

**WQD7005 DATA MINING**

**ASSIGNMENT 2 REPORT**

**SEMESTER 2, SESSION 2023/2024**

**ASSIGNMENT: FINDING FREQUENT ITEM-SETS USING FP GROWTH ALGORITHM**

**Group: G2**

**Teacher: Tutut Herawan**

|  |  |
| --- | --- |
| **Name** | **Matric Number** |
| SAGAL MOHAMED YUSUF | 22093010 |
| TARSVINI A/P RAVINTHER | 17193844 |
| RABITA BHUIYA TANAYA | 22111473 |
| ZHANGXUAN | 22091244 |

**TASK 1**

**Association Rule Mining:**

The goal of association rules mining, a robust data mining technique, is to unearth interesting correlations and associations between variables in large datasets. Finding common if-then relationships, or association rules, between collections of objects in relational or transactional databases helps to reveal patterns, correlations, or causal structures. Market basket analysis is a prime illustration of this method since it examines consumer buying habits to determine the most frequently purchased items in a set [1].

In association rules mining, several key concepts play crucial roles:

* **Support:** This statistic shows what percentage of all database transactions contain a given itemset. It gives you a sense of how popular or common an itemset is by measuring how frequently it shows up in the dataset.
* **Confidence:** This is the probability that, for every purchase of item A, item B is also bought. To emphasize the close relationship between the two things, it measures the conditional likelihood of purchasing item B if item A is already part of the transaction.
* **Lift:** This represents the proportion of actual support to the amount of support that would be expected if items A and B were not related. The likelihood of buying item A and item B together is quantified relative to the likelihood of buying them separately. If the lift value is bigger than 1, it means that the items are positively correlated, indicating that they are purchased together more often than what would be predicted by chance.

Through these concepts, association rules mining helps in uncovering hidden patterns in data, providing valuable insights for decision-making in various domains.

**FP-Growth (Frequent Pattern Growth):**

For massive datasets, FP-Growth is an efficient technique that can mine frequent item sets without generating candidates [2]. A data structure called an FP-tree (Frequent Pattern Tree) is used to compress the input database, which is how the process works. The original data contains crucial information regarding the itemset associations, and this tree structure preserves that knowledge. Next, the FP-Growth algorithm systematically extracts common item sets from the FP-tree by recursive traversal, revealing valuable patterns in the data.

**Importance in Data Mining:**

FP-Growth plays a pivotal role in data mining due to several key advantages:

**Reduction in Computational Cost:** FP-Growth Sidesteps the Generation and Testing of Candidate Item sets Necessitated by the Apriori Algorithm [3]. It improves performance and cuts down on computing overhead by not generating candidates or scanning databases several times.

**Making Effective Use of Massive Datasets:** For efficient management and processing of massive datasets, FP-Growth employs the FP-tree structure. Compared to other common itemset mining techniques, it is both faster and more scalable thanks to its data compression and recursive pattern extraction capabilities.

**Business-Useful Insights:** FP-Growth's pattern and association detection capabilities yield useful information that can guide numerous company choices. To illustrate the point, it can help with inventory management and product placement in the retail sector by revealing the product combinations that customers often buy together. By gaining insight into consumer habits and preferences, it aids marketers in creating more precise and relevant advertising campaigns. If used consistently, FP-Growth can improve operational efficiency and bottom-line results by facilitating data-driven decision-making.

**Real-World Applications of FP-Growth:**

**Market Basket Analysis:** Retailers can use FP-Growth to learn more about their customers' buying habits and the items that go together most often [4]. By using this data, stores may better arrange their products, with related items shown next to each other to boost sales. On top of that, with the help of data-driven insights, stores can create hyper-targeted marketing campaigns that highlight frequently purchased items, increase sales, and delight customers.

**Recommendation Systems:** To improve their recommendation systems, e-commerce companies use FP-Growth. This is because these platforms predict what customers would buy by looking at the common item sets purchased by people who are like them. Customers are more engaged, have a better shopping experience, and have more successful purchases due to this tailored recommendation method.

**Fraud Detection:** To detect suspicious trends in customer transactions that could point to fraud, financial institutions use FP-Growth [5]. Banks and credit card firms can improve their ability to detect possibly fraudulent transactions by detecting frequent item sets that differ from regular transaction behavior. By taking a preventative stance against fraud, we can keep our financial processes secure, and our customers' assets protected.

**Healthcare:** In medical research, FP-Growth plays a crucial role in identifying links between different symptoms and diseases. Investigators can find commonalities between symptoms and diseases by examining patient records. By improving diagnostic methods, these insights allow healthcare providers to make faster and more accurate illness diagnoses. Furthermore, FP-Growth helps improve patient care and results by helping develop more effective treatment plans through the revelation of common treatment responses among patients with similar profiles.

