Association rule/pattern mining for recommender system

Group:

Hengyi Ma a1875198

Weiyou Liu a1872800

The University of Adelaide

4433_COMP_SCI_7306 Mining Big Data

Lecturer: Dr. Alfred Krzywicki

Table of Contents

1. Executive Summary

Pattern mining is a key subfield of data mining that aims at developing algorithms to discover interesting patterns in databases [1]. This technology can be combined with common recommendation systems nowadays to solve many practical problems, such as recommending similar users on social platforms and recommending products on e-commerce platforms. We take the sale of goods in a grocery store as a case study, using pattern mining technology and designing a recommendation system to process the store's sales information. This helps the store find intuitive and useful sales information for operators and provide suitable product recommendations for customers to help the store increase sales. Our method reveals some useful patterns in product sales, which can provide guidance for the store's product purchase strategy. At the same time, our designed item-based recommendation system has been successfully tested on the provided real dataset. We have also discovered some patterns to improve the performance of the recommendation system, which will help us further improve the recommendation system and ultimately apply it directly to help the store increase profits.

2. Introduction

With the continuous development of big data technology and the advancement of artificial intelligence, models designed for massive data are showing excellent performance in various fields. One of the most popular and well-known models is the recommendation system. In this project, we will attempt to design a recommendation system based on the sales data of a grocery store to provide recommendations for different customers' purchase information, helping customers find items of interest faster, which may increase sales. In addition to directly using the recommendation system to process data and generate recommendations, we will also use pattern mining methods for frequent itemset mining and association rule mining to find combinations of items that frequently appear together, which can provide more hidden information in the data to help store operators better understand purchasing patterns and formulate useful procurement strategies. We will also combine the generated association rules with the recommendation system to observe whether there is a difference in using or not using association rules in the process of constructing the recommendation system, which will help us build a more accurate recommendation system and further increase sales.

2.1. Recommender system

A recommender system, or a recommendation system, is a subclass of information filtering system that provides suggestions for items that are most pertinent to a particular user [2]. Recommendation systems analyze data using models to calculate the similarity between targets, and then recommend similar items to each target, thereby increasing user interest and helping users find things they might be interested in. Recommendation systems have many specific

applications, most of which are concentrated in the commercial field. For example, e-commerce platforms can use recommendation systems to recommend items that users may be interested in, thereby increasing transaction volume. Social media platforms recommend other users that users may like or know, thereby enhancing user stickiness. Various streaming media platforms that we commonly see in life also try to discover works that users may like in this way, thereby increasing user frequency of use.

2.2. Association rule/pattern mining

Association rule mining and frequent item set mining are commonly used techniques in the field of data mining, which are used to discover associations between items in data sets. Association rule mining is a technique used to discover associations between items in a data set. It is often used to discover relationships between items that appear frequently in a data set. The goal is to find rules with sufficient support and confidence. Frequent itemset mining is a subtask of association rule mining, and its goal is to discover frequently occurring item sets in the data set. These two technologies are widely used in recommendation systems and can help analyze correlations in data sets to provide decision support and insights.

