal variable P, their relationship may create mutual exclusivity. This makes it impossible to find matching pairs with the same control set, as the association between P and Q alters Q when P varies. This is illustrated in Table 4 below.

Table 4. Relationship between antecedent, control group, and consequent

P (Antecedent)	Q (Control group)	Z (Consequent)
1	1	0
0	0	1

If the control group Q is excluded from the rule, it can lead to a false positive, where spurious relationships are mistaken for causal association odds ratio rules. For example, in the case of $P \rightarrow Q \rightarrow Z$ and $Q \rightarrow P \cap Q \rightarrow Z$, where Q is the primary antecedent of Z and P merely appears to be associated with Q, the rule $P \rightarrow Z$ is misleading. If Q, the primary antecedent, is not included in the control group when checking whether $P \rightarrow Z$ is a causal association odds ratio rule, a false positive may arise.

This study focuses on eliminating false negatives first, ensuring that no false positives occur. It then examines rules with an odds ratio of zero. By exploring causal association odds ratio rules involving multiple variables, including mutually exclusive variables in the causal variable set, the study aims to uncover missing causal relationships.

Multi-variable causal association odds ratio rules can recover false negatives and reveal relationships missed by single-variable methods. Some causal effects require multiple attributes to interact. This study explores these hidden connections, uncovering rules overlooked in single-variable analysis or candidate filtering.

4 Experiments

In this study, "activity expenditure" refers to the total amount spent during the brand's promotional activities in that year. "Expenditure on each product category" is the total amount spent on each product category, while "total expenditure" indicates the overall spending on the brand. "Total number of transactions" refers to the total number of transactions a customer made with the brand in that year.

To prepare the data for causal association odds ratio rule analysis, this study categorizes fields such as expenditure on each product category, total expenditure, and total number of transactions into five levels, using the median as the midpoint. Level 1 represents the lowest expenditure, while Level 5 represents the highest. Activity expenditure is categorized into two levels: "high" if it exceeds half of the total expenditure, and "low" if it is less than half of the total expenditure.

The required computation time and dataset size for each dataset are shown in Table 5. Below, we discuss some of the causal association odds ratio rules identified in the datasets for each brand and year.

Table 5. The required computation time and dataset size

Dataset	Causal association odds ratio rule verification time	Number of causal association odds ratio rule
2018 Brand A	181 minutes 5 seconds	425
2019 Brand A	108 minutes 7 seconds	245
2020 Brand A	120 minutes 19 seconds	393
2021 Brand A	146 minutes 15 seconds	474
2018 Brand B	1 minutes 38 seconds	139
2019 Brand B	1 minutes 35 seconds	137
2020 Brand B	79 minutes 45 seconds	289
2021 Brand B	38 minutes 2 seconds	300

In the 2018 dataset for Brand A, Table 6 the causal association odds ratio rules with higher odds ratios are mainly related to Category C products, Category B products, and activity expenditure. Specifically, customers whose activity expenditure accounts for a higher proportion of their total expenditure tend to have a total number of transactions at Level 1 and a total expenditure also at Level 1. These customers usually spend at Level 1 on Category C or Category B products. Additionally, customers with a Level 1 expenditure on Category C products and a Level 3 total number of transactions tend to spend at Level 4 on Category B products, with their total expenditure at Level 3. Customers with a Level 4 expenditure on Category B products typically have an activity expenditure ratio lower than half of their total expenditure, and both their total expenditure and total number of transactions reach Level 4. Customers with a Level 4 total expenditure usually also have a total number of transactions at Level 4 and a low activity expenditure ratio relative to total expenditure. Finally, customers with a Level 4 total number of transactions usually also have a total expenditure at Level 4 and spend at Level 4 on Category B products.

From these three causal association odds ratio rules, we can infer that the three features-Level 4 expenditure, Level 4 total number of transactions, and Level 4 Category B product expenditure-exhibit a strong causal relationship with one another.