**Challenges in Developing FP-Growth Model:**

**Memory Usage:** The memory utilization of the FP-Growth model is one of the major obstacles to its development. The FP-tree, an integral part of the FP-Growth algorithm, can grow to enormous sizes when working with datasets that contain many unique objects or a substantial number of transactions. The memory usage is high due to the enormous size, which puts a burden on system resources and restricts the algorithm's efficiency when dealing with large datasets.

**Complexity in Tree Construction:** The FP-tree construction is an involved and painstaking procedure. To build the tree structure and properly record the item sets' frequencies, many passes through the dataset are necessary. To keep the FP-tree intact, it is also necessary to handle the node links carefully. Because of its intricacy, building trees can be a computationally and time-consuming process, which presents issues with implementation and performance optimization.

**Scalability:** Even while FP-Growth outperforms the Apriori approach, it is still not easy to scale it up to handle massive datasets or run it in a distributed setting. The algorithm must handle and process large volumes of data, and it must work well on various scales, which calls for advanced methods and solid infrastructure. To use FP-Growth in huge data situations, we must solve this scaling problem.

**Data Sparsity:** There is also a substantial problem of data sparsity. Many rare but distinct objects in a dataset might make it such that building an FP-tree doesn't significantly improve compression. The FP-tree may not be as efficient or compact due to the sparse data, which means that performance improvements will be decreased. To overcome this obstacle and keep the algorithm's efficiency, it must be optimized for sparse datasets.

When compared to older methods like the Apriori algorithm, FP-Growth provides clear benefits as a robust and efficient algorithm for mining frequent item sets. From healthcare and fraud detection to retail and recommendation systems, it finds use in a broad variety of domains. Memory consumption, tree construction difficulty, scalability, and data sparsity are some of the issues that the FP-Growth model encounters, despite its virtues [6]. It is essential to tackle these problems if we want to use FP-Growth to its fullest capacity in big data applications and keep it as a useful tool for finding interesting connections and patterns in massive datasets.

# **Task 2: Dataset and Tool Evaluation**

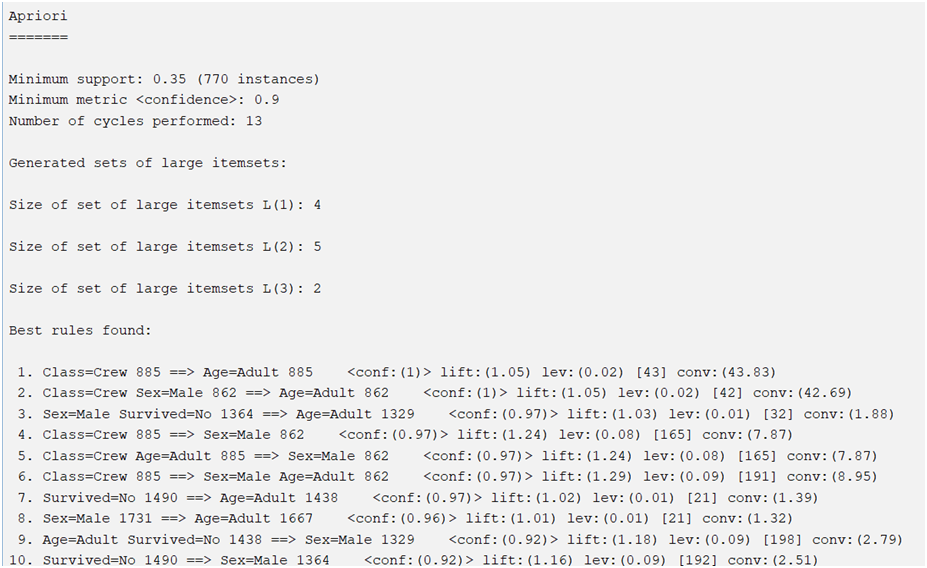
1. Motivation to Select an Appropriate Dataset for Association Rule Mining

Market basket research uses a lot of association rule mining, especially using methods like FP-Growth, to find links between items in big transaction datasets. The requirement to obtain insightful and useful information that can enhance business choices and tactics is the main driver behind choosing a suitable dataset.

The necessity to conduct applicable and significant market basket analysis drives the selection of a suitable dataset. For replication, the chosen dataset should be publicly accessible, reasonable in terms of size and complexity, and indicative of actual business transactions. These requirements are well met in our dataset, which makes it a great example for illustrating how to apply FP-Growth in programs like Weka and RapidMiner. This guarantees that the study produces useful, business-relevant findings.