3. Exploratory Analysis

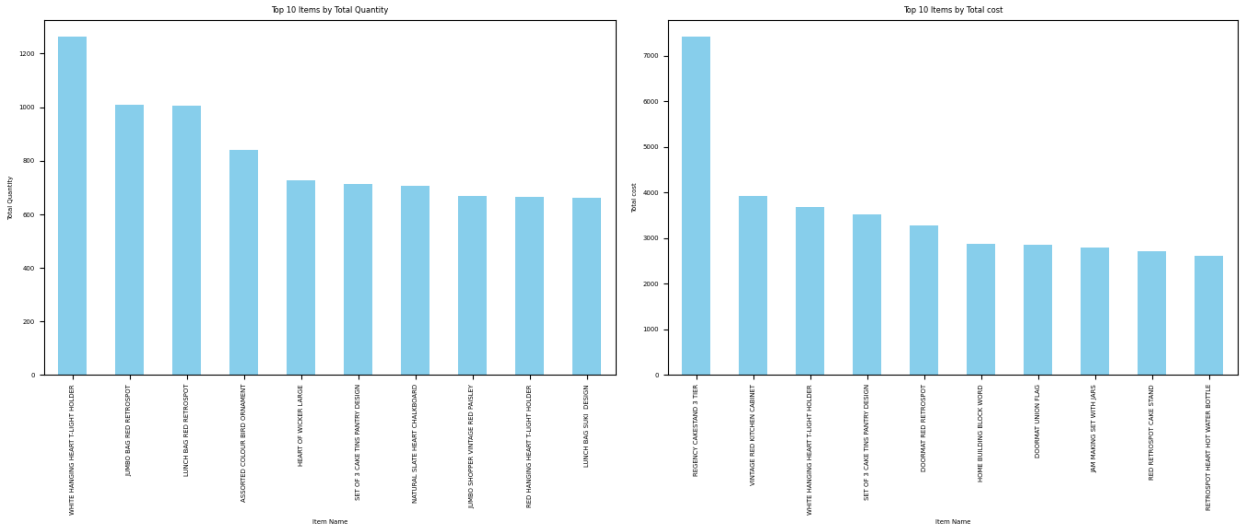
3.1. Content check

First, we need to determine the contents of the file to establish the data structure and identify potentially useful information, which will guide us in subsequent data processing. Upon reviewing the data, we found the following columns: BillNo, Itemname, Quantity, Date, Price, CustomerID, and Cost. Our tasks involve association rule mining and recommendation system design, which require constructing item similarity and customer profiles. Our ultimate goal is to increase sales, so the most important data for our design includes Itemname, Quantity, CustomerID, and Cost. Additionally, we noticed that the original data structure has each item purchased by each customer listed in a separate column. Therefore, we will need to redesign the structure of the dataset for further analysis.

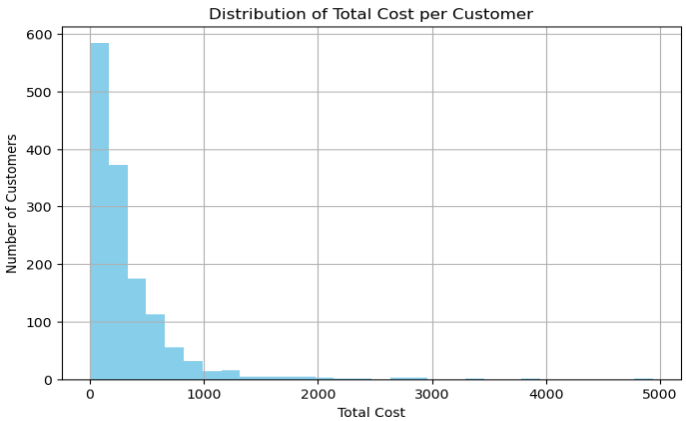
3.2. Visualization Analysis

We then further analyze the data using visualization techniques. Firstly, we rank the products based on total quantity and total spending, respectively. We use histograms to represent the top ten products in each ranking. These visualizations help us identify a group of products with the highest sales volume and revenue. These products represent the most popular and profitable items for the store and can serve as a basis for guiding store purchases. The top ten products ranked by quantity and spending, along with the quantity/cost, are shown in the following figure. Taking "WHITE HANGING HEART T-LIGHT HOLDER" as an example, we can see that this

product appears in both tables, indicating its high popularity.



We also created a distribution plot for the total spending of individual customers in the store. Although this does not directly assist in pattern mining or the recommendation system, it has potential benefits for our client. For example, it can help the store understand customer retention rates (whether there are more new or returning customers, which can be reflected in the distribution of total spending) and assist the store in determining its consumer positioning (high-end, mid-range, low-end). The distribution plot of total spending is shown in the figure below. It can be observed that the majority of customers spend between 0 and 1000, but there are also a considerable number of loyal customers who spend more than 1000.



3.3. Data preprocessing

Firstly, we check whether there are missing values in the training set. If there are, we will delete the row directly. However, the results show that there are no missing values in the data and can be used directly.