Table 6. Excerpt from the 2018 Brand A causal association odds ratio rules dataset

Antecedents	Consequents	Odds ratio
Activity expenditure: Level high	Category C product expenditure: Level 1, Total number of transactions: Level 1, Total expenditure: Level 1	377774.5
	Category B product expenditure: Level 1, Total number of transactions: Level 1, Total expenditure: Level 1	327583.4
Category C product expenditure: Level 1, Total number of transactions: Level 3	Total expenditure: Level 3, Category B product expenditure: Level 4	21811.28
Category B product expenditure: Level 4	Activity expenditure: Low, Total expenditure: Level 4, Total number of transactions: Level 4	1633.901
Total number of transactions: Level 4	Total expenditure: Level 4, Category B product expenditure: Level 4	617.3958
Total expenditure: Level 4	Activity expenditure: Low, Total number of transactions: Level 4	276.5027

Table 7. Excerpt from the 2019 Brand A causal association odds ratio rules dataset

Antecedents	Consequents	Odds ratio
Activity expenditure: Level high	Category B product expenditure: Level 1, Total number of transactions: Level 1, Total expenditure: Level 1	4743538.364
Activity expenditure: Level high	Category C product expenditure: Level 1, Total number of transactions: Level 1, Total expenditure: Level 1	2457542.5
Category D product expenditure: Level 4	Category C product expenditure: Level 3	41721.70815
Category C product expenditure: Level 3	Category D product expenditure: Level 4	34495.22737
Activity expenditure: Level low, Category C product expenditure: Level 1, Age: 20-29 years	Category B product expenditure: Level 4	9827.26066
Activity expenditure: Level low, Category A product expenditure: Level 1, Total expenditure: Level 4		
Activity expenditure: Level low, Category D product expenditure: Level 1, Total expenditure: Level 3		
Category C product expenditure: Level 4	Activity expenditure: Level low, Total number of transactions: Level 4, Total expenditure: Level 4	7994.40406

In the 2019 dataset for Brand A, Table 7 the causal association odds ratio rules with higher odds ratios are still mainly related to Category C products, Category B products, and activity expenditure. However, causal association odds ratio rules related to Category D products appear more frequently than in the previous year. Specifically, customers whose activity expenditure accounts for a higher proportion of their total expenditure tend to have total number of transactions at Level 1 and total expenditure at Level 1. These customers generally spend at Level 1 on Category C or Category B products. Additionally, customers with Category C product expenditure at Level 3 tend to spend at Level 3 on Category D products. Similarly, customers with Category D product expenditure at Level 3 also tend to spend at Level 3 on Category C products, indicating a reciprocal causal relationship between these two. The reason for customers having Level 4 expenditure on Category B products is related to a lower activity expenditure ratio, with Category A, D, and C product expenditure at Level 1, customers aged between 20-29 years, and total expenditure at Level 3 as well as total number of transactions at Level 4. Finally, customers with Category C product expenditure at Level 4 tend to have lower activity expenditure ratios, with both total expenditure and total number of transactions at Level 4.

In the 2020 dataset for Brand A, Table 8 the causal association odds ratio rules with higher odds ratios are primarily related to Category C products, Category D products, Category B products, and activity expenditure. Specifically, customers whose activity expenditure accounts for a higher proportion of their total expenditure tend to have total number of transactions at Level 1 and total expenditure at Level 1. These customers typically spend at Level 1 on Category C or Category B products. Moreover, the expenditure on Category C products, Category D products, and total expenditure not only influence each other's causal relationships but are also affected by the total number of transactions. Customers with total number of transactions at Level 4 typically have total expenditure at Level 4 as well as expenditure on Category C products at Level 4. Additionally, in that year, customers aged between 20 and 29 years and with total expenditure at Level 3 tend to spend at Level 4 on Category B products, and their total number of transactions reaches Level 3.

In the 2021 dataset for Brand A, Table 9 the causal association odds ratio rules with higher odds ratios primarily involve Category C products, Category D products, Category B products, and activity expenditure. Specifically, customers whose activity expenditure accounts for a higher proportion of their total expenditure tend to have total number of transactions at Level 1 and total expenditure at Level 1. These customers typically spend at Level 1 on Category C or Category B products. The causal association odds ratio rule with the highest odds ratio in this dataset indicates that customers with Category C product expenditure at Level 1 and total number of transactions at Level 3 generally have total expenditure at Level 3 and spend at Level 4 on Category C products. Compared to previous years, the 2021 dataset introduces several causal association odds ratio rules related to membership Level 1. For instance, customers spending at Level 3 on Category D products may have a lower activity expenditure ratio, are likely to be at membership Level 1, and also spend at Level 3 on Category C products. Additionally, customers with Level 3 expenditure on Category D or Category C products and total expenditure at Level 2 are likely to have membership Level 1 and total number of transactions at Level 2.