1. FP Growth Using Weka

|  |  |
| --- | --- |
|  |  |
|  | Figure 1: Results in WEKA |

s

The decision between Weka and RapidMiner, two potent data mining tools with FP-Growth capabilities, is influenced by many variables, including user interface preferences, dataset complexity, and unique features provided by each tool. I'll go over how to utilize both tools for mining frequent patterns using FP-Growth below, and I'll provide some hypothetical datasets to support my tool selection.

The size of the dataset, the difficulty of the necessary analysis, and the user's familiarity with data mining tools all influence the decision between Weka and RapidMiner. Weka's simplicity and emphasis on basic methods make it ideal for small to medium datasets for instructional applications. Because of its comprehensive capabilities and scalability, RapidMiner is better suited for complicated analytical needs and larger datasets.

**Task 3**

**Results**

Based on the WEKA results, minimum support is 0.35 (770 instances). This means that any itemset included in the results must appear in at least 35% of the total transactions. Minimum confidence is 0.9. This indicates that the rules included in the results must have confidence of at least 90%. The number of cycles performed is 13. This refers to the number of iterations the algorithm performed to find the frequent itemsets.

Besides, there are 4 large itemsets of size 1, 5 large itemsets of size 2 and large itemsets of size 3. Based on figure 1, there are 10 best rules found. But we will just explain the rules with 100% confidence.

Rule 1: Class=Crew 885 ==> Age=Adult 885

Rule 2: Class=Crew Sex=Male 862 ==> Age=Adult 862

Rule 1 and Rule 2 have 100% confidence.

Whenever 'Class=Crew' is true, 'Age=Adult' is also true 100% of the time. There are no instances where a crew member is not an adult in the dataset.

Whenever both 'Class=Crew' and 'Sex=Male' are true, 'Age=Adult' is also true 100% of the time. There are no instances where a male crew member is not an adult in the dataset.

**Summary**

Association rule mining is a powerful data mining technique used to discover interesting correlations and associations between variables in large data sets. Key concepts in association rule mining include support, confidence, and lift, which help reveal hidden patterns in decisions. FP-Growth is an efficient technique for mining frequent itemsets without generating candidates, using FP trees to compress the input database and systematically extracting common itemsets through recursive traversal. FP-Growth has significant advantages in reducing computational costs, effectively using massive data sets, and providing business-useful insights, although challenges in developing FP-Growth models include high memory usage, complexity of tree construction, scalability issues and data sparsity. But FP-Growth remains a valuable tool for discovering interesting connections and patterns in massive data sets.

The dataset is used to illustrate the application of FP-Growth in tools such as Weka and RapidMiner. The choice between Weka and RapidMiner depends on factors such as user interface preferences, dataset complexity, and the specific capabilities of each tool. Weka is better suited for small to medium-sized data sets and educational purposes, while RapidMiner is better suited for complex analysis needs and larger data sets.

WEKA results show that the algorithm performed 13 cycles to find frequent itemsets with a minimum support of 0.35 and a minimum confidence of 0.9. Among the rules discovered with 100% confidence: "Class=Crew" always corresponds to "Age=Adult" and "Class=Crew" combined with "Sex=Male" always corresponds to "Age= Adult", indicating that there are no crew members or male crew members who are not adults in the data set.

**References:**

[1] R. Sethi, “Market Basket Analysis of Instacart,” 2023. <http://www.ir.juit.ac.in:8080/jspui/handle/123456789/9972>

[2] N. Boyko and O. Tkachyk, “Model for Finding Frequent Sets in FP-growth for Multimodal Data,” vol. 3137, pp. 126–143, Jan. 2022, doi: 10.32782/cmis/3137-11.

[3] M. SMythili and A. R. M. Shanavas, “Performance Evaluation of Apriori and FP-Growth Algorithms,” *International Journal of Computer Applications*, vol. 79, no. 10, pp. 34–37, Oct. 2013, doi: 10.5120/13779-1650.

[4] M. R. Pradana, M. Syafrullah, H. Irawan, N. Irawan, J. C. Chandra, and A. Solichin, “Market Basket Analysis Using FP-Growth Algorithm On Retail Sales Data,” Oct. 2022, doi: 10.23919/eecsi56542.2022.9946478.

[5] X. Luo, “Suspicious Transaction Detection for Anti-Money Laundering,” *International Journal of Security and Its Applications*, vol. 8, no. 2, pp. 157–166, Mar. 2014, doi: 10.14257/ijsia.2014.8.2.16.

[6] A. Çoğalan, “FP-Growth Algorithm: How to Analyze User Behavior and Outrank Your Competitors,” *Medium*, Mar. 22, 2023. [Online]. Available: <https://medium.com/@anilcogalan/fp-growth-algorithm-how-to-analyze-user-behavior-and-outrank-your-competitors-c39af08879db>