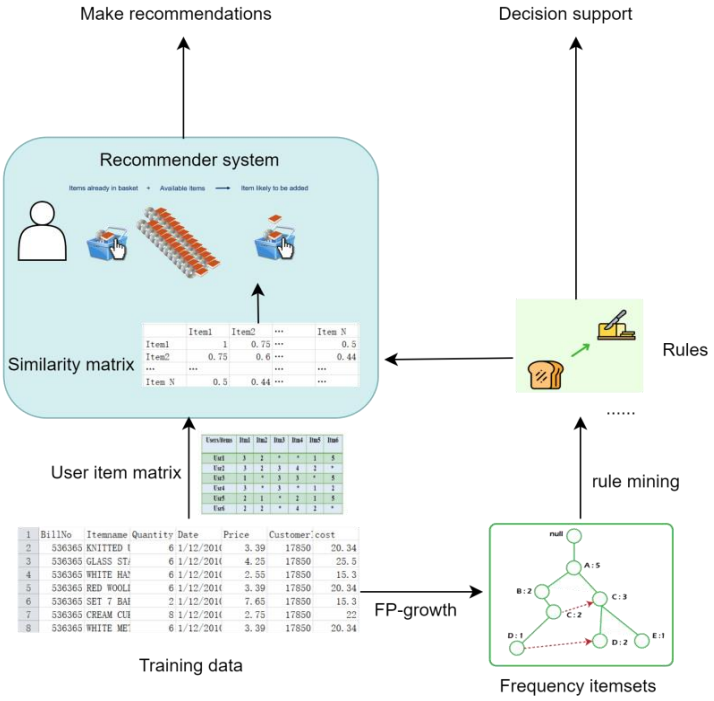
Then, we hope to integrate all the purchase records of each user into one record and separate the statistics according to the product name. This will allow us to intuitively view the user's purchase

of the product, and will also help us to establish a user-item matrix in the future. But the way we reconstruct data is slightly different in rule mining and recommendation systems. In rule mining, we are more concerned with whether another item exists when the current item exists, so the purchase situation is determined with True and False. But in the recommendation system, in order to establish a more accurate user-item matrix, we specifically consider the purchase quantity of the item.

4. Implementation and Testing

4.1. Method introduction

In this section, we present the details of the algorithm used in our task. The entire algorithm is divided into two parts. The first part is to conduct pattern mining based on data to generate frequent item sets and association rules that meet the conditions to find the potential rules of the data. The second part is to establish a suitable recommendation system algorithm to recommend items to users. The flow chart of the entire algorithm is shown in the figure below.



4.1.1. Method for pattern mining

In the pattern mining part, we use FP-growth to generate frequent item sets and association rules that meet the conditions. There are several commonly used algorithms for pattern mining, like Apriori, Eclat, TreeProjection and FP-growth [3]. Among them, FP-Growth is fast and good for large datasets, there are also some other advantages using FP-Growth, like Scalability and Flexibility, we get some of them from Anıl Çoğalan [4].

The usage process of FP-growth algorithm is summarized as follows [5]:

1. Scan the database, calculate the support (frequency) of each item, and obtain the frequent item set.

2. Construct an FP tree and insert transactions into the tree in descending order of support.

3. Based on the FP tree, a conditional pattern base (Conditional Pattern Base) is generated.

4. For each frequent item set, the conditional pattern base is extracted from the FP tree, and the conditional FP tree is recursively constructed to obtain the frequent item set.

5. Repeat steps 3 and 4 until no more frequent item sets can be generated.

4.1.2.Method for recommender system

We build an item-based recommendation system based on the code from workshop5 [6]. In this process, we use cosine similarity to calculate the similarity between items, and calculate different items and recommendation scores based on the similarity between user purchase records and nearby items, and sort the top ten as our recommended items. We also added logic to generate recommendations based on the pattern generated by task1. For the cold start problem of new users, we use a simple but effective logic, which is to select the most popular items for recommendation. In general, our recommendation system process can be summarized as the following process:

1. Initialization: In the initialization phase, the user needs to provide training data and test data. The recommendation system will convert this data into a user-item matrix, where each element represents the quantity of an item purchased by a user.

2. Construct User-Item Matrix: Through the users_items_matrix method, convert the training data and test data into a user-item matrix. This matrix represents the items purchased by each user and their quantities.

3. Use Patterns (Optional): If the user chooses to use patterns, the system will read the rules.csv file containing rules and modify the user-item matrix based on these rules to introduce pattern information.