Table 8. Excerpt from the 2020 Brand A causal association odds ratio rules dataset

Antecedents	Consequents	Odds ratio
Total number of transactions: Level 2, Category C product expenditure: Level 3	Category D product expenditure: Level 3, Total expenditure: Level 2	124375.5
Total number of transactions: Level 2, Category D product expenditure: Level 1	Category C product expenditure: Level 1, Total expenditure: Level 2	116622
Activity expenditure: Level high	Category B product expenditure: Level 1, Total number of transactions: Level 1, Total expenditure: Level 1	42525.94
Activity expenditure: Level high	Category C product expenditure: Level 1, Total number of transactions: Level 1, Total expenditure: Level 1	4046.066
Total number of transactions: Level 4	Category C product expenditure: Level 4, Total expenditure: Level 4	4776.43
Activity expenditure: Level low, Total number of transactions: Level 2	Category C product expenditure: Level 3, Total expenditure: Level 2	2835.444
Age: 20-29 years, Total expenditure: Level 3	Category B product expenditure: Level 4, Total expenditure: Level 2	361.6174

Table 9. Excerpt from the 2021 Brand A causal association odds ratio rules dataset

Antecedents	Consequents	Odds ratio
Category C product expenditure: Level 1, Total number of transactions: Level 3	Category B product expenditure: Level 4, Total expenditure: Level 2	28927.77
Activity expenditure: Level high	Category B product expenditure: Level 1, Total number of transactions: Level 1, Total expenditure: Level 1	27617.27
Activity expenditure: Level high	Category C product expenditure: Level 1, Total number of transactions: Level 1, Total expenditure: Level 1	2155.695
Category D product expenditure: Level 3	Category C product expenditure: Level 3, Activity expenditure: Level low, Membership Level: Level 1	1223.898
Category D product expenditure: Level 3, Total expenditure: Level 2	Membership Level: Level 1, Total number of transactions: Level 2	351
Category C product expenditure: Level 3, Total expenditure: Level 2	Membership Level: Level 1, Total number of transactions: Level 2	231.3529

Table 10. Excerpt from the 2018 Brand B causal association odds ratio rules dataset

Antecedents	Consequents	Odds ratio
Category D product expenditure: Level 4	Category A product expenditure: Level 4, Activity expenditure: Level low, Total number of transactions: Level 4	835396
	Category A product expenditure: Level 4, Membership Level: Level 3, Activity expenditure: Level low	125316
	Category A product expenditure: Level 4, Gender: Female, Activity expenditure: Level low	120409
Category A product expenditure: Level 4	Category D product expenditure: Level 4, Membership Level: Level 3	52524.33
	Category D product expenditure: Level 4, Total number of transactions: Level 4, Total expenditure: Level 4	50993.89
Total expenditure: Level 3	Total number of transactions: Level 3, Activity expenditure: Level low	7859.451
Total expenditure: Level 4	Activity expenditure: Level low, Membership Level: Level 3, Total number of transactions: Level 4	1038.8
Membership level: Level 3, Activity expenditure: Level low	Total number of transactions: Level 4, Total expenditure: Level 4	486.2654
Membership level: Level 1, Activity expenditure: Level high	Total number of transactions: Level 1, Total expenditure: Level 1	122.0982

In the 2018 dataset for Brand B, Table 10 the causal association odds ratio rules with higher odds ratios primarily involve Category D products and Category A products. Specifically, customers with Category D product expenditure at Level 4 also tend to spend at Level 4 on Category A products, indicating that customers who spend more on Category D products also tend to spend relatively more on Category A products compared to other categories. Customers whose expenditure reaches Level 3 or higher generally also have a total number of transactions at Level 3 or higher. Those with total expenditure at Level 4 often have a lower activity expenditure ratio, yet their spending and transaction levels still reach Level 4, and their membership level is typically 3. On the other hand, customers with higher activity expenditure ratios and membership Level 1 usually have fewer transactions and lower expenditure. Conversely, customers with lower activity expenditure ratios and membership Level 3 tend to have higher spending and more transactions. This suggests that customers with higher membership levels and lower activity expenditure ratios are a more stable customer group, characterized by frequent transactions and higher spending.

In the 2019 dataset for Brand B, Table 11 the causal association odds ratio rules with higher odds ratios primarily involve Category D products and Category A products. Specifically, customers whose Category D product expenditure reaches Level 4 also tend to have Category A product expenditure at Level 4, indicating that customers who spend more on Category D products also spend relatively more on Category A products compared to other categories. Customers with Category L or Category G product expenditure at Level 1 and total expenditure at Level 2 tend to have a higher activity expenditure ratio and a transaction count at Level 2 during the activity period. Conversely, customers with a higher activity expenditure ratio typically have transaction counts at Level 1 and total expenditures at Level 1 as well.

Table 11. Excerpt from the 2019 Brand B causal association odds ratio rules dataset

Antecedents	Consequents	Odds ratio
Category D product expenditure: Level 4	Category A product expenditure: Level 4, Activity expenditure: Level low, Total number of transactions: Level 4	985056
Category D product expenditure: Level 4	Category A product expenditure: Level 4	433876
Category A product expenditure: Level 4	Category D product expenditure: Level 4	329183.1351
Total expenditure: Level 3	Activity expenditure: Level low, Total number of transactions: Level 3	9386.045
Category L product expenditure: Level 1, Total expenditure: Level 2	Activity expenditure: Level high, Total number of transactions: Level 2	3461.571
Total expenditure: Level 4	Activity expenditure: Level low, Membership Level: Level 3, Total number of transactions: Level 4	1420.208
Activity expenditure: Level high	Membership Level: Level 1, Total number of transactions: Level 2	343.6088
Category G product expenditure: Level 1, Total expenditure: Level 2	Activity expenditure: Level high, Total number of transactions: Level 2	290.2216

Table 12. Excerpt from the 2020 Brand B causal association odds ratio rules dataset

Antecedents	Consequents	Odds ratio
Category D product expenditure: Level 4	Category A product expenditure: Level 4	6027392
Category A product expenditure: Level 4	Category D product expenditure: Level 4	4020576
Category D product expenditure: Level 4	Category A product expenditure: Level 4, Total number of transactions: Level 4	3428052
Activity expenditure: Level high, Total number of transactions: Level 2	Category G product expenditure: Level 1, Total expenditure: Level 2	1124660
Activity expenditure: Level low, Total number of transactions: Level 2	Category G product expenditure: Level 1, Total expenditure: Level 3	6784
Gender: Female, Total expenditure: Level 3	Activity expenditure: Level low, Total number of transactions: Level 3	2222.202
Category G product expenditure: Level 1, Age: 30-39 years	Activity expenditure: Level high, Total number of transactions: Level 1	5.847269
Category F product expenditure: Level 1, Age: 30-39 years	Total number of transactions: Level 1	1.031177

In the 2020 dataset for Brand B, Table 12 the causal association odds ratio rules with higher odds ratios primarily involve Category D products and Category A products. Specifically, customers whose Category D product expenditure reaches Level 4 also tend to have Category A product expenditure at Level 4, indicating that those who spend more on Category D products also spend relatively more on Category A products compared to other categories. In this dataset, rules involving Category G product expenditure at Level 1 appear more frequently than in the previous two years. This suggests a shift in customer purchasing behavior, with a trend of spending smaller amounts on Category G products emerging in this year. Customers aged 30-39 years who spend relatively less on Category G or Category F products tend to have transaction counts at Level 1, indicating fewer purchases overall. Furthermore, customers in the same age group who spend less on Category G products typically have transaction counts at Level 1 but exhibit a higher activity expenditure ratio, with a larger portion of their spending occurring during promotional periods.