4. Calculate Item Similarity: Calculate the similarity between items through the cos_similarity method. Cosine similarity is used here to measure the similarity between items.

5. Generate Recommendations: Use the recommend method to generate recommended items for users. If the user provides shopping basket information, the system will calculate recommendations based on the items in the basket; otherwise, it will use the most popular items as recommendations.

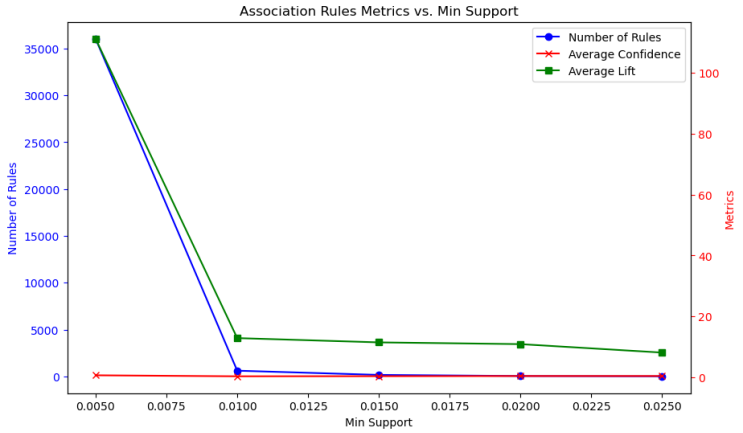
6. Train the Recommendation System: Use the train method to calculate item similarity on the

training data and introduce pattern information if necessary.

7. Evaluate the Recommendation System: Use the evaluate method to evaluate the performance of the recommendation system. For existing users and new users, the system calculates metrics.

4.2. Hyper-parameter study

Deciding the threshold for support is an important step for FP-Growth in pattern mining. This will determine the number and quality of frequent item sets we generate using the algorithm. We selected different min_support thresholds for experiments and used the results to find a relatively suitable threshold. Based on the experimental results from the figure below, we chose threshold=0.01, because it can be seen that this is the inflection point of the entire data, and a trade-off between the number of rules, lift and threshold is reached here.



For generating the rules, confidence is another hyper-parameter which is used to measure the reliability of the rules. In the process of generating rules, the threshold is set to 0.7, so that the generated rules are all high-confidence rules, and further adjustments to the relevant rules can be omitted when building the recommendation system later.

For recommender system, there are two parameters we can choose to vary, including n (number of neighbors) and N (the number of recommendations that the model provide). The former is an important factor affecting the score calculation, and the latter determines how many similar items our model will recommend. For different numbers of n, we will conduct controlled experiments and our baseline model choose N=3, and this part of the research will be shown in Section 5. For the number of recommended items N for the model, we choose TOP 10 recommendations.

4.3. Evaluation metrics

We use Precision, recall and F1-score as evaluation metrics for recommender system.

Precision rate refers to the proportion of items recommended to users that are actually liked by users. The higher the accuracy, the higher the proportion of items recommended by the system

that truly meet the user's preferences.

Recall rate refers to the proportion of items that users actually like that are successfully recommended by the system. The higher the recall rate, the more comprehensive the system can cover the items that users like.

The F1 value is the harmonic average of precision and recall, which takes into account the accuracy and comprehensiveness of the recommendation results. The closer the value is to 1, the better the recommendation effect of the system.