In the 2021 dataset for Brand B, Table 13 the causal association odds ratio rules with higher odds ratios continue to prominently feature Category D products and Category A products. Specifically, customers whose Category D product expenditure reaches Level 4 also tend to have Category A product expenditure at Level 4, indicating that those who spend more on Category D products also allocate relatively higher spending to Category A products compared to other categories. Customers in their 20s who spend Level 1 amounts on both Category A and Category D products typically have lower transaction counts. This suggests that the brand's pricing might be higher, making it less accessible for customers in their early 20s, who may have just entered the workforce, to make frequent purchases. In this dataset, rules involving Category G product expenditure at Level 1 remain significant, suggesting that the shift in purchasing behavior observed in 2020-where customers spent smaller amounts on Category G products-persisted into 2021, reflecting a continued habit of minimal spending in this category.

Table 13. Excerpt from the 2021 Brand B causal association odds ratio rules dataset

Antecedents	Consequents	Odds ratio
Category D product expenditure: Level 4	Category A product expenditure: Level 4	4143260
Category A product expenditure: Level 4	Category D product expenditure: Level 4	4141225
Category A product expenditure: Level 4	Category D product expenditure: Level 4, Total number of transactions: Level 4	1397124
Category D product expenditure: Level 4	Category A product expenditure: Level 4, Category G product expenditure: Level 1	887364
Category A product expenditure: Level 4	Category D product expenditure: Level 4, Category G product expenditure: Level 1	887364
Category G product expenditure: Level 1, Total expenditure: Level 2	Activity expenditure: Level high, Total number of transactions: Level 1	3638.541
Total expenditure: Level 1	Category G product expenditure: Level 1, Activity expenditure: Level high, Total number of transactions: Level 1	1882.012
Activity expenditure: Level high	Total number of transactions: Level 1, Total expenditure: Level 1	949.606
Age: 30-39 years, Category A product expenditure: Level 1, Category D product expenditure: Level 1	Total number of transactions: Level 1	97.53375527

5 Discussion and Conclusions

5.1 Discussion

In the Brand A dataset, customer purchases from 2018 to 2021 show a strong bidirectional causal relationship between products in Categories B and C, with Category D also influencing purchases starting in 2019. The causal association odds ratio rules for spending across these categories vary each year, but the relationship between Categories B, C, and D remains consistently strong, suggesting these products may have overlapping uses or relevance.

Additionally, a causal relationship is observed between promotional spending and total spending. Customers with a higher promotional spending ratio tend to fall into spending level 1, but there is no significant rule indicating that lower promotional spending ratios lead to higher total spending levels. This suggests that promotional offerings may have attracted customers who typically spend little, pushing them to higher levels of spending during promotional periods.

In 2019 and 2020, a causal relationship is observed between customers aged 20 to 29 and spending level 4 on Category B products, indicating that this age group was particularly attracted to Category B products during those years. In 2021, spending levels 3 and 2 in Categories D and C, as well as total spending level 2, show a causal relationship with membership level 1 customers with purchase frequency level 2. This suggests that lower-tier or new members were attracted to these products, leading to higher purchase frequencies.

For Brand B, a strong bidirectional causal relationship is observed between Category A and Category D products from 2018 to 2021. Customers with spending level 4 on Category A products are highly likely to also have spending level 4 on Category D products, indicating that these categories may be interdependent or frequently sold together.

Additionally, there is a causal relationship between promotional spending and total spending. Generally, customers with a higher ratio of promotional spending to total spending tend to have a lower total spending level, while those with a lower promotional spending ratio tend to have a higher total spending level. This suggests that the brand's products may be relatively high-priced, and some customers might choose to purchase only a limited quantity of products during promotional periods at discounted prices, avoiding purchases during non-promotional periods.

In datasets from 2020 onward, there is an increase in

causal association odds ratio rules involving spending level 1 on Category G products compared to 2018 and 2019. This change suggests that starting in 2020, customers' lifestyles may have shifted, influencing their purchasing habits for beauty products, leading them to spend lower amounts on Category G products.

Across the four-year datasets, both Brand A and Brand B exhibit unique factor combinations with strong causal relationships. Notably, spending on Category D products consistently appears as a key factor for both brands, indicating its potential impact on sales performance. However, post-2020, Brand B shows noticeable shifts in customer spending habits, a trend absent in Brand A. This contrast may stem from differences in marketing strategies, product characteristics, or target customer segments.