5. Results analysis

5.1. Frequent patterns

We have discovered 698 frequent item sets in the training set and 974 frequent item sets in the test set using FP-growth. In the training set, the top five frequent item sets and their support are WHITE HANGING HEART T-LIGHT HOLDER (0.105388), KNITTED UNION FLAG HOT WATER BOTTLE (0.039223), RED WOOLLY HOTTIE WHITE HEART (0.035658), SET 7 BABUSHKA NESTING BOXES (0.031300), and CREAM CUPID HEARTS COAT HANGER (0.029319). In the test set, the top five frequent item sets and their support are JUMBO SHOPPER VINTAGE RED PAISLEY (0.051071), LOVE BUILDING BLOCK WORD (0.042834), EDWARDIAN PARASOL NATURAL (0.019769), RED HEART SHAPE LOVE BUCKET (0.014827), and WOODEN HAPPY BIRTHDAY GARLAND (0.014827). And the support of top five frequent patterns which in training set in test set are WHITE HANGING HEART T-LIGHT HOLDER (0.084020), KNITTED UNION FLAG HOT WATER BOTTLE (0.011532), RED WOOLLY HOTTIE WHITE HEART (<0.01), SET 7 BABUSHKA NESTING BOXES (<0.01), and CREAM CUPID HEARTS COAT HANGER (0.011532). We can find that the support of the five items with the highest frequency in the training set has changed significantly in the test set, and some cannot even reach the threshold of 0.01, which indicates that there may be a large difference between the training set and the test set data.

Both datasets show differences in data size and the displayed product names. For example, in the training set, WHITE HANGING HEART T-LIGHT HOLDER has a support indicating that this product is much more frequent than others, while in the test set, JUMBO SHOPPER VINTAGE RED PAISLEY and LOVE BUILDING BLOCK WORD are much more frequent than others.

5.2. Recommendations for each pattern

Next, we select two customers each who have purchased specific items under five different patterns from train set, and use the model involving patterns to make recommendations. This result in a total of 5 sets of recommendations, each containing 2 customers, for a total of 10

recommendations. We choose 2 customers who have purchased WHITE HANGING HEART T-LIGHT HOLDER, KNITTED UNION FLAG HOT WATER BOTTLE, RED WOOLLY HOTTIE WHITE HEART, SET 7 BABUSHKA NESTING BOXES, and CREAM CUPID HEARTS COAT HANGER respectively (all of which come from different patterns mentioned above). The recommendation results are as follows. We can see that customers who purchase different items will receive completely different recommendations.

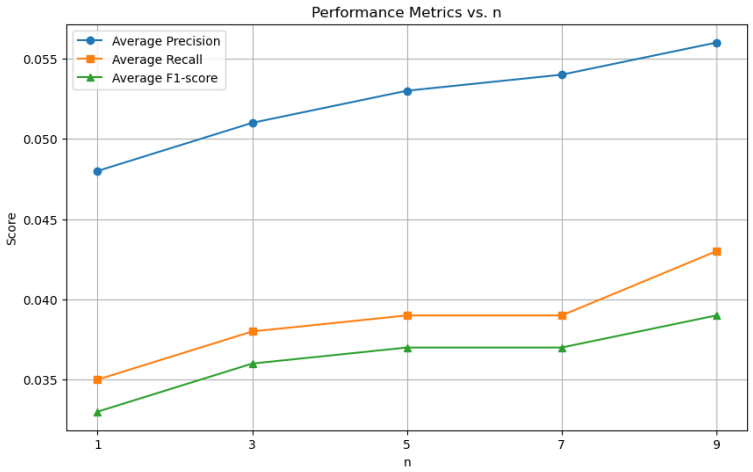
Pattern	CustomerID	Recommendation
1	17338	72 SWEETHEART FAIRY CAKE CASES, HEART OF WICKER SMALL, REX CASH+CARRY JUMBO SHOPPER, 10 COLOUR SPACEBOY PEN, 12 DAISY PEGS IN WOOD BOX, 12 EGG HOUSE PAINTED WOOD, 12 IVORY ROSE PEG PLACE SETTINGS, 12 MESSAGE CARDS WITH ENVELOPES, 12 PENCIL SMALL TUBE WOODLAND, 12 PENCILS SMALL TUBE RED
	16014	HEART OF WICKER SMALL, NATURAL SLATE HEART CHALKBOARD, RETROSPOT RED WASHING UP GLOVES, RECIPE BOX RETROSPOT, REX CASH+CARRY JUMBO SHOPPER, CLOTHES PEGS RETROSPOT PACK 24, HEART OF WICKER LARGE, JUMBO BAG RED RETROSPOT, SMALL POPCORN HOLDER, HAND WARMER BABUSHKA DESIGN
2	17338	72 SWEETHEART FAIRY CAKE CASES, HEART OF WICKER SMALL, REX CASH+CARRY JUMBO SHOPPER, 10 COLOUR SPACEBOY PEN, 12 DAISY PEGS IN WOOD BOX, 12 EGG HOUSE PAINTED WOOD, 12 IVORY ROSE PEG PLACE SETTINGS, 12 MESSAGE CARDS WITH ENVELOPES, 12 PENCIL SMALL TUBE WOODLAND, 12 PENCILS SMALL TUBE RED
	13077	10 COLOUR SPACEBOY PEN, 12 COLOURED PARTY BALLOONS, 12 DAISY PEGS IN WOOD BOX, 12 EGG HOUSE PAINTED WOOD, 12 IVORY ROSE PEG PLACE SETTINGS, 12 MESSAGE CARDS WITH ENVELOPES, 12 PENCIL SMALL TUBE WOODLAND, 12 PENCILS SMALL TUBE RED RETROSPOT, 12 PENCILS SMALL TUBE SKULL, 12 PENCILS TALL TUBE POSY
3	12709	HAND WARMER RED POLKA DOT, RETRO COFFEE MUGS ASSORTED, VINTAGE BILLBOARD DRINK ME MUG, KNITTED UNION FLAG HOT WATER BOTTLE, VINTAGE BILLBOARD LOVE/HATE MUG, GLASS STAR FROSTED T-LIGHT HOLDER, IVORY EMBROIDERED QUILT, SAVE THE PLANET MUG, WHITE METAL LANTERN, HAND WARMER UNION JACK
	17817	FELTCRAFT CUSHION RABBIT, 72 SWEETHEART FAIRY CAKE CASES, IVORY KITCHEN SCALES, ENAMEL WASH BOWL CREAM, LOVE BUILDING BLOCK WORD, HEART OF WICKER LARGE, HOME BUILDING BLOCK WORD, SET OF 3 CAKE TINS PANTRY DESIGN, HEART OF WICKER SMALL, RETROSPOT TEA SET CERAMIC 11 PC
4	12509	JUMBO BAG OWLS, SMALL FOLKART STAR CHRISTMAS DEC, JUMBO BAG RED RETROSPOT, HEART OF WICKER LARGE, PAPER BUNTING WHITE LACE, LUNCH BAG SPACEBOY DESIGN, WICKER STAR, SMALL PURPLE BABUSHKA NOTEBOOK, JUMBO BAG CHARLIE AND LOLA TOYS, WOOD BLACK BOARD ANT WHITE FINISH
	15241	HOT WATER BOTTLE BABUSHKA, POPPY'S PLAYHOUSE KITCHEN, PINK CREAM FELT CRAFT TRINKET BOX, FELTCRAFT DOLL MARIA, FELTCRAFT CUSHION OWL, FELTCRAFT CHRISTMAS FAIRY, FELTCRAFT CUSHION BUTTERFLY, GREY FLORAL FELTCRAFT SHOULDER BAG, FELTCRAFT CUSHION RABBIT, PINK BLUE FELT CRAFT
5	17338	72 SWEETHEART FAIRY CAKE CASES, HEART OF WICKER SMALL, REX CASH+CARRY JUMBO SHOPPER, 10 COLOUR SPACEBOY PEN, 12 DAISY PEGS IN WOOD BOX, 12 EGG HOUSE PAINTED WOOD, 12 IVORY ROSE PEG PLACE SETTINGS, 12 MESSAGE CARDS WITH ENVELOPES, 12 PENCIL SMALL TUBE WOODLAND, 12 PENCILS SMALL TUBE RED RETROSPOT
	16907	DOORMAT RED RETROSPOT, HEART OF WICKER LARGE, RECIPE BOX PANTRY YELLOW DESIGN, REGENCY CAKESTAND 3 TIER, SET OF 3 CAKE TINS PANTRY DESIGN, REX CASH+CARRY JUMBO SHOPPER, HEART OF WICKER SMALL, PARTY BUNTING, PINK HEART SHAPE EGG FRYING PAN, HOME BUILDING BLOCK WORD

5.3. Metric comparisons

In this section, we used Precision, Recall, and F1-score to directly compare the performance of recommendation systems trained under different conditions.

Firstly, we compared the performance of the baseline model trained with and without the mined patterns. The model trained with pattern mining achieved an Average Precision of 0.051, Average Recall of 0.038, and Average F1-score of 0.036, while the model without pattern mining achieved an Average Precision of 0.051, Average Recall of 0.037, and Average F1-score of 0.035. It can be observed that the model trained with pattern mining slightly outperformed the model without pattern mining.

Next, we conducted a study on the hyper-parameter n in the recommendation system. We compared the test results of five recommendation systems with n=1, 3, 5, 7, and 9, as shown in the following figure. It can be seen that as n increases, the recommendation system considers more neighboring items, and the performance of the model also improves.



6. Conclusion and Recommendations

In this project, we used pattern mining to obtain potential sales information based on the store's past sales data, and designed an item-based recommendation system. Our goal is to use pattern mining to obtain hidden sales patterns and provide input on future sales strategies based on these patterns. At the same time, we hope to design the recommendation system to make it easier for customers to find items they may like, thereby potentially increasing sales.

Our pattern mining results revealed some very popular products, such as WHITE HANGING HEART T-LIGHT HOLDER, which can be considered to focus on such products in subsequent purchase strategies. The association rules we generated reveal the strong correlation between some products, such as buying POPPY'S PLAYHOUSE BEDROOM and preferring to buy POPPY'S PLAYHOUSE KITCHEN at the same time.

Based on these generation rules, we further design an item-based recommendation system. Our experiments show that baseline recommender systems that introduce patterns outperform recommender systems without patterns, suggesting that using pattern-assisted training in recommender systems can help improve the performance of recommender systems. The precision of our recommendation system is 0.051, and if the recommendation is made randomly, the precision is about 10/2500=0.004, which shows that our recommendation system can make it easier for customers to find the items they want, thereby potentially increasing sales.

We also studied the impact of changing the hyperparameter n of the recommendation model on the performance of the recommendation system. The experimental results show that as n increases, the performance of the recommendation system will improve, which means that referring to more neighbor items helps to improve the performance of the model. , this is also a method that can further improve the recommendation system in the future. However, it should also be noted that if n is too large, overfitting and high computational costs may occur. In

addition, there are some other methods that can be used to improve our recommendation system, such as changing the calculation method of similarity scores.

7. Reflection

In this project, we first applied pattern mining and recommendation system techniques to real-world problems. Combining experimental design and results, we have gained the following insights:

1. Data mining techniques enhance our ability to interpret data. By using pattern mining and other data mining techniques, we were able to grasp the inherent attributes of the data, such as data distribution and relationships between data. These attributes help us better understand the data and extract useful information for decision-making, which directly assists us in solving practical problems

2. Understanding of recommendation systems. We attempted to build a recommendation system model and trained and tested it on data. We became familiar with the entire design process of the recommendation system, as well as several important factors that affect the recommendation effect, such as the number of neighbors referenced. This knowledge will help us design more effective models in different practical problems in the future.

3. Improvements. We realized that a deep understanding of the data is crucial for solving practical problems effectively: 1. The size of the dataset determines which algorithm is more suitable, and different tasks require different types of models, which is important in practical problems. 2. The content of the dataset and the interaction with the model will generate specific problems. Understanding these issues is equally important. For example, the cold start problem in recommendation systems requires specialized strategies to solve. Therefore, in order to improve future performance, by understanding the task content based on data, we can better discover potential problems of the task and choose appropriate solutions.

8. References

1. Fournier-Viger, P., Gan, W., Wu, Y., Nouioua, M., Song, W., Truong, T., & Duong, H. (2022, April). Pattern mining: Current challenges and opportunities. In *International Conference on Database Systems for Advanced Applications* (pp. 34-49). Cham: Springer International Publishing.

2. Ricci, F., Rokach, L., & Shapira, B. (2021). Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, 1-35.

3. Aggarwal, C. C., Bhuiyan, M. A., & Hasan, M. A. (2014). *Frequent pattern mining algorithms: A survey* (pp. 19-64). Springer International Publishing.

4.https://medium.com/@pressiclinicalogalan/invoice-growth-algorithm-how-to-analyze-user-behavior-and-outrank-someone-competitors-from-39affan08879neither

5. https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/

6. Workshop 5: Frequent Itemset Mining and Recommender Systems