5.2 Conclusion

Since database data is dynamic and frequently updated, schema changes may impact analysis accuracy and efficiency. As data volume increases, identifying causal association odds ratio rules requires more time and computational resources. To enhance practical applications, a progressive system is needed to deliver real-time analysis, reducing delays and improving decision-making flexibility. This study reveals that customer spending amount and frequency are key factors in causal association analysis. However, using the median to categorize spending and purchase frequency into five levels lacks precise range definitions.

Academically, this research proposes a causal analysis framework operating between raw datasets and multidimensional data warehouses, enhancing data processing flexibility. By extending traditional association rule mining, it uncovers deeper causal insights to support decision-making. Additionally, it optimizes data warehouse design by identifying effective dimensions and improving interpretability in complex data environments, bridging the gap between raw data and high-level analysis in causal inference.

In the future, in collaboration with industry, can refine grading methods by gaining deeper dataset insights, leading to more precise analysis. Instead of the current five-level division based on the median, future plans include categorizing customer spending and purchase frequency into 10 to 20 levels using actual data distribution. This approach will enhance the interpretation of purchasing behaviors and produce more intuitive and comprehensible causal association odds ratio rules.

Acknowledgements

This research was sponsored by the National Science and Technology Council, Taiwan under the Contract No. MOST 110-2622-E-008-013.

References

[1] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, S. Belfkih, Big Data technologies: A survey, *Journal of King Saud University-Computer and Information Sciences*, Vol. 30, No. 4, pp. 431-448, October, 2018. https://doi.org/10.1016/j.jksuci.2017.06.001

[2] U. Awan, S. Shamim, Z. Khan, N. U. Zia, S. M. Shariq, M. N. Khan, Big data analytics capability and decision-making: The role of data-driven insight on circular economy performance, *Technological Forecasting and Social Change*, Vol. 168, Article No. 120766, July, 2021. https://doi.org/10.1016/j.techfore.2021.120766

[3] S. Akter, K. Michael, M. R. Uddin, G. McCarthy, M. Rahman, Transforming business using digital innovations: The application of AI, blockchain, cloud and data analytics, *Annals of Operations Research*, Vol. 308, No. 1-2, pp. 7-39, January, 2022. https://doi.org/10.1007/s10479-020-03620-w

[4] J. Wang, C. Xu, J. Zhang, R. Zhong, Big data analytics for intelligent manufacturing systems: A review, *Journal of Manufacturing Systems*, Vol. 62, pp. 738-752, January, 2022. https://doi.org/10.1016/j.jmsy.2021.03.005

[5] A. Labrinidis, H. V. Jagadish, Challenges and opportunities with big data, *Proceedings of the VLDB Endowment*, Vol. 5, No. 12, pp. 2032-2033, August, 2012. https://doi.org/10.14778/2367502.2367572

[6] J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, C. Carlsson, Past, present, and future of decision support technology, *Decision support systems*, Vol. 33, No. 2, pp. 111-126, June, 2002. https://doi.org/10.1016/S0167-9236(01)00139-7

[7] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, S. A. Tsaftaris, Causal machine learning for healthcare and precision medicine, *Royal Society Open Science*, Vol. 9, No. 8, Article No. 220638, August, 2022. https://doi.org/10.1098/rsos.220638

[8] J. W. Tukey, *Exploratory data analysis*, Addison-Wesley, 1977.

[9] N. Miloslavskaya, A. Tolstoy, Big data, fast data and data lake concepts, *Procedia Computer Science*, Vol. 88, pp. 300-305, 2016. https://doi.org/10.1016/j.procs.2016.07.439

[10] A. Bogatu, A. A. Fernandes, N. W. Paton, N. Konstantinou, *Dataset discovery in data lakes*, *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, Dallas, TX, USA, 2020, pp. 709-720. https://doi.org/10.1109/ICDE48307.2020.00067

[11] J. Han, J. Pei, H. Tong, *Data mining: concepts and techniques*, Morgan Kaufmann, 2022.

[12] P. Spirtes, Introduction to causal inference, *Journal of Machine Learning Research*, Vol. 11, No. 54, pp. 1643-1662, March, 2010. https://www.jmlr.org/papers/v11/spirtes10a.html

[13] X. Shu, Y. Ye, Knowledge Discovery: Methods from data mining and machine learning, *Social Science Research*, Vol. 110, Article No. 102817, February, 2023. https://doi.org/10.1016/j.ssresearch.2022.102817

[14] B. K. Rimer, Correlation is not causation, *American Journal of Public Health*, Vol. 88, No. 5, pp. 832-835, May, 1998. https://doi.org/10.2105/AJPH.88.5.832

[15] M. Van Hul, T. Le Roy, E. Prifti, M. C. Dao, A. Paquot, J. D. Zucker, N. M. Delzenne, G. Muccioli, K. Clément, P. D. Cani, From correlation to causality: the case of Subdoligranulum, *Gut microbes*, Vol. 12, No. 1, pp. 1-13, November, 2020. https://doi.org/10.1080/19490976.2020.1849998

[16] J. Pearl, D. Mackenzie, *The book of why*, Penguin Books, 2019.

[17] F. A. Tan, To Know the Causes of Things: Text Mining for Causal Relations, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 21, pp. 23425-23426, March, 2024. https://doi.org/10.1609/aaai.v38i21.30413

[18] M. G. Majumder, S. D. Gupta, J. Paul, Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis, *Journal of Business Research*, Vol. 150, pp. 147-164, November, 2022. https://doi.org/10.1016/j.jbusres.2022.06.012

[19] T. A. Brown, M. T. Moore, Confirmatory factor analysis, in R. H. Hoyle (Ed.), *Handbook of structural equation modeling*, The Guilford Press, 2012, pp. 361-379.

[20] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, B. Sun, S. Ma, From observational studies to causal rule mining, *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 7, No. 2, pp. 1-27, January, 2016. https://doi.org/10.1145/2746410

[21] N. Pawlowski, D. Coelho de Castro, B. Glocker, Deep structural causal models for tractable counterfactual inference, *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 2020, pp. 857-869.

[22] C. Borgelt, An Implementation of the FP-growth Algorithm, *OSDM'05: Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, Chicago, Illinois, USA, 2005, pp. 1-5. https://doi.org/10.1145/1133905.1133907

[23] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, San Francisco, CA, USA, 1994, pp. 487-499.

[24] S. Brin, R. Motwani, C. Silverstein, Beyond market baskets: Generalizing association rules to correlations, *ACM SIGMOD Record*, Vol. 26, No. 2, pp. 265-276, June, 1997. https://doi.org/10.1145/253262.253327

[25] J. W. Song, K. C. Chung, Observational studies: cohort and case-control studies, *Plastic and reconstructive surgery*, Vol. 126, No. 6, pp. 2234-2242, December, 2010. https://doi.org/10.1097/PRS.0b013e3181f44abc

[26] J. M. Bland, D. G. Altman, *The odds ratio*, BMJ, Vol. 320, p. 1468, May, 2000. https://doi.org/10.1136/bmj.320.7247.1468

Biographies

Ting-Yan Lin, currently is a Ph.D. student at the Department of Computer Science and Information Engineering, National Central University, Taiwan. His research interests include data mining, causal inference, statistical and big data in education.

Meng-Feng Tsai received the Ph.D. degree in Computer Science from UCLA, USA in 2004. He then joined the Department of Computer Science and Information Engineering, National Central University, Taiwan, as Assistant Professor in 2004. His current research interests include data warehouse, database systems, data mining, scientific computation and distributed computation.

Chi-Sheng Huang received the Ph.D. degree in Computer Science and Information Engineering from NCU, Taiwan, in 2018. He then joined the Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, Taiwan, as Assistant Professor in 2022. His research interests include data mining and cloud computation.

Liang-Han Lin, he received the master degree in Computer Science and Information Engineering from National Central University, Taiwan, in 2022. His research interests include data mining, causal inference, statistical and big data in education.

Jiun-Yi Tsai, she graduated with a Master's degree from National Central University in Taiwan in 2022. Her research interests include data mining and big data analysis, with a focus on Institutional Research and open